U.P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 302N/ MASTAT – 302N

Multivariate Analysis

Block – 1 : Multivariate Normal Distribution and Estimation of Parameter

Unit – 1 : Multivariate Normal Distribution

Unit – 2 : MLE of Parameters and Different Coefficients

Unit – 3 : Sampling Distributions

Block - 2 : Distributions Related to MND and Their Applications

Unit – 4 : Wishart Distribution

Unit – 5 : Hoteling's T2 Statistic

Unit – 6 : Mahalnobis D2

Unit – 7 : Discriminant Analysis

Block – 3 : Advance Multivariate Analysis

Unit – 8 : Advance Analysis

Unit – 9 : Principal Component Analysis

Unit – 10 : Factor Analysis

Unit – 11 : Tests of Hypothesis

Unit – 12 : Linear Regression Model

Course Design Committee

Dr. Ashutosh Gupta — **Chairman**
Director, School of Sciences
U. P. Rajarshi Tandon Open University, Prayagraj

Prof. Anoop Chaturvedi — **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. S. Lalitha — **Member**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Himanshu Pandey — **Member**
Department of Statistics
D. D. U. Gorakhpur University, Gorakhpur.

Prof. Shruti — **Member-Secretary**
Professor, School of Sciences
U.P. Rajarshi Tandon Open University, Prayagraj

Course Preparation Committee

Dr. Anupma Singh — **Writer**
Department of Statistics
Ewing Christian College,
University of Allahabad, Prayagraj

Prof. Anoop Chaturvedi — **Editor**
Ex. Head, Department of Statistics
University of Allahabad, Prayagraj

Prof. Shruti — **Course Coordinator**
School of Sciences,
U. P. Rajarshi Tandon Open University, Prayagraj

MScSTAT – 302 N/ MASTAT – 302 N **MULTIVARIATE ANALYSIS**

First Edition: ***July 2024***
ISBN :

Printed By:

Blocks & Units Introduction

The present SLM on ***Multivariate Analysis*** consists of twelve units with three Blocks.

The ***Block 1- Multivariate Normal Distribution and Estimation of Parameters*** is the first block of said SLM, which is divided into three units.

The ***Unit 1 - Multivariate Normal Distribution***, is the first unit of present self-learning material, which describes Multivariate normal distribution, Moment generating function, Characteristic function, marginal and conditional distributions, multiple and partial correlation coefficient.

In ***Unit 2 - MLE of Parameters and Different Coefficients,*** deals with the Maximum Likelihood Estimation of Parameters. In particular the maximum likelihood estimators of the mean vector and covariance matrix, sample Multiple and partial correlation coefficients, regression coefficient have been derived.

In ***Unit 3 - Sampling Distributions,*** we discuss Sampling Distributions of sample mean vector, Null sampling distributions of Multiple and Partial Correlations, distribution of sample regression coefficient. Distribution of the matrix of sample regression coefficients and the matrix of residual sum of squares and cross products, Rao's U-statistic, its distribution, and applications.

The ***Block 2 - Distributions Related to the Multivariate Normal Distribution and Their Applications*** is the second block of said SLM, which is divided into four units.

The ***Unit 4 - Wishart Distribution***, is discusses the Wishart distribution derives its characteristic function, proves the additive property of Wishart distribution and Cochran theorem and derives the distribution of characteristic roots and vectors of Wishart matrices.

The ***Unit 5 - Hotelling's T^2 Statistic***, is discusses Hotelling's T^2 Statistic and obtains its distribution. The unit also gives various applications in tests for the mean vector of one and more multivariate normal population.

The ***Unit 6 - Mahalnobis D^2***, discusses Equality of the component of a mean vector in a multivariate normal population, discusses the Mahalanobis D^2 and its various applications.

The ***Unit 7 - Discriminant Analysis***, is focuses on the Discriminant analysis and classification and discrimination. The procedures for discrimination between two multivariate normal populations, the sample discriminant function, and the tests associated with discriminant functions are given. The probabilities of miss classification and their estimation and classification into more than two multivariate normal populations are also discussed along with the Fisher-Behrens Problem.

The ***Block 3 – Advance Multivariate Analysis*** discusses some advanced level topics of Multivariate Analysis.

In ***Unit 8 – Advance Analysis,*** we consider an improved shrinkage estimator for the multivariate normal mean vector. The inadmissibility of maximum likelihood estimator of mean vector of multivariate normal distribution is shown when dimension is greater than three, James-Stein estimator of the mean vector and improved estimation of dispersion matrix of a MN is given and its dominance over the MLE is established.

The ***Unit 9 – Principal Component Analysis,*** considers the principal component analysis for the multivariate data. The interpretation of principal components, their maximum likelihood estimators, sample variances, canonical correlation and variable, the procedures for selecting appropriate number of principal components have been discussed. Also discussed the Interference on canonical correlations.

The ***Unit 10 – Factor Analysis,*** discusses the factor Analysis, linear factor models, estimation of factor loadings, factor rotation, and the estimation of factor scores.

The objective of ***Unit 11 – Tests of Hypothesis,*** is to discuss the tests for the equality of covariance matrices, sphericity tests for covariance matrix, equality of mean vector and covariance matrix to specified vector and matrix.

The ***Unit 12 – Linear Regression Model,*** considers the Multivariate analysis of variance [MANOVA] of one-way classified data. Wilk's lambda criterion and other testing criterion for testing the equality of means of different groups of categorical data.

At the end of every block/unit the summary, self-assessment questions and further readings are given.

UNIT-1 MULTIVARIATE NORMAL DISTRIBUTION

Structure

1.1 Introduction

In Multivariate analysis we consider statistical analysis of data consisting of sets of measurements on a number of individuals or objects. If each member of the population exhibits a set of values, one for each of the variables under consideration, then such type of population is called a Multivariate population. A sample drawn from such type of population is called a multivariate sample.

Example:

1. A dietician collects patient data on cholesterol, blood pressure, sugar levels and weight. She also collects data on dietary habits. Using Multivariate Data Analysis, she can determine how much each element of diet influences health outcomes.

2. A researcher has collected data on three demographic variables and four academic variables (let's say standardized test scores) for $1{,}000$ students along with the programmes they are enrolled for. The researcher wants to determine how demographics and academic variables are related with the choice of program.

3. The football league table is an example of multivariate data. Here $W =$ number of wins, $D =$ number of draws, $F =$ number of goals scored and $A =$ number of goals conceded for four teams. In this example we have $p = 4$ variables $(W, D, F, A)'$ measured on $n = 4$ cases (teams).

Team	W	D	F	A
Argentina	1	2	4	3
Portugal	1	2	2	1
USA	1	1	3	3
France	0	1	0	2

The data vector for the Argentina is $x^T = (1,2,4,3)$.

1.1.1 Notations of Multivariate Distribution

1. If $X_1, X_2 \ldots, X_p$ are p random variables then Cumulative distribution function (cdf) is given by

$$F(x_1, x_2, \ldots, x_p) = P\{X_1 \leq x_1, X_2 \leq x_2, .., X_p \leq x_p\}$$

Defined for every set of real numbers $x_1, x_2, \cdots x_p$.

2. If $F(x_1, x_2, \ldots, x_p)$ is absolutely continuous, the joint density function of $X_1, X_2 \ldots, X_p$is

$$\frac{\partial^p F(x_1, x_2, \ldots, x_p)}{\partial(x_1, x_2, \ldots, x_p)} = f\,(x_1, x_2, \ldots, x_p)$$

and

$$F(x_1, x_2, \ldots, x_p) = \int_{-\infty}^{x_p} \ldots \int_{-\infty}^{x_1} f(u_1, u_2, \ldots, u_p) du_1, du_2, \ldots, du_p$$

3. The probability of falling in any measurable set R in the p-dimensional Euclidean space:

$$P\left\{(X_1, X_2, \ldots, X_p) \in R\right\} = \int_R f(x_1, x_2, \ldots, x_p) dx_1, dx_2, \ldots, dx_p$$

4. Joint moments:

$$E\left[X_1^{h_1}, X_2^{h_2}, \ldots, X_p^{h_p}\right] = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} x_1^{h_1}, x_2^{h_2}, \ldots, x_p^{h_p}\ f(x_1, x_2, \ldots, x_p) dx_1, dx_2, \ldots, dx_p$$

5. The marginal cdf of $X_1, X_2, \dots, X_p$ $(r < p)$ is given by

$$\Pr\{X_1 \le x_1, \dots, X_r \le x_r, X_{r+1} < \infty, \dots, X_p < \infty\} = F(x_1, x_2, \dots, x_r, \infty, \dots, \infty)$$

$$= \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_r} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(u_1, u_2, \dots, u_p)\, du_1, du_2, \dots, du_p$$

6. Random variables $X_1, X_2 \dots, X_p$ are said to be mutually independent if

$$F(x_1, x_2, \dots, x_p) = F_1(x_1)F_2(x_2) \dots F_p(x_p)$$

where $F_i(x_i)$ is the marginal cdf of $X_i (i = 1,2, \dots, p)$.

Similarly, the set $X_1, X_2 \dots, X_r$ is said to be independent of set $X_{r+1}, \dots, X_p$ if

$$F(x_1, x_2, \dots, x_p) = F(x_1, x_2, \dots, x_r, \infty, \dots, \infty)F(\infty, \infty, \dots, \infty, x_{r+1}, \dots, x_p)$$

If $X_1, X_2 \dots, X_p$ are mutually independent

$$E\left[x_1^{h_1} x_2^{h_2} \dots x_p^{h_p}\right] = \prod_{i=1}^{p} E(x_i^{h_i})$$

7. The conditional density of $X_1, X_2, \dots, X_r$ given $X_{r+1} = x_{r+1}, \dots, X_p = x_p$, is

$$\frac{f(x_1, x_2, \dots, x_p)}{\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(u_1, u_2, \dots, u_r, x_{r+1}, x_{r+2}, \dots, x_p) du_1, du_2, \dots, du_r}$$

If we write $X = \left(X_{(1)}' \;\; X_{(2)}'\right)'$, where $X_{(1)}$ is consist of first r elements of X and $X_{(2)}$ is consist of last $(p - r)$ elements of X then we have

$$f(x_{(2)}|x_{(1)}) = \frac{f(x_{(1)}, x_{(2)})}{f_1(x_{(1)})}$$

$f(x_{(1)}, x_{(2)})$: Joint pdf of $X_{(1)}, X_{(2)}$.

$f_1(x_{(1)})$: Marginal pdf of $X_{(1)}$.

8. Transformation of variables: Let

$$y_i = y_i(x_1, x_2, \dots, x_p); \qquad i = 1,2, \dots, p$$

We assume that transformation from the x-space to the y- space is one to one.

The inverse transformation is

$$x_i = x_i(y_1, y_2, \dots, y_p); \qquad i = 1,2, \dots, p$$

The random variables $Y_1, Y_2, \dots, Y_p$ are defined as

$$Y_i = Y_i(X_1, X_2, \dots, X_p); \qquad i = 1,2, \dots, p$$

The joint density of $Y_1, Y_2, \dots, Y_p$ is

$$g(y_1, y_2, \dots, y_p) = f(x_1, x_2, \dots, x_p) J(y_1, y_2, \dots, y_p)$$

The Jacobean of the transformation is given by

$$J(y_1, y_2, \dots, y_p) = mod \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \dots & \frac{\partial x_1}{\partial y_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial x_p}{\partial y_1} & \frac{\partial x_p}{\partial y_2} \dots & \frac{\partial x_p}{\partial y_p} \end{vmatrix}$$

1.2 Objectives

After going through this unit, learner should be able to:

• Understand the basic concepts of multivariate normal distribution

• Obtain the moment generating function and characteristic function

• Finding the marginal distribution and conditional distribution

• Get basic concept of multiple and partial correlation coefficient

1.3 Multivariate Normal Distribution

The pdf of univariate normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad ; \quad -\infty < x < \infty$$

X: random vector of dimension p

$$X = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix},$$

$$E(X) = \mu$$

$$= \begin{bmatrix} E(X_1) \\ \vdots \\ E(X_p) \end{bmatrix}$$

$$= \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix}$$

Variance: Measure of variation of random variable.

Covariance: Measure of joint variation of two random variables simultaneously.

$$Cov(X_i, X_j) = E\left[(X_i - \mu_i)(X_j - \mu_j)\right]$$

Variance-Covariance Matrix:

$$\Sigma = \begin{bmatrix} Cov(X_1, X_1) & \cdots & Cov(X_1, X_p) \\ \vdots & \ddots & \vdots \\ Cov(X_p, X_1) & \cdots & Cov(X_p, X_p) \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

Σ is a symmetric, positive definite matrix and is non-singular.

A $p \times 1$ vector $X = \left(X_1, X_2, \dots, X_p\right)'$ is said to have a p-variate (non-singular) normal distribution if its p.d.f. is of the form

$$f(x) = k.exp\left\{-\frac{1}{2}(x-b)'A(x-b)\right\};\ -\infty < x_i < \infty \quad (i = 1,2,\dots,p) \tag{1.1}$$

where $b = \left(b_1, b_2, \dots, b_p\right)'$ is a $p \times 1$ vector $(-\infty < b_i < \infty, i = 1,2,\dots,p)$,

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & \ddots & \vdots \\ a_{p1} & \cdots & a_{pp} \end{pmatrix} \tag{1.2}$$

is a positive definite, symmetric matrix of order p.

Further, $k\ (> 0)$ is a constant chosen so that the integral of $f(x)$ over the entire p-dimensional Euclidean space of $x_1, x_2, \dots, x_p$ is unit.

b, A: parameters of this distribution

Result 1.3.1: The value of the normalizing constant $k = |A|^{\frac{1}{2}} (2\pi)^{-\frac{p}{2}}$

Proof: Since A is positive definite we have

$$(x - b)'A(x - b) \geq 0 \tag{1.3}$$

There exists a non-singular matrix C such that

$C'A\,C = I$

where I denote a $p \times p$ identity matrix.

Let $(x - b) = Cy$

$$\left(x_i = b_i + \sum_{j=1}^{p} c_{ij} y_j \right)$$

where $y = (y_1, y_2, \dots, y_p)'$.

Then

$(x - b)' A (x - b)$

$= y' C' A C y$

$= y' y$

The Jacobean of the transformation from x to y is

$J(y_1, y_2, \dots, y_p)$

$$= mod \begin{bmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_p} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial x_p}{\partial y_1} & \frac{\partial x_p}{\partial y_2} & \cdots & \frac{\partial x_p}{\partial y_p} \end{bmatrix}$$

$= mod|C|$

Since $|C'AC| = 1$, we obtain $|C| = |A|^{-1/2}$.

The integral of (1.1) over the p dimensional space is one, we shall evaluate

$$k^* = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x-b)' A (x-b)} \, dx_1 \dots dx_p \; mod \; |C|$$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-\frac{1}{2}y'y} \; dy_1 \ldots dy_p \, |A|^{-\frac{1}{2}}$$

$$= |A|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left(\prod_{j=1}^{p} e^{-\frac{1}{2}y_j^2} \right) dy_1 \ldots dy_p$$

$$= |A|^{-\frac{1}{2}} \prod_{j=1}^{p} \int_{-\infty}^{\infty} e^{-\frac{1}{2}y_j^2} dy_j$$

$$= |A|^{-\frac{1}{2}} \prod_{j=1}^{p} \sqrt{(2\pi)}$$

$$= |A|^{-\frac{1}{2}} (2\pi)^{\frac{p}{2}} \qquad (1.4)$$

Substituting the values of integral from (4) into (3), we obtain

$$k = \frac{1}{k^*} \Rightarrow k = |A|^{\frac{1}{2}} (2\pi)^{-\frac{p}{2}} \qquad (1.5)$$

Therefore, the density function of p-variate normal distribution is

$$f(x) = |A|^{\frac{1}{2}} (2\pi)^{-\frac{p}{2}} \exp\left\{-\frac{1}{2}(x - b)'A(x - b)\right\}$$

The pdf of Y is

$$f(y) = \frac{|A|^{\frac{1}{2}}}{(2\pi)^{\frac{p}{2}}} \left\{|A|^{-\frac{1}{2}} e^{-\frac{1}{2}y'y}\right\}$$

$$= \frac{1}{(2\pi)^{\frac{p}{2}}} \left\{e^{-\frac{1}{2}y'y}\right\}$$

$$= \prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}} \; e^{-\frac{1}{2}y_j^2}$$

Therefore $y_1, y_2, \dots, y_p$ are independently distributed.

Theorem 1.3.2: The density of p dimensional random vector X is

$$f(x) = |A|^{\frac{1}{2}} (2\pi)^{-\frac{p}{2}} \exp\left\{-\frac{1}{2}(x - b)'A(x - b)\right\}$$

The expected value of X is b and the covariance matrix is A^{-1}. Conversely, given a mean vector μ and positive definite vector Σ, the multivariate normal density is,

$$f(x_1, x_2, \dots, x_p) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \, exp\left\{-\frac{1}{2}(x - \mu)' \, \Sigma^{-1}(x - \mu)\right\}$$

Show that $E(X) = \mu$ and $Var(X) = \Sigma$

Proof: Let

$$X = CY + b \tag{1.6}$$

where C and y are as defined in theorem 1.3.1. Taking expectation on both sides, we get

$$E(X) = CE(Y) + b \tag{1.7}$$

The expected value of i^{th} component y_i of Y is

$$E(y_i) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} y_i \left\{\prod_{j=1}^{p} \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2}y_j^2}\right\} \, dy_1 \cdots dy_p$$

$$= \left\{\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i \, e^{-\frac{1}{2}y_i^2} dy_i\right\} \left\{\prod_{j(\neq i)=1}^{p} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \, e^{-\frac{1}{2}y_j^2} \, dy_j\right\}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i \, e^{-\frac{1}{2}y_j^2} \, dy_j$$

$$= 0$$

Notice that $y_i\, e^{-\frac{1}{2}y_j^2}$ is an odd function of y_i leading to

$$\int_{-\infty}^{\infty} y_i\; e^{-\frac{1}{2}y_j^2}\, dy_j = 0.$$

Thus $E(Y) = 0$. Putting this value in (1.7), we have

$E(X) = b = \mu$ (say).

The variance-covariance matrix of X is

$$E(X - \mu)\,(X - \mu)' = CE(YY')\,C' \qquad (1.8)$$

If $i = j$,

$$EY_i^2 = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} y_i^2 \left\{ \prod_{k=1}^{p} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}y_k^2} \right\} dy_1 \cdots dy_p$$

$$= \left\{ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i^2\, e^{-\frac{1}{2}y_i^2} dy_i \right\} \left\{ \prod_{k=1\neq i}^{p} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}y_k^2}\, dy_k \right\}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} y_i^2\, e^{-\frac{1}{2}y_i^2} dy_i$$

$$= \frac{1}{\sqrt{2\pi}} 2 \int_{0}^{\infty} \mathrm{y_i^2}\, e^{-\frac{1}{2}y_i^2} dy_i \qquad (since\ y_i^2 e^{-\frac{1}{2}y_j^2}\ is\ an\ even\ function\ of\ y_i)$$

$$= \frac{2\sqrt{2}}{\sqrt{2\pi}} \int_{0}^{\infty} z_i\, e^{-z_i}\, dz_i \qquad \left(\frac{1}{2}y_i^2 = z_i \;\Rightarrow\; y_i\, dy_i = dz_i\right)$$

$$= \frac{1}{\sqrt{2\pi}}\, 2\sqrt{2}\Gamma\left(\frac{3}{2}\right) = \frac{1}{\sqrt{2\pi}}\, 2\, \frac{1}{2}\Gamma\left(\frac{1}{2}\right) = 1$$

If $i \neq j$

$$E(Y_iY_j)$$

$$= \left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y_i\ e^{-\frac{1}{2}y_i^2} dy_i\right\}\left\{\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} y_j\ e^{-\frac{1}{2}y_j^2}\ dy_j\right\}$$

$$= 0$$

Therefore,

$$E(YY')$$

$$= \begin{bmatrix} E(Y_1^2) & E(Y_1Y_2) & \cdots & E(Y_1Y_p) \\ E(Y_2Y_1) & E(Y_2^2) & \cdots & E(Y_2Y_p) \\ \vdots & \vdots & \vdots & \vdots \\ E(Y_pY_1) & E(Y_pY_2) & \cdots & E(Y_p^2) \end{bmatrix}$$

$$= I$$

Since $C'AC = I \Rightarrow A^{-1} = CC'$

Then

$$E(X-b)\ (X-b)'$$

$$= CC'$$

$$=\ A^{-1}$$

If we write μ as the mean vector of x and Σ as the variance covariance matrix of x, then

$$\mu = b,$$

$$\Sigma =\ A^{-1} = \left((\sigma_{ij})\right)$$

$$\sigma_{ij} = E(X_i -\ \mu_i)(X_j -\ \mu_j)$$

We denote the p.d.f.

$$f(x_1, x_2, \dots, x_p) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\ (x-\mu\ )'\Sigma^{-1}(x-\ \mu)\right\}$$

Example: Find mean vector μ and variance covariance matrix or dispersion matrix Σ of the following density functions:

(i) $f(x,y) = \frac{1}{2\pi} exp\left[-\frac{1}{2}\ \{(x-1\ )^2 + (y-\ 2)^2\}\right]$

(ii) $f(x,y) = \frac{1}{2.4\,\pi} exp\left\{-\frac{1}{0.72}\ \left(\frac{x^2}{4} - 1.6\frac{xy}{2} + y^2\right)\right\}$

Solution: If $(X,Y) \sim N_2(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \rho)$, then

$f(x,y)$

$$= \frac{1}{(2\pi)\sigma_1\sigma_2\sqrt{1-\rho^2}} exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right\}\right]$$

(i) We have

$f(x,y)$

$$= \frac{1}{2\pi} exp\left[-\frac{1}{2}\ \{(x-1\ )^2 + (y-\ 2)^2\}\right]$$

$\Rightarrow \mu_1 = 1, \mu_2 = 2, \sigma_1 = 1, \sigma_2 = 1, \rho = 0$

i.e. $\mu = \begin{pmatrix}\mu_1 \\ \mu_2\end{pmatrix} = \begin{pmatrix}1 \\ 2\end{pmatrix}$ and $\Sigma = \begin{pmatrix}\sigma_1^2 & \rho\ \sigma_1\sigma_2 \\ \rho\ \sigma_2\sigma_1 & \sigma_2^2\end{pmatrix} = \begin{pmatrix}1 & 0 \\ 0 & 1\end{pmatrix}$

(ii) We have

$$f(x,y) = \frac{1}{2.4\,\pi} exp\left\{-\frac{1}{0.72}\ \left(\frac{x^2}{4} - 1.6\frac{xy}{2} + y^2\right)\right\}$$

or $f(x,y)$

$$= \frac{1}{(2\pi)(2)(1)(0.6)} exp\left[-\frac{1}{2 \times 0.36}\left\{\left(\frac{x}{2}\right)^2 - 2 \times 0.8\left(\frac{x}{2}\right)\left(\frac{y}{1}\right) + \left(\frac{y}{1}\right)^2\right\}\right]$$

$\Rightarrow 1 - \rho^2 = 0.36$

$\Rightarrow \rho^2 = 1 - 0.36 = 0.64$

$\Rightarrow \rho = \pm 0.8$

Hence $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 2, \sigma_2 = 1, \rho = 0.8$

i.e., $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\, \sigma_1\sigma_2 \\ \rho\, \sigma_2\sigma_1 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 4 & 1.6 \\ 1.6 & 1 \end{pmatrix}$

1.4 Moment Generating Function

The moment generating function of X is

$\Psi(t) == M_X(t) = E\left[e^{t'X}\right]$

Let $X = \mu + CY$

where C is the non-singular matrix such that $C'\Sigma^{-1}C = I$

Therefore $\Sigma = CC'$. Then, $Y \sim N_p(y|0, I_p)$

Now

$\Psi(t) = E\left[e^{t'X}\right]$

$= E\left[e^{t'\mu + t'CY}\right]$

$= e^{t'\mu}\, E\left[e^{t'CY}\right]$

Let $u = C't$

Then

$$\Psi(t) = e^{t'\mu}E\left[e^{u'Y}\right] = e^{t'\mu}E\left[\prod_{j=1}^{p} e^{u_jY_j}\right]$$

$$= e^{t'\mu}\prod_{j=1}^{p} E(e^{u_jY_j}) \qquad \left(since\ Y_1, Y_2, \dots, Y_p\ are\ independently\ distributed\right)$$

$$= e^{t'\mu}\prod_{j=1}^{p} e^{\frac{1}{2}u_j^2} \qquad (since\ y_j \sim N(0,1))$$

$$\Rightarrow \Psi(t) = e^{t'\mu}e^{\frac{1}{2}u'u} = e^{t'\mu+\frac{1}{2}t'CC't}$$

$$\Rightarrow \Psi(t) = e^{t'\mu+\frac{1}{2}t'CC't}$$

1.5 Characteristic Function

The characteristic function of X is

$$\phi(t) = E\left[e^{it'x}\right]$$

$$= E\left[e^{it'(Cy+\mu)}\right]$$

$$= e^{it'\mu}\,E\left[e^{iu'y}\right]$$

where $u = C't = \left(u_1, u_2, \dots, u_p\right)'$.

Since $y_j \sim N(0,1)$ and $y_j's$ are independently distributed

$$E\left[e^{iu'y}\right]$$

$$= E\left[\prod_{j=1}^{p} e^{iu_j y_j}\right]$$

$$= \prod_{j=1}^{p} E\left[e^{iu_j y_j}\right]$$

$$= \prod_{j=1}^{p} e^{-\frac{1}{2}u_j^2}$$

(Characteristic function of y_j with u_j as the argument is $e^{-\frac{1}{2}u_j^2}$)

$$\Rightarrow E\left[e^{iu'Y}\right] = e^{-\frac{1}{2}u'u} = e^{-\frac{1}{2}t'\,\Sigma t}$$

Therefore

$$\phi(t) = exp\left(it'\mu - \frac{1}{2}t'\,\Sigma t\right).$$

We can obtain the first two moments of X using the characteristic function as follows:

$$E(X_j) = \frac{1}{i}\left[\frac{\partial\phi(\underline{t})}{\partial t_j}\right]_{t=0} = \mu_j$$

$$E(X_j X_k)$$

$$= \frac{1}{i^2}\left[\frac{\partial\phi}{\partial t_j \partial t_k}\right]_{t=0}$$

$$= \sigma_{jk} + \mu_j\mu_k$$

1.5.1 Correlation Coefficient

The i^{th} diagonal element of Σ, σ_{ii}, is the variance of X_i, and the $(i,j)^{th}$ element of Σ, σ_{ij}, is the covariance between X_i and X_j. The correlation coefficient between X_i and X_j is given by

$$\rho_{ij} = \frac{\sigma_{ij}}{\left(\sigma_{ii}.\sigma_{jj}\right)^{\frac{1}{2}}}$$

1.5.2 Multiple Correlation Coefficients

If X_1 is the first component of X and $X^{(2)}$ the vector of remaining $(p-1)$ components. Here first expression X_1 as a linear combination of $X^{(2)}$ defined by the relation

$$X_1^* = \mu_1 + \beta'\left(X^{(2)} - \mu^{(2)}\right) \tag{1.9}$$

The coefficient vector β is determined by minimizing

$$U = E[X_1 - X_1^*]^2 = E\left[X_1 - \mu_1 - \beta'\left(X^{(2)} - \mu^{(2)}\right)\right]^2$$

Differentiating with respect to β and equating to zero, we have

$$\frac{\partial U}{\partial \beta} = 0$$

$$\Rightarrow -2E\left[(X_1 - \mu_1) - \beta'\left(X^{(2)} - \mu^{(2)}\right)\right]\left(X^{(2)} - \mu^{(2)}\right)' = 0$$

$$\Rightarrow E(X_1 - \mu_1)\left(X^{(2)} - \mu^{(2)}\right)' - \beta' E\left(X^{(2)} - \mu^{(2)}\right)\left(X^{(2)} - \mu^{(2)}\right)' = 0$$

$$\Rightarrow \sigma_{12}' = \beta'\Sigma_{22}$$

Or

$$\hat{\beta}' = \sigma_{12}'\Sigma_{22}^{-1}$$

Here $\sigma_{12}' = E(X_1 - \mu_1)\left(X^{(2)} - \mu^{(2)}\right)'$,

Putting this value in (1.9), we get

$$X_1^* = \mu_1 + \sigma_{12}'\Sigma_{22}^{-1}\left(X^{(2)} - \mu^{(2)}\right) = \hat{X}_1$$

The correlation coefficient between X_1 and $X^{(2)}$ is called Multiple correlation between X_1 and $X_2, X_3, \dots, X_p$. It is denoted by

$$\rho_{1.(2,3,\dots,p)} = \frac{Cov\left(X_1, \hat{X}_1\right)}{\sqrt{Var(X_1)\, Var\left(\hat{X}_1\right)}} \tag{1.10}$$

Now

$$Var(X_1) = E[X_1 - E(X_1)]^2 = \sigma_{11}$$

$$Var\left(\hat{X}_1\right) = E\left[\hat{X}_1 - E\left(\hat{X}_1\right)\right]\left[\hat{X}_1 - E\left(\hat{X}_1\right)\right]'$$

$$= E\left[\mu_1 + \sigma_{12}'\Sigma_{22}^{-1}\left(X^{(2)} - \mu^{(2)}\right) - \mu_1\right]\left[\mu_1 + \sigma_{12}'\Sigma_{22}^{-1}\left(X^{(2)} - \mu^{(2)}\right) - \mu_1\right]'$$

$$= \sigma_{12}'\Sigma_{22}^{-1}\, E\left[\sigma_{12}'\Sigma_{22}^{-1}\left(X^{(2)} - \mu^{(2)}\right)\left(X^{(2)} - \mu^{(2)}\right)'\right]\Sigma_{22}^{-1}\, \sigma_{12}$$

$$\Rightarrow\ Var\left(\hat{X}_1\right) = \sigma_{12}'\Sigma_{22}^{-1}\, \sigma_{12}$$

$$Cov\left(X_1, \hat{X}_1\right) = E[X_1 - E(X_1)]\left[\hat{X}_1 - E\left(\hat{X}_1\right)\right]'$$

$$= E[X_1 - \mu_1]\left[\mu_1 + \sigma_{12}\Sigma_{22}^{-1}\left(X^{(2)} - \mu^{(2)}\right) - \mu_1\right]'$$

$$= E\left[(X_1 - \mu_1)\left(X^{(2)} - \mu^{(2)}\right)'\right]\Sigma_{22}^{-1}\sigma_{21}$$

$$\Rightarrow\ Cov\left(X_1, \hat{X}_1\right) = \sigma_{12}'\Sigma_{22}^{-1}\, \sigma_{12}$$

Putting this value in (1.10), we get

$$\rho_{1.(2,3,\dots,p)}$$

$$= \frac{\sigma_{12}'\Sigma_{22}^{-1}\, \sigma_{12}}{\sqrt{\sigma_{11}(\, \sigma_{12}'\Sigma_{22}^{-1}\, \sigma_{12})}}$$

$$= \sqrt{\frac{\sigma_{12}' \Sigma_{22}^{-1} \sigma_{12}}{\sigma_{11}}}$$

$$= \sqrt{\frac{\beta' \Sigma_{22} \beta}{\sigma_{11}}}$$

1.5.3 Partial Correlation Coefficient

If X_1 and X_2 are considered in conjunction with $(p-2)$ other variables $X_3, X_4, \cdots, X_p$, we may regard the variation of X_1 and X_2 as to certain extents due to the variation of the other variables. Let $X_{1.3,\cdots p}$ and $X_{2.3,\cdots p}$ represent these parts of variation of X_1 and X_2 respectively, which remains after subtraction of the best linear estimate in terms of $X_3, X_4, \cdots, X_p$. Thus, the correlation coefficient between $X_{1.3,\cdots p}$ and $X_{2.3,\cdots p}$ as a measure of correlation between X_1 and X_2 after removal of any part of the variation due to the influence of $X_3, X_4, \cdots, X_p$, is called partial correlation of X_1 and X_2 with respect to $X_3, X_4, \cdots, X_p$. It is denoted by

$$\rho_{12.(3,\ldots,p)} = \frac{Cov\left(X_{1.3,\cdots p}, X_{2.3,\cdots p}\right)}{\sqrt{Var\left(X_{1.3,\cdots p}\right) Var\left(X_{2.3,\cdots p}\right)}} \tag{1.11}$$

Let

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X^{(3)} \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \underline{\sigma}_{13}' \\ \sigma_{21} & \sigma_{22} & \underline{\sigma}_{23}' \\ \underline{\sigma}_{31} & \underline{\sigma}_{32} & \Sigma_{33} \end{pmatrix}$$

Without loss of generality, we assume that $\mu = 0$.

The best linear estimates of X_1 and X_2 in terms of $X^{(3)}$ are $\hat{X}_1 = \underline{\sigma}_{13}' \Sigma_{33}^{-1} X^{(3)}$ and $\hat{X}_2 = \underline{\sigma}_{23}' \Sigma_{33}^{-1} X^{(3)}$ respectively.

Define,

$X_{1.3,\cdots p} = X_1 - \hat{X}_1$ and $X_{2.3,\cdots p} = X_2 - \hat{X}_2$, then

$$Var(X_{1.3,\cdots p}) = E\left[X_1 - \hat{X}_1\right]\left[X_1 - \hat{X}_1\right]'$$

$$= E\left[X_1 - \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, X^{(3)}\right]\left[X_1 - \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, X^{(3)}\right]'$$

$$= E\left[X_1^2 - 2\underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, X_1 \, X^{(3)} + \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, X^{(3)} X^{(3)'} \, \Sigma_{33}^{-1} \underline{\sigma}_{13} \right]$$

$$= E(X_1^2) - 2\underline{\sigma}_{13}' \, \Sigma_{33}^{-1} E\left[X_1 \, X^{(3)}\right] + \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} E\left[X^{(3)} X^{(3)'}\right] \Sigma_{33}^{-1} \underline{\sigma}_{13}$$

$$= \sigma_{11} - 2\underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, \underline{\sigma}_{13} + \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, \Sigma_{33} \Sigma_{33}^{-1} \underline{\sigma}_{13}$$

$$\Rightarrow \; Var(X_{1.3,\cdots p}) = \sigma_{11} - \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, \underline{\sigma}_{13}$$

$$Var(X_{2.3,\cdots p}) = E\left[X_2 - \hat{X}_2\right]\left[X_2 - \hat{X}_2\right]'$$

$$= E\left[X_2 - \underline{\sigma}_{23}' \, \Sigma_{33}^{-1} \, X^{(3)}\right]\left[X_2 - \underline{\sigma}_{23}' \, \Sigma_{33}^{-1} \, X^{(3)}\right]'$$

$$= E\left[X_2^2 - 2\underline{\sigma}_{23}' \, \Sigma_{33}^{-1} \, X_1 \, X^{(3)} + \underline{\sigma}_{23}' \, \Sigma_{33}^{-1} \, X^{(3)} X^{(3)'} \, \Sigma_{33}^{-1} \underline{\sigma}_{23} \right]$$

$$= E(X_2^2) - 2\underline{\sigma}_{23}' \, \Sigma_{33}^{-1} E\left[X_1 \, X^{(3)}\right] + \underline{\sigma}_{23}' \, \Sigma_{33}^{-1} E\left[X^{(3)} X^{(3)'}\right] \Sigma_{33}^{-1} \underline{\sigma}_{23}$$

$$= \sigma_{22} - 2\underline{\sigma}_{23}' \, \Sigma_{33}^{-1} \, \underline{\sigma}_{23} + \underline{\sigma}_{23}' \, \Sigma_{33}^{-1} \, \Sigma_{33} \Sigma_{33}^{-1} \underline{\sigma}_{23}$$

$$\Rightarrow \; Var(X_{2.3,\cdots p}) = \sigma_{22} - \underline{\sigma}_{23}' \, \Sigma_{33}^{-1} \, \underline{\sigma}_{23}$$

$$Cov\left(X_{1.3,\cdots p}, X_{2.3,\cdots p}\right)$$

$$= E\left[X_1 - \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, X^{(3)}\right]\left[X_2 - \underline{\sigma}_{23}' \, \Sigma_{33}^{-1} \, X^{(3)}\right]'$$

$$= E\left[X_1 X_2' - \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, X^{(3)} X_2' - \; \underline{\sigma}_{23} \, \Sigma_{33}^{-1} \, X^{(3)'} X_1 + \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, X^{(3)} X^{(3)'} \Sigma_{33}^{-1} \, \underline{\sigma}_{23}\right]$$

$$= E(X_1 X_2') - \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, E\left[X^{(3)} X_2'\right] - \; \underline{\sigma}_{23} \, \Sigma_{33}^{-1} \, E\left[X^{(3)'} X_1\right] + \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, E\left[X^{(3)} X^{(3)'}\right] \Sigma_{33}^{-1} \, \underline{\sigma}_{23}$$

$$= \sigma_{12} - \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \; \underline{\sigma}_{23} - \; \underline{\sigma}_{23} \, \Sigma_{33}^{-1} \, \underline{\sigma}_{13}' + \underline{\sigma}_{13}' \, \Sigma_{33}^{-1} \, \Sigma_{33} \, \Sigma_{33}^{-1} \, \underline{\sigma}_{23}$$

$\Rightarrow Cov\left(X_{1.3,\cdots p}, X_{2.3,\cdots p}\right) = \sigma_{12} - \underline{\sigma}'_{13}\, \Sigma_{33}^{-1}\ \underline{\sigma}_{23}$

Putting these values in (1.11), we get

$$\rho_{12.(3,\ldots,p)} = \frac{\sigma_{12} - \underline{\sigma}'_{13}\, \Sigma_{33}^{-1}\ \underline{\sigma}_{23}}{\sqrt{\left(\sigma_{11} - \underline{\sigma}'_{13}\, \Sigma_{33}^{-1}\, \underline{\sigma}_{13}\right)\left(\sigma_{22} - \underline{\sigma}'_{23}\, \Sigma_{33}^{-1}\, \underline{\sigma}_{23}\right)}}.$$

Theorem 1.5.1: Let $X \sim \mathrm{N_p}(x|\mu, \Sigma)$ and $Y = AX$ where A is any $m \times p$ matrix of rank m($\leq p$). Then the distribution of Y is $N_m(Y|A\mu, A\Sigma A'')$.

Proof: Let C be a non-singular matrix such that

$C'\Sigma^{-1}C = I$

or, $\Sigma = CC'$

Then

$A\Sigma A' = (AC)(AC)'$

Since the post multiplication by a non-singular matrix does not alter the rank

$rank\ (A\Sigma A')$

$= rank(AC)$

$= rank\ (A)$

$= m$

Thus $A\Sigma A'$ is a positive definite matrix.

Now characteristic function of $Y = AX$ is given by

$E\left(e^{it'Y}\right) = E\left[e^{it'AX}\right]$

$= E\left[e^{iu'X}\right]; \qquad t = (t_1, t_2, \ldots, t_m)'$

$$= exp\left(iu'\underline{\mu} - \frac{1}{2}u'\Sigma\, u\right) \qquad (where\ u = A't)$$

$$= \exp\left(it'A\underline{\mu} - \frac{1}{2}t'A\Sigma A't\right)$$

which is the characteristic function of $N_m(Y|A\mu, A\Sigma A'')$.

Theorem 1.5.2: If $X \sim N_p(\mu, \Sigma)$, then $Z = DX$ and $Z \sim N_q(D\mu, D'\Sigma\mathrm{D})$, where D is $q \times p$ matrix of rank $q, q \leq p$.

Proof: Consider, the transformation, $Z = DX$. Here, Z has q-components and D is $q \times p$ real matrix. The expected value of Z is,

$E(Z) = E(DX) = DE(X) = D\mu$

The variance-covariance matrix is

$E[(Z - D\mu)(Z - D\mu)'] = D'\Sigma\mathrm{D}$

If $q = p$ and D is non-singular has been prove.

If $q < p$ and D is a $q \times p$ matrix of rank q, then we can find a $(p - q) \times p$ matrix E. Such that

$$\begin{bmatrix} Z \\ W \end{bmatrix} = \begin{bmatrix} D \\ E \end{bmatrix} X$$

is a non-singular transformation thus Z and W have a joint normal distribution and Z hs marginal distribution $Z = DX$, i.e. $Z \sim N_q(D\mu,\ D'\Sigma\mathrm{D})$.

Theorem1.5.3: If $X \sim N_p(x|\mu, \Sigma)$, a necessary and sufficient condition that one subset of X and the subset consisting of the remaining variables be independent is that each covariance of a variable from one set and a variable from the other set be 0.

Proof: Without loss of generality, we assume that the first set consists of first q variables $X_1, \dots, X_q$ and the other set consists of remaining $(p - q)$ variables $X_{q+1}, \dots, X_p$. Let

$X^{(1)} = (X_1, \dots, X_q)', X^{(2)} = (X_{q+1}, \dots, X_p)'$

so that, $X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}$

$E(X) = \mu = \begin{bmatrix} \mu^{(1)} \\ \mu^{(2)} \end{bmatrix}$

$E(X-\mu)(X-\mu)' = \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$

where $E\big(X^{(1)} - \mu^{(1)}\big)\big(X^{(1)} - \mu^{(1)}\big)' = \Sigma_{11}$

$E\big(X^{(1)} - \mu^{(1)}\big)\big(X^{(2)} - \mu^{(2)}\big)' = \Sigma_{12}$

$E\big(X^{(2)} - \mu^{(2)}\big)\big(X^{(1)} - \mu^{(1)}\big)' = \Sigma_{21} = \Sigma_{12}'$

$E\big(X^{(2)} - \mu^{(2)}\big)\big(X^{(2)} - \mu^{(2)}\big)' = \Sigma_{22}$

Necessary: Let two sets be independent so that

$f\big(X_1, \dots, X_p\big) = f\big(X_1, \dots, X_q\big) f\big(X_{q+1}, \dots, X_p\big)$

where $f\big(X_1, \dots, X_q\big)$ is the Marginal pdf of $X_1, \dots, X_q$

$f\big(X_{q+1}, \dots, X_p\big)$ is the Marginal pdf of $X_{q+1}, \dots, X_p$

Therefore, for $1 \le i \le q, q+1 \le j \le p$

$$\sigma_{ij} = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (X_i - \mu_i)\big(X_j - \mu_j\big) f\big(X_1, \dots, X_p\big) dX_1 \cdots dX_p$$

$$= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} (X_i - \mu_i) f\big(X_1, \dots, X_q\big) dX_1 \dots dX_p$$

$$= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \big(X_j - \mu_j\big) f\big(X_{q+1}, \dots, X_p\big) dX_{q+1} \cdots Xx_p$$

$= 0$

Since $\sigma_{ij} = \rho_{ij}\sqrt{\sigma_{ii}\sigma_{jj}}$ and by the assumption that Σ is non-singular, $\sigma_{ii} \neq 0, \sigma_{jj} \neq 0$ the condition that $\sigma_{ij} = 0 \Rightarrow \rho_{ij} = 0$, i.e., one set of variates is uncorrelated with the remaining variates.

Suppose the two sets are uncorrelated, i.e., $\Sigma_{12} = 0, \Sigma_{21} = \Sigma_{12}' = 0$. then

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

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix}$$

$$Q = (X - \mu)'\Sigma^{-1}(X - \mu)$$

$$= \begin{bmatrix} \left(X^{(1)} - \mu^{(1)}\right)' & \left(X^{(2)} - \mu^{(2)}\right)' \end{bmatrix} \begin{bmatrix} \Sigma_{11}^{-1} & 0 \\ 0 & \Sigma_{22}^{-1} \end{bmatrix} \begin{bmatrix} \left(X^{(1)} - \mu^{(1)}\right) \\ \left(X^{(2)} - \mu^{(2)}\right) \end{bmatrix}$$

$$= Q_1 + Q_2$$

where

$$Q_1 = \left(X^{(1)} - \mu^{(1)}\right)'\Sigma_{11}^{-1}\left(X^{(1)} - \mu^{(1)}\right)$$

$$Q_2 = \left(X^{(2)} - \mu^{(2)}\right)'{\Sigma_{22}}^{-1}\left(X^{(2)} - \mu^{(2)}\right)$$

$$|\Sigma| = |\Sigma_{11}||\Sigma_{22}|$$

Therefore, pdf of X can be written as

$$f(X)$$

$$= \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|} exp\left(-\frac{1}{2}Q\right)$$

$$= \frac{1}{(2\pi)^{\frac{q}{2}}|\Sigma_{11}|} exp\left(-\frac{1}{2}Q_1\right) \times \frac{1}{(2\pi)^{\frac{(p-q)}{2}}|\Sigma_{22}|} exp\left(-\frac{1}{2}Q_2\right)$$

$= f\{X^{(1)}\} f\{X^{(2)}\}$

Now, marginal distribution of $X^{(1)}$ is

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(X)\, dX_{q+1\dots} dX_p = f\{X^{(1)}\} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f\{X^{(2)}\}\, dX_{q+1\dots} dX_p$$

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(X)\, dX_{q+1\dots} dX_p = f\{X^{(1)}\}$$

Similarly marginal distribution of $X^{(2)}$ is $f\{X^{(2)}\}$.

Thus, the joint distribution of X is the product of marginal distribution of $X^{(1)}$ and $X^{(2)}$. Therefore, the two sets of random variables are independently distributed.

$X^{(1)} \sim N_q(X^{(1)} | \mu^{(1)}, \Sigma_{11})$

$X \sim N_q\, X^{(2)} | \mu^{(2)}, \Sigma_{22})$

1.5 Marginal Distribution

The Marginal distribution of X is

$F(x_1, x_2, \dots, x_r)$

$$= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_r} \left[\int_{x_{(r+1)}=-\infty}^{\infty} \cdots \int_{x_p=-\infty}^{\infty} f(x_1, x_2, \dots, x_p) dx_{r+1} \dots dx_p \right] dx_1 \dots dx_r$$

$= P(X_1 \le x_1, X_2 \le x_2, \dots, X_r \le x_r, X_{r+1} \le \infty \dots < p \le \infty)$

$= F(x_1, x_2, \dots, x_r, \infty, \dots, \infty)$

Differentiating partially, we will get pdf. If we integrate pdf of $(x_1, x_2, \dots, x_p)$ for whole range then we will get equal to one.

Theorem 1.6.1: If $X \sim N_p(x|\mu, \Sigma)$, the marginal distribution of any set of components of X is multivariate normal with mean vector and variance-covariance matrix obtained by taking the proper components of μ and Σ respectively.

Proof: Let

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}_{p-q}^{q}$$

Since the numbering of the components of X is arbitrary, without loss of generality we assume that we must obtain the marginal distribution of last $(p-q)$ components of X, i.e., marginal distribution of $X^{(2)}$. Consider the non-singular linear transformation to sub vectors

$$y^{(1)} = x^{(1)} + Mx^{(2)}$$

$$y^{(2)} = x^{(2)}$$

Matrix M is chosen so that components of $y^{(1)}$ are uncorrelated with the components of $y^{(2)} = x^{(2)}$, i.e.,

$$0 = E\left[y^{(1)} - E\left(y^{(1)}\right)\right]\left[y^{(2)} - E\left(y^{(2)}\right)\right]$$

$$= E\left[x^{(1)} + Mx^{(2)} - \mu^{(1)} - M\mu^{(2)}\right]\left[x^{(2)} - \mu^{(2)}\right]'$$

$$= \Sigma_{12} + M\Sigma_{22}$$

Thus $M = -\Sigma_{12}\Sigma_{22}^{-1}$ and

$$y^{(1)} = x^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}x^{(2)}$$

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix} = y = \begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix} x$$

The vector Y is a non-singular transform of X. Therefore Y follows a normal distribution with mean vector

$$E\begin{bmatrix} y^{(1)} \\ y^{(2)} \end{bmatrix}$$

$$= E\begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{bmatrix}\begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix}$$

$$= \begin{bmatrix} \mu^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\mu^{(2)} \\ \mu^{(2)} \end{bmatrix}$$

$$= \begin{bmatrix} \nu^{(1)} \\ \nu^{(2)} \end{bmatrix} = \nu \ \ (say)$$

and variance covariance matrix

$$E[Y-\nu][Y-\nu]' = \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix}$$

Since

$$E\left[Y^{(1)} - \nu^{(1)}\right]\left[Y^{(1)} - \nu^{(1)}\right]'$$

$$= E\left[\left(x^{(1)} - \mu^{(1)}\right) - \Sigma_{12}\Sigma_{22}^{-1}\left(x^{(2)} - \mu^{(2)}\right)\right]\left[\left(x^{(1)} - \mu^{(1)}\right) - \Sigma_{12}\Sigma_{22}^{-1}\left(x^{(2)} - \mu^{(2)}\right)\right]'$$

$$= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

Thus $Y^{(1)}$ and $Y^{(2)}$ are independent and hence the marginal distribution of

$$X^{(2)} = Y^{(2)} \sim N_{p-q}\left(x^{(2)} | \mu^{(2)}, \Sigma_{22}\right)$$

Example: Let $Y = (y_1, y_2, y_3)' \sim N_3(\mu, \Sigma)$, where

$$\mu = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \Sigma = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix}.$$

(i) Find the marginal of $y_1, y_2\ and\ y_3$.

(ii) Find the marginal of $Z_1 = \begin{pmatrix} y_1 \\ y_3 \end{pmatrix}$.

Solution:

(i) $f(y_1) = \frac{1}{(2\pi)^{\frac{1}{2}}|4|^{\frac{1}{2}}}\ exp\left\{-\frac{1}{2\times4}(y_1-1)^2\right\} = \frac{1}{2(2\pi)^{\frac{1}{2}}}\ exp\left\{-\frac{1}{8}(y_1-1)^2\right\}$

$$f(y_2) = \frac{1}{2(2\pi)^{\frac{1}{2}}}\ exp\left\{-\frac{1}{8}(y_2+1)^2\right\}$$

$$f(y_3) = \frac{1}{2(2\pi)^{\frac{1}{2}}}\ exp\left\{-\frac{1}{8}(y_3)^2\right\}$$

$y_1 \sim N(1,4), y_2 \sim N(-1,4), y_2 \sim N(0,4)$

(ii) $f\begin{bmatrix} y_1 \\ y_3 \end{bmatrix} = \frac{1}{(2\pi)|\Sigma|^{\frac{1}{2}}}\ exp\left\{-\frac{1}{2}(x-\mu)'\ \Sigma^{-1}(x-\mu)\right\}$

$$\mu = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}, \Sigma^{-1} = \begin{pmatrix} 4 & -2 \\ -2 & 4 \end{pmatrix} = \frac{1}{12}$$

$$\begin{bmatrix} y_1 \\ y_3 \end{bmatrix} \sim N_2\left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \begin{pmatrix} 4 & 2 \\ 2 & 4 \end{pmatrix}\right]$$

1.6 Conditional Distribution

Let

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}_{p-q}^{q}$$

The conditional distribution $X^{(1)}|X^{(2)}$ is

$$f\{X^{(1)}|X^{(2)}\} = \frac{f\{X^{(1)}, X^{(2)}\}}{f\{X^{(2)}\}}$$

Now, consider the transformation,

$y^{(1)} = x^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}x^{(2)}$

$= x^{(1)} + Mx^{(2)}$

$y^{(2)} = x^{(2)}$

The joint distribution of $X^{(1)}, X^{(2)}$ is

$$f\{x^{(1)}, x^{(2)}\} = f(x)$$

$$= \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left[-\frac{1}{2}\{(x-\mu)'\Sigma^{-1}(x-\mu)\}\right] \quad (1.12)$$

Now, consider the transformation and the joint pdf of $Y^{(1)}$ and $Y^{(2)}$ is

$$g\left(y^{(1)}, y^{(2)}\right) = \frac{1}{(2\pi)^{\frac{q}{2}}|\Sigma_{11.2}|^{\frac{1}{2}}} exp\left[-\frac{1}{2}\left(y^{(1)} - \nu^{(1)}\right)'\Sigma_{11.2}^{-1}\left(y^{(1)} - \nu^{(1)}\right)\right]$$

$$\times \frac{1}{(2\pi)^{\frac{(p-q)}{2}}|\Sigma_{22}|^{\frac{1}{2}}} exp\left[-\frac{1}{2}\left(y^{(2)} - \nu^{(2)}\right)'\Sigma_{22}^{-1}\left(y^{(2)} - \nu^{(2)}\right)\right] \times 1$$

$g\left(y^{(1)}, y^{(2)}\right)$

$$= \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_{11.2}|^{\frac{1}{2}}|\Sigma_{22}|^{\frac{1}{2}}} exp\left[-\frac{1}{2}\left\{\left(y^{(1)} - \nu^{(1)}\right)'\Sigma_{11.2}^{-1}\left(y^{(1)} - \nu^{(1)}\right)\right.\right.$$

$$\left.\left. + \left(y^{(2)} - \nu^{(2)}\right)'\Sigma_{22}^{-1}\left(y^{(2)} - \nu^{(2)}\right)\right\}\right]$$

$$= \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_{11.2}|^{\frac{1}{2}}|\Sigma_{22}|^{\frac{1}{2}}} exp\left[-\frac{1}{2}\left(y^{(1)} - \nu^{(1)}\right)'\Sigma_{11.2}^{-1}\left(y^{(1)} - \nu^{(1)}\right)\right]$$

$$exp\left[-\frac{1}{2}\left(y^{(2)} - \nu^{(2)}\right)'\Sigma_{22}^{-1}\left(y^{(2)} - \nu^{(2)}\right)\right]$$

Hence the joint pdf of $X^{(1)}, X^{(2)}$ is

$$f\left(x^{(1)}, x^{(2)}\right) = g\left(y^{(1)}, y^{(2)}\right)|J|$$

$$= \frac{1}{(2\pi)^{p/2}|\Sigma_{11.2}|^{1/2}|\Sigma_{22}|^{1/2}} \; exp\left[-\frac{1}{2}\left(y^{(1)} - \nu^{(1)}\right)'\Sigma_{11.2}^{-1}\left(y^{(1)} - \nu^{(1)}\right)\right]$$

$$\times exp\left[-\frac{1}{2}\left(y^{(2)} - \nu^{(2)}\right)'\Sigma_{22}^{-1}\left(y^{(2)} - \nu^{(2)}\right)\}\right] \tag{1.13}$$

The marginal density of $X^{(2)}$ is

$$f\left(x^{(2)}\right) = \frac{1}{(2\pi)^{(p-q)/2}|\Sigma_{22}|^{1/2}} \times exp\left[-\frac{1}{2}\left(y^{(2)} - \nu^{(2)}\right)'\Sigma_{22}^{-1}\left(y^{(2)} - \nu^{(2)}\right)\right] \tag{1.14}$$

From (1.13) and (1.14), we get the conditional distribution,

$$f\left(x^{(1)}|x^{(2)}\right) = \frac{1}{(2\pi)^{q/2}|\Sigma_{11.2}|^{1/2}} exp\left[-\frac{1}{2}\left\{\left(y^{(1)} - \nu^{(1)}\right)'\Sigma_{11.2}^{-1}\left(y^{(1)} - \nu^{(1)}\right)\right\}\right] \tag{1.15}$$

Now consider,

$$\left(y^{(1)} - \nu^{(1)}\right)'\Sigma_{11.2}^{-1}\left(y^{(1)} - \nu^{(1)}\right)$$

$$= \left\{x^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}x^{(2)} - \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\mu^{(2)}\right\}'\Sigma_{11.2}^{-1}\left\{X^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}x^{(2)} - \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\mu^{(2)}\right\}$$

$$= \left\{\left(x^{(1)} - \mu^{(1)}\right) - \Sigma_{12}\Sigma_{22}^{-1}\left(x^{(2)} - \mu^{(2)}\right)\right\}'\Sigma_{11.2}^{-1}\left\{\left(x^{(1)} - \mu^{(1)}\right) - \Sigma_{12}\Sigma_{22}^{-1}\left(x^{(2)} - \mu^{(2)}\right)\right\}$$

Putting these values in (1.15), we get

$$f\left(x^{(1)}|x^{(2)}\right)$$

$$= \frac{1}{(2\pi)^{q/2}|\Sigma_{11.2}|^{1/2}} \times$$

$$exp\left[-\frac{1}{2}\left\{x^{(1)} - \mu^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\left(x^{(2)} - \mu^{(2)}\right)\right\}'\Sigma_{11.2}^{-1}\left\{x^{(1)} - \mu^{(1)} - \Sigma_{12}\Sigma_{22}^{-1}\left(x^{(2)} - \mu^{(2)}\right)\right\}\right]$$

This is the pdf of a multivariate normal distribution with mean vector

$$E\left[X^{(1)}|X^{(2)}\right] = \mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left(X^{(2)} - \mu^{(2)}\right)$$

The conditional variance covariance matrix is $Cov\left(X^{(1)}\middle|X^{(2)}\right) = \Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Example: Let $X \sim N_4(\mu, \Sigma)$, where

$$\mu = \begin{bmatrix} 5 \\ 6 \\ 7 \\ 8 \end{bmatrix}, \Sigma = \begin{bmatrix} 2 & 0 & 1 & 0 \\ 0 & 3 & 2 & 0 \\ 1 & 2 & 4 & 0 \\ 0 & 0 & 0 & 9 \end{bmatrix}$$

(i) Find the distribution of $\begin{pmatrix} X_2 \\ X_4 \end{pmatrix}$.

(ii) Find the distribution of $(X_1 - X_4)$.

(iii) Find the conditional distribution of $(X_1, X_2)|X_3$.

Solution:

(i) Let $C = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$, then $\begin{pmatrix} X_2 \\ X_4 \end{pmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{bmatrix}$, so $C\mu = \begin{pmatrix} 6 \\ 8 \end{pmatrix}, C\Sigma C' = \begin{pmatrix} 3 & 0 \\ 0 & 9 \end{pmatrix}$.

Thus

$$\begin{pmatrix} X_2 \\ X_4 \end{pmatrix} \sim N_2\left[\begin{pmatrix} 6 \\ 8 \end{pmatrix}, \begin{pmatrix} 3 & 0 \\ 0 & 9 \end{pmatrix}\right]$$

(ii) Let $C = (1 \quad 0 \quad 0 \quad -1)$, then $X_1 - X_4 = CX$, So $C\mu = 5 - 8 = -3, C\Sigma C' = 11$. Thus

$(X_1 - X_4) \sim N(-3,11)$.

(iii) Let $X_1 = (X_1, X_2)'$ and $X_3 = X_3$, then$\{(X_1, X_2)'|X_3 = x_3\}, then\ \mu_1 = \begin{pmatrix} 5 \\ 6 \end{pmatrix}, \mu_2 = 7,$

$\Sigma_{12} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \Sigma_{22} = 4, \Sigma_{11} = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$. Mean of $\{(X_1, X_2)'|X_3 = x_3\}$ is

$$\mu^{(1)} + \Sigma_{12}\Sigma_{22}^{-1}\left(X^{(2)} - \mu^{(2)}\right)$$

$$= \binom{5}{6} + \binom{1}{2}\frac{1}{4}(x_3 - 7)$$

$$= \begin{pmatrix} \frac{1}{4}x_3 + \frac{13}{4} \\ \frac{1}{2}x_3 + \frac{5}{2} \end{pmatrix}$$

The covariance matrix of $\{(x_1, x_2)' | X_3 = x_3\}$ is

$$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}{\Sigma_{22}}^{-1}\Sigma_{12}' = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} - \binom{1}{2}\frac{1}{4}(1 \quad 2) = \frac{1}{4}\begin{pmatrix} 7 & -2 \\ -2 & 8 \end{pmatrix}$$

1.7 Summary

The multivariate normal distribution is an extension of the univariate normal distribution to higher dimensions. It is characterized by its mean vector and covariance matrix. The moment generating function and characteristic function provide alternative representations of the distribution, and the properties of marginal and conditional distributions allow for the study of subsets and dependencies within the multivariate distribution.

This unit covers the basic concepts of multivariate normal distribution. The procedure of finding the moment generating function and characteristic function is discussed in detail. Also, the marginal distribution and conditional distribution are derived and their properties are studied.

1.8 Self-Assessment Exercises

1. Find μ mean vector and Σ variance covariance matrix or dispersion matrix of the following density functions:

   (i) $f(x, y) = \frac{1}{2\pi} exp\left[-\frac{1}{2}\{x^2 + y^2 + 4x - 6y + 13\}\right]$

   (ii) $f(x, y) = \frac{1}{2\pi} exp\left\{-\frac{1}{2}(2x^2 + y^2 + 2xy - 22x - 14y + 65)\right\}$

2. Let $f(x) = C \exp\left(-\frac{Q}{2}\right)$, where

(i) $Q = 3x^2 + 2y^2 - 2xy - 32x + 4y + 92$, and

(ii) $Q = 2x_1^2 + 3x_2^2 + 4x_3^2 + 2x_1x_2 - 2x_1x_3 - 4x_2x_3 - 6x_1 - 2x_2 + 10x_3 + 8$

Find μ and Σ.

3. Derive the characteristic function of a multivariate normal distribution. Using the characteristic function, obtain the mean vector.

4. Let X be partitioned as $X = \left(X^{(1)\prime}, X^{(2)\prime}\right)'$. Derive the marginal distribution of $X^{(1)}$ and conditional distribution of $X^{(2)}$ given $X^{(1)}$.

5. Let $X = (X_1, X_2, X_3)'$ follows a 3-variate normal distribution with mean vector 0 and variance-covariance matrix

$$\begin{bmatrix} 1 & \frac{1}{3} & 0 \\ \frac{1}{3} & 1 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{2} \end{bmatrix}$$

Find

(i) The marginal distribution of $(X_1, X_2)'$

(ii) The conditional distribution of $(X_1, X_2)'$ given X_3

(iii) $E(X_3|X_1, X_2)$

(iv) $E[(X_1 + X_2 - X_3)X_3]$

(v) Obtain the correlation coefficient between X_2 and X_3.

6. Prove that $X^{(1)}$ and $X^{(2)}$ are independently distributed if and only if $\Sigma_{12} = 0$.

7. Obtain the mean vector and variance covariance matrix of the random variable $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ with pdf $f(X) = \frac{1}{2\pi} exp(2X_1^2 + X_1^2 + 2X_1X_2 - 22X_1 - 14X_2 + 65)$.

8. Suppose $y \sim N_4(\mu, \Sigma)$, where

$$\mu = \begin{bmatrix} -2 \\ 3 \\ -1 \\ 5 \end{bmatrix}, \Sigma = \begin{bmatrix} 11 & -8 & 3 & 9 \\ -8 & 9 & -3 & 6 \\ 3 & -3 & 2 & 3 \\ 9 & 6 & 3 & 9 \end{bmatrix}.$$

(i) Find the distribution of $Z = 4y_1 - 2y_2 + y_3 - 3y_4$.

(ii) Find the joint distribution of $Z_1 = y_1 + y_2 + y_3 + y_4$ and $Z_2 = -2y_1 + 3y_2 + y_3 - 2y_4$.

(iii) Find the joint distribution of $Z_1 = 3y_1 + y_2 - 4y_3 - y_4, Z_2 = -y_1 - 3y_2 + y_3 - 2y_4$ and $Z_3 = 2y_1 + 2y_2 + 4y_3 - 5y_4$.

(iv) What is the distribution of y_3.

(v) What is the joint distribution of y_2 and y_4.

(vi) Find the joint distribution of $y_1, \frac{1}{2}(y_1 + y_2), \frac{1}{3}(y_1 + y_2 + y_3)$ and $\frac{1}{4}(y_1 + y_2 + y_3 + y_4)$.

9. Suppose $y \sim N_3(\mu, \Sigma)$, where

(i) Find the distribution of $Z = 2y_1 - y_2 + 3y_3$.

(ii) Find the joint distribution of $Z_1 = y_1 + y_2 + y_3$ and $Z_2 = y_1 - y_2 + 2y_3$.

(iii) Find the distribution of y_2.

(iv) Find the joint distribution of y_1 and y_3.

(v) Find the joint distribution of y_1, y_3 and $\frac{1}{2}(y_1 + y_2)$

10. If $\underline{X} \sim N_p(\underline{\mu}, \Sigma)$. Show that $M_{\underline{X}-\underline{\mu}}(\underline{t}) = \exp\left(\frac{1}{2}\underline{t}'\Sigma\,\underline{t}\right)$.

11. Prove that

(i) $E(x_j - \mu_j)(x_k - \mu_k) = \sigma_{jk} \qquad \forall j = 1,2, \ldots\ldots, p;\ \ k = 1,2, \ldots\ldots, p$

(ii) $E(x_j - \mu_j)(x_k - \mu_k)(x_l - \mu_l) = 0$

(iii) $E(x_j - \mu_j)(x_k - \mu_k)(x_l - \mu_l)(x_m - \mu_m) = \sigma_{jk}\sigma_{lm} + \sigma_{jl}\sigma_{km} + \sigma_{jm}\sigma_{kl}$

1.9 References

- Johnson, R. A., Wichern, D. W. (2019): Applied Multivariate Statistical Analysis. United Kingdom: Pearson.

- Muirhead, R. J. (2009): Aspects of Multivariate Statistical Theory. Germany: Wiley.
- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Giri Narayan C. (1995): Multivariate Statistical Analysis
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.
- Mardia, K. V., Bibby, J. M., Kent, J. T. (1979): Multivariate Analysis. United Kingdom: Academic Press.
- Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York.

1.10 Further Reading

- Kotz, S., Balakrishnan, N. and Johnson, N.L.: Continuous Multivariate Distribution Models and Applications (Second Edition). Volume 1, Wiley - Inter science, New York.
- Khatri, C. G.: Multivariate Analysis.
- Mardia, K. V.: Multivariate Analysis.
- Seber, G.A.F.: *Multivariate Observations.* Wiley, New York.
- Rencher, Alvin C.: Multivariate Statistical Inference and Applications. John Wiley. New York, New York.

UNIT:2 MLE OF PARAMETERS

Structure

2.1 Introduction

Maximum Likelihood Estimator is a method of 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 regression coefficients, covariance matrices, and mean vectors.

Here is a step-by-step explanation:

1. Specify the model: Define the multivariate model, such as a multivariate normal distribution or a linear regression model.

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 Expectation-Maximization algorithm or gradient-based methods.

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.

2.2 Objectives

After studying this unit, you should be able to:

- Describe the likelihood function and the role of maximum likelihood estimation in deriving the estimators of parameters of multivariate normal distribution.
- Derive the sufficient statistic for the multivariate normal distribution.

- Compute the sample multiple correlation coefficients, partial correlation coefficients and regression coefficient.

2.3 Estimation of Parameters in Multivariate Normal Distribution

The multivariate normal distribution is completely specified if its mean vector μ and dispersion matrix Σ. In case of unknown parameters, the problem of their estimation arises. We can estimate these parameters by the method of maximum likelihood estimation.

Let $x_1, x_2, \ldots, x_N$, be a random sample of size N from $N_p(\mu, \Sigma)$, where $N > p$ and x_α is $p \times 1$ vector, $\alpha = 1, 2, \cdots, N$.

Notation

Suppose observations on p characteristics $X_1, X_2, \cdots, X_p$ of N individuals $\alpha = 1, 2, \cdots, N$ are as given in the following table:

Characteristic	**Individuals**						**Mean**
	1	**2**	$\cdots$	$\boldsymbol{\alpha}$	$\cdots$	$\boldsymbol{N}$	
X_1	x_{11}	x_{12}	$\cdots$	$x_{1\alpha}$	$\cdots$	x_{1N}	$\bar{x}_1$
X_2	x_{21}	x_{22}	$\cdots$	$x_{2\alpha}$	$\cdots$	x_{2N}	$\bar{x}_2$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
X_i	x_{i1}	x_{i2}	$\cdots$	$x_{i\alpha}$	$\cdots$	x_{iN}	$\bar{x}_i$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
X_p	x_{p1}	x_{p2}	$\cdots$	$x_{p\alpha}$	$\cdots$	x_{pN}	$\bar{x}_p$
	x_1	x_2	$\cdots$	x_α	$\cdots$	x_N	

Therefore, the sample mean vector is

$$\bar{x} = \frac{1}{N}\sum_{\alpha=1}^{N} x_\alpha = \begin{pmatrix} \frac{1}{N}\sum_{\alpha=1}^{N} x_{1\alpha} \\ \vdots \\ \frac{1}{N}\sum_{\alpha=1}^{N} x_{i\alpha} \\ \vdots \\ \frac{1}{N}\sum_{\alpha=1}^{N} x_{p\alpha} \end{pmatrix} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_i \\ \vdots \\ \bar{x}_p \end{pmatrix}$$

The sample variance and covariance matrix is

$$\mathrm{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{pmatrix}$$

Where

$$s_{ij} = \frac{1}{N-1}\sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)\left(x_{j\alpha} - \bar{x}_j\right), \qquad \forall\, i \; and \; j$$

Also

$$\mathrm{S} = \frac{1}{N-1}\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} = \frac{A}{N-1}$$

The matrix A is called the sum of squares and cross products of deviations about the mean.

Here

$$a_{ij} = \sum_{\alpha=1}^{N} (x_{i\alpha} - \bar{x}_i)\left(x_{j\alpha} - \bar{x}_j\right), \qquad \forall\, i \; and \; j$$

Remark:

(i) For a quadratic form

$$Q = x' A x = \sum_{i,j=1}^{p} a_{ij} x_i x_j$$

We have

$$\frac{\partial Q}{\partial x} = \begin{pmatrix} \frac{\partial Q}{\partial x_1} \\ \frac{\partial Q}{\partial x_2} \\ \vdots \\ \frac{\partial Q}{\partial x_p} \end{pmatrix} = 2\,Ax$$

(ii) If $Q = (x-b)'A(x-b) = (b-x)'A(b-x)$, then

$$\frac{\partial Q}{\partial x} = 2\,A(x-b)$$

And

$$\frac{\partial Q}{\partial b} = 2\,A(b-x)$$

(iii) A submatrix of A is a rectangular array obtained from A by deleting rows and columns. A minor is the determinant of the square submatrix of A.

$$|A| = \sum_{i=1}^{p} a_{ij}\,A_{ij} = \sum_{j=1}^{p} a_{jk}\,A_{jk}$$

where A_{ij}, is $(-1)^{i+j}$ times the minor of a_{ij}, and the minor of an element a_{ij} is the determinant of the submatrix of a square matrix A obtained by deleting the i^{th} row and j^{th} column.

If $|A| \neq 0$, there exists a unique matrix B such that $AB = I, B$ is called the inverse of A and it is denoted by A^{-1}.

Let a_{ij} be the element of A^{-1} in the i^{th} row and j^{th} column, then

$$a^{ij} = \frac{A_{ji}}{|A|} \quad and \quad a_{ij} = \frac{A^{ji}}{|A^{-1}|}$$

Example 2.3.1: Let

$$A = \begin{bmatrix} 2 & 4 \\ 3 & 7 \end{bmatrix}$$

The Determinant of Matrix A is

$$|A| = 2 \times 7 - 3 \times 4 = 14 - 12 = 2$$

The inverse of the matrix A is

$$A^{-1} = \frac{adjA}{|A|}$$

the adjoint matrix is

$$adj\ A = \begin{bmatrix} 7 & -4 \\ -3 & 2 \end{bmatrix}$$

$$\therefore A^{-1} = \frac{\begin{bmatrix} 7 & -4 \\ -3 & 2 \end{bmatrix}}{2} = \begin{bmatrix} \frac{7}{2} & -\frac{4}{2} \\ -\frac{3}{2} & \frac{2}{2} \end{bmatrix} = \begin{bmatrix} \frac{7}{2} & -2 \\ -\frac{3}{2} & 1 \end{bmatrix}$$

$$= \frac{7}{2} - \frac{6}{2} = \frac{1}{2}$$

Also

$$a_{11} = \frac{A^{11}}{|A^{-1}|} = \frac{1}{1/2} = 2$$

$$a_{12} = \frac{A^{21}}{|A^{-1}|} = \frac{2}{1/2} = 4$$

$$a_{21} = \frac{A^{12}}{|A^{-1}|} = \frac{3/2}{1/2} = 3$$

$$a_{22} = \frac{A^{22}}{|A^{-1}|} = \frac{7/2}{1/2} = 7$$

(iv) A square matrix $A_{m\times m}$ is said to be orthogonal if $A'A = AA' = I$, and if the transformation $Y = A\,X$ transform $X'X$ to $Y'Y$. Also

$$\sum_{k=1}^{m} a_{ik} \times \frac{1}{\sqrt{n}} = 0$$

$$\Rightarrow \sum_{k=1}^{m} a_{ik} = 0 \quad , \quad i = 1,2,\cdots,(m-1)$$

and

$$\sum_{k=1}^{m} a_{ik} \times a_{jk} = \begin{cases} 0\,, & i \neq j \\ 1\,, & i = j \end{cases} \; for \quad i,j = 1,2,\cdots,m$$

Example 2.3.2: Consider an orthogonal matrix of order 3

$$A = \begin{bmatrix} \frac{1}{\sqrt{2\times 1}} & \frac{-1}{\sqrt{2\times 1}} & \frac{1\times 0}{\sqrt{2\times 1}} \\ \frac{1}{\sqrt{2\times 3}} & \frac{1}{\sqrt{2\times 3}} & \frac{-2}{\sqrt{2\times 3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

The transpose of the matrix A is

$$A^T = \begin{bmatrix} \frac{1}{\sqrt{2\times 1}} & \frac{1}{\sqrt{2\times 3}} & \frac{1}{\sqrt{3}} \\ \frac{-1}{\sqrt{2\times 1}} & \frac{1}{\sqrt{2\times 3}} & \frac{1}{\sqrt{3}} \\ \frac{1\times 0}{\sqrt{2\times 1}} & \frac{-2}{\sqrt{2\times 3}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

Now

$$AA^T = \begin{bmatrix} \frac{1}{\sqrt{2\times1}} & \frac{-1}{\sqrt{2\times1}} & \frac{1\times0}{\sqrt{2\times1}} \\ \frac{1}{\sqrt{2\times3}} & \frac{1}{\sqrt{2\times3}} & \frac{-2}{\sqrt{2\times3}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2\times1}} & \frac{1}{\sqrt{2\times3}} & \frac{1}{\sqrt{3}} \\ \frac{-1}{\sqrt{2\times1}} & \frac{1}{\sqrt{2\times3}} & \frac{1}{\sqrt{3}} \\ \frac{1\times0}{\sqrt{2\times1}} & \frac{-2}{\sqrt{2\times3}} & \frac{1}{\sqrt{3}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I$$

Here $m = 3$

$$\sum_{k=1}^{3} a_{ik} = a_{11} + a_{12} + a_{13}$$

$$= \frac{1}{\sqrt{2\times1}} - \frac{1}{\sqrt{2\times1}} + 0 = 0$$

$$\sum_{k=1}^{m} a_{ik} \times a_{jk} = a_{11}a_{11} + a_{12}a_{12} + a_{13}a_{13}$$

$$= \frac{1}{2} + \frac{1}{2} + 0 = 1$$

$$\sum_{k=1}^{m} a_{ik} \times a_{jk}$$

$$= a_{11}a_{21} + a_{12}a_{22} + a_{13}a_{23} = 0$$

$$\sum_{k=1}^{m} a_{ik} \times a_{jk}$$

$$= a_{11}a_{31} + a_{12}a_{32} + a_{13}a_{33} = 0$$

$$\sum_{k=1}^{m} a_{ik} \times a_{jk}$$

$= a_{21}a_{11} + a_{22}a_{12} + a_{23}a_{13} = 0$

$$\sum_{k=1}^{m} a_{ik} \times a_{jk}$$

$= a_{21}a_{21} + a_{22}a_{22} + a_{23}a_{23} = \frac{3}{3} = 1$

$$\sum_{k=1}^{m} a_{ik} \times a_{jk}$$

$= a_{21}a_{31} + a_{22}a_{32} + a_{23}a_{33} = 0$

$$\sum_{k=1}^{m} a_{ik} \times a_{jk}$$

$= a_{31}a_{31} + a_{32}a_{32} + a_{33}a_{33} = \frac{3}{3} = 1$

$$\sum_{k=1}^{m} a_{ik} \times a_{jk}$$

$= a_{31}a_{21} + a_{32}a_{22} + a_{33}a_{23} = 0$

$$\sum_{k=1}^{m} a_{ik} \times a_{jk}$$

$= a_{31}a_{11} + a_{32}a_{12} + a_{33}a_{13} = 0$

(v) **Trace:** The sum of diagonal elements of a matrix is called its trace. We have the following results for the trace of matrices:

1. $tr(A + B) = tr(A) + tr(B)$
2. $tr\ (AB) = tr(BA)$

3. The trace of a scalar quantity will be the same number. For example: $trace\ of\ [6]_{1\times1} = 6$

(vi) Differentiation Rules:

1. $\frac{\partial|X|}{\partial X} = |X|X^{-1}$

2. $\frac{\partial tr\ AX}{\partial X} = A$

3. $\frac{\partial(X'AX)}{\partial X} = 2AX$ or $\frac{\partial(X'AX)}{\partial X'} = 2X'A$

Lemma 2.3.1.: Let $f(\theta)$ be a real-valued function defined on a certain set S and let ϕ be a single-valued function, with a single-valued inverse on S to some other set S^*, i.e., to each $\theta \in S$ there corresponds a unique $\theta^* = \phi(\theta) \in S^*$ and conversely to each $\theta^* \in S^*$ there corresponds a unique $\theta = \phi^{-1}(\theta^*) \in S$. Let

$g(\theta^*) = f[\phi^{-1}(\theta^*)]$

Then if $f(\theta)$ attains a maximum at $\theta = \theta_0$, $g(\theta^*)$ attains a maximum at $\theta^* = \theta_0^* = \phi(\theta_0)$. If the maximum of $f(\theta)$ at θ_0 is unique so as the maximum of $g(\theta^*)$ at θ_0^*.

Proof: $f(\theta_0) \geq f(\theta)\ \forall\ \theta \in S$

Then $\forall \theta^* \in S^*$

$g(\theta^*) = f[\phi^{-1}(\theta^*)] = f(\theta) \leq f(\theta_0)$

$= g[\phi(\theta_0)] = g(\theta_0^*)$

Hence $g(\theta^*)$ attains a maximum at θ_0^*. If maximum of $f(\theta)$ at θ_0 is unique, there is a strict inequality above for $\theta \neq \theta_0$ and the maximum of $g(\theta^*)$ is unique.

We have the following corollary:

Corollary 2.3.1.: Let $\widehat{\theta_1}, \widehat{\theta_2}, \dots, \widehat{\theta_m}$ are the MLE of the parameters $\theta_1, \theta_2, \dots, \theta_m$ and the transformation from $\theta_1, \theta_2, \dots, \theta_m$ to $\phi_1, \dots, \phi_m$ is one to one. Then

$\phi_1(\widehat{\theta_1}, \widehat{\theta_2}, \dots, \widehat{\theta_m}), \dots, \phi_m(\widehat{\theta_1}, \widehat{\theta_2}, \dots, \widehat{\theta_m})$

Are, respectively, the MLE's of

$\phi_1(\theta_1, \theta_2, \dots, \theta_m), \dots, \phi_m(\theta_1, \theta_2, \dots, \theta_m)$.

If the estimators of $\theta_1, \theta_2, \dots, \theta_m$ are unique then the estimators of $\phi_1, \dots, \phi_m$ are also unique.

2.3.1 Maximum Likelihood Estimates of the Mean Vector when Variance-Covariance Matrix is Known

Let $x_1, x_2, \dots, x_N$, be a random sample of size N from $N_p(\mu, \Sigma)$. The likelihood function is

$$L(\mu, \Sigma \mid x_\alpha) = \frac{|\Sigma|^{\frac{-N}{2}}}{(2\pi)^{\frac{Np}{2}}} \exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N}\{(x_\alpha - \mu)'\Sigma^{-1}(x_\alpha - \mu)\}\right]$$

Taking log on both sides, we get

$$\log L = \frac{N}{2}\log|\Sigma^{-1}| - \frac{Np}{2}\log 2\pi - \left[\frac{1}{2}\sum_{\alpha=1}^{N}\{(x_\alpha - \mu)'\Sigma^{-1}(x_\alpha - \mu)\}\right]$$

Consider

$$\sum_{\alpha=1}^{N}(x_\alpha - \mu\ )'\,\Sigma^{-1}(x_\alpha - \mu) = tr\ \sum_{\alpha=1}^{N}\Sigma^{-1}(x_\alpha - \mu)(x_\alpha - \mu\ )'$$

$$= tr\ \Sigma^{-1}\sum_{\alpha=1}^{N}(x_\alpha - \mu)(x_\alpha - \mu\ )' \qquad (2.1)$$

Now

$$\sum_{\alpha=1}^{N}(x_\alpha - \mu)\,(x_\alpha - \mu)' + (\bar{x} - \mu)(\bar{x} - \mu)'\}$$

$$= \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' + (\bar{x} - \mu) \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})' + \left\{ \sum_{\alpha=1}^{N} (x_\alpha - \bar{x}) \right\} (\bar{x} - \mu)'$$

$$+ N(\bar{x} - \mu)(\bar{x} - \mu)'$$

Since $\sum_{\alpha=1}^{N} (x_\alpha - \bar{x}) = \sum_{\alpha=1}^{N} x_\alpha - N\bar{x} = 0$, we have

$$\sum_{\alpha=1}^{N} (x_\alpha - \mu\ )(x_\alpha - \mu)' = A + N(\bar{x} - \mu)(\bar{x} - \mu)'$$

Putting this value in (2.1), we have

$$\sum_{\alpha=1}^{N} (x_\alpha - \mu\ )'\ \Sigma^{-1}(x_\alpha - \mu)$$

$$= N(\bar{x} - \mu)'\ \Sigma^{-1}(\bar{x} - \mu) + tr\ \Sigma^{-1}A$$

Hence

$$logL$$

$$= \frac{N}{2} log|\Sigma^{-1}| - \frac{Np}{2} log\ 2\pi - \left[\frac{1}{2} N(\bar{x} - \mu)'\ \Sigma^{-1}(\bar{x} - \mu) + tr\ \Sigma^{-1}A \right]$$

Differentiating $log\ L$ with respect to μ, and equating it to 0, we get

$$\frac{\partial}{\partial \mu} log\ L = 0$$

$$\Rightarrow -0 - 0 - \frac{1}{2} 2\Sigma^{-1}(\mu - \bar{x}) - 0 = 0 \qquad \left(\because \frac{\partial X'AX}{\partial X} = 2AX \right)$$

$$\Rightarrow \Sigma^{-1}(\mu - \bar{x}) = 0$$

$$\Rightarrow \hat{\mu} = \bar{x}$$

2.3.2 Maximum Likelihood Estimates of Variance-Covariance Matrix when the Mean Vector is Known

The log-likelihood function is

$$logL$$

$$= \frac{N}{2} log|\Sigma^{-1}| - \frac{Np}{2} log\ 2\pi - \left[\frac{1}{2} N(\bar{x} - \mu)'\ \Sigma^{-1}(\bar{x} - \mu) + tr\ \ \Sigma^{-1} A\right]$$

Now substituting the value of $\hat{\mu} = \bar{x}$, in the above equation, it can be written as

$$logL$$

$$= \frac{N}{2} log|\Sigma^{-1}| - \frac{Np}{2} log\ 2\pi - \frac{1}{2} tr\ \Sigma^{-1} \sum_{\alpha=1}^{N} (x_\alpha - \mu\ )(x_\alpha - \ \mu)'$$

Differentiating the log-likelihood with respect to Σ^{-1} and equating it to 0, we get

$$\frac{\partial\ log\ L}{\partial \Sigma^{-1}} = 0$$

$$\Rightarrow \frac{N(|\Sigma^{-1}|)}{2|\Sigma^{-1}|} \hat{\Sigma} - 0 - \frac{1}{2} \sum_{\alpha=1}^{N} (x_\alpha - \mu)(x_\alpha - \ \mu)' = 0 \qquad \text{(from differantiation rules)}$$

$$\Rightarrow \frac{N}{2} \hat{\Sigma} = \frac{1}{2} \sum_{\alpha=1}^{N} (x_\alpha - \mu)(x_\alpha - \ \mu)'$$

$$\Rightarrow \hat{\Sigma} = \frac{1}{N} \sum_{\alpha=1}^{N} (x_\alpha - \mu\ )(x_\alpha - \ \mu)'.$$

2.3.3 Maximum Likelihood Estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ when both are Unknown

We can write the log likelihood function as

$$\log \mathrm{L} = \frac{N}{2} log|\Sigma^{-1}| - \frac{Np}{2} log\ 2\pi - \left[\frac{1}{2} N(\bar{x} - \mu)' \Sigma^{-1}(\bar{x} - \mu) + tr\ \Sigma^{-1} A\right]$$

Differentiating $log\ L$ with respect to μ, and equating it to 0, we get

$$\frac{\partial}{\partial \mu} log\ L = 0$$

$$\Rightarrow \Sigma^{-1}(\mu - \bar{x}) = 0$$

$$\Rightarrow \hat{\mu} = \bar{x}$$

Further differentiating $log\ L$ with respect to Σ, equating it to 0, and replacing μ by $\hat{\mu} = \bar{x}$, we get

$$\frac{N}{2}\hat{\Sigma} = \frac{1}{2}\sum_{\alpha=1}^{N}(x_\alpha - \bar{x})(x_\alpha - \bar{x})'$$

$$\Rightarrow \hat{\Sigma} = \frac{1}{N}\sum_{\alpha=1}^{N}(x_\alpha - \bar{x}\ )(x_\alpha - \bar{x})'$$

$$= \frac{1}{N} A$$

$$A = \sum_{\alpha=1}^{N}(x_\alpha - \bar{x}\ )(x_\alpha - \bar{x})'$$

Theorem 2.3.1.: Let $x_1, x_2, \ldots, x_N$, be a random sample of size N from $N_p(\mu, \Sigma)$. Then

$E(\bar{x}) = \mu$,

$E(\bar{x} - \mu)(\bar{x} - \mu)' = \frac{1}{N}\Sigma$.

Proof: We have

$E(\bar{x})$

$$= E\left(\frac{1}{N}\sum_{\alpha=1}^{N} x_\alpha\right)$$

$$= \frac{1}{N} E(\, x_1 + x_2 + \cdots + x_N)$$

$$= \frac{1}{N} (N\, \mu)$$

$$= \mu$$

Further

$$E(\bar{x} - \mu)(\bar{x} - \mu)'$$

$$= E\left[\frac{1}{N}(\, x_1 + x_2 + \cdots +\, x_N) - \mu\right]\left[\frac{1}{N}(\, x_1 + x_2 + \cdots +\, x_N) - \mu\right]'$$

$$= \frac{1}{N^2} E[(\, x_1 + x_2 + \cdots +\, x_N) - N\mu][(\, x_1 + x_2 + \cdots +\, x_N) - N\mu]'$$

$$= \frac{1}{N^2}[E(x_1 - \mu)(x_1 - \mu)' + E(x_2 - \mu)(x_2 - \mu)' + \cdots + E(x_N - \mu)(x_N - \mu)' + 0]$$

$$= \frac{1}{N^2} (N\, \Sigma)$$

$$= \frac{1}{N} \Sigma$$

2.4 Sufficient Statistics

Let $x_1, \ldots, x_N$ be a random sample of size N from a distribution having p.d.f. $f(x, \theta)$, where θ is an unknown population parameter. Then $t = \underline{t}(x_1, \ldots, x_N)$ is a sufficient statistic for θ if

$$\prod_{\alpha=1}^{N} f(x_\alpha, \theta) = g\ (t, \theta) h(x_1, \ldots, x_N),$$

where $h(x_1, \ldots, x_N)$ does not depend on θ.

2.4.1 Sufficient Statistics for the Parameters of a Multivariate Normal Distribution

The p.d.f. of a random vector $X \sim N_p(x|\mu, \Sigma)$ is given by

$$f(x|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\,(x-\ \mu\ )'\,\Sigma^{-1}(x-\ \mu\ )\right\}$$

Now, the likelihood function of $x_1, \ldots, x_N$ can be written as

$$L = \prod_{\alpha=1}^{N} f\,(x|\mu,\Sigma)$$

$$= \frac{1}{(2\pi)^{\frac{Np}{2}}|\Sigma|^{\frac{N}{2}}} \exp\left\{-\frac{1}{2}\sum_{\alpha=1}^{N}(x_\alpha-\mu\ )'\,\Sigma^{-1}(x_\alpha-\ \mu)\right\}$$

It can be shown that

$$\sum_{\alpha=1}^{N}(x_\alpha-\mu\ )'\,\Sigma^{-1}(x_\alpha-\ \mu) = tr\ \sum_{\alpha=1}^{N}\ \Sigma^{-1}(x_\alpha-\mu\ )(x_\alpha-\ \mu)'$$

$$= tr\ \Sigma^{-1}\sum_{\alpha=1}^{N}(x_\alpha-\mu\ )(x_\alpha-\ \mu)' \qquad (2.2.)$$

Consider

$$\sum_{\alpha=1}^{N}(x_\alpha-\mu)(x_\alpha-\ \mu)' = \sum_{\alpha=1}^{N}\{(x_\alpha-\bar{x})+(\bar{x}-\ \mu)\}\{(x_\alpha-\bar{x})+(\bar{x}-\mu)\}'$$

$$= \sum_{\alpha=1}^{N}\{(x_\alpha-\bar{x})(x_\alpha-\bar{x})' + (\bar{x}-\mu)(x_\alpha-\bar{x})' + (x_\alpha-\bar{x})(\bar{x}-\mu)'$$

$$+(\bar{x}-\ \mu)(\bar{x}-\ \mu)'$$

$$= \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})' + (\bar{x} - \mu) \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})' + \left\{ \sum_{\alpha=1}^{N} (x_\alpha - \bar{x}) \right\} (\bar{x} - \mu)'$$

$$+N(\bar{x} - \mu)(\bar{x} - \mu)'$$

Since $\sum_{\alpha=1}^{N} (x_\alpha - \bar{x}) = \sum_{\alpha=1}^{N} x_\alpha - N\bar{x} = \underline{0}$

Thus $\sum_{\alpha=1}^{N} (x_\alpha - \mu)(x_\alpha - \mu)' = A + N(\bar{x} - \mu)(\bar{x} - \mu)'$

Where

$$A = \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'$$

Putting this value in (2.2), we get

$$\sum_{\alpha=1}^{N} (x_\alpha - \bar{x})' \Sigma^{-1} (x_\alpha - \bar{x}) = tr[\, \Sigma^{-1} \{A + N(\bar{x} - \mu)(\bar{x} - \mu)'\}]$$

$$\Rightarrow \sum_{\alpha=1}^{N} (x_\alpha - \bar{x})' \Sigma^{-1} (x_\alpha - \bar{x}) = N(\bar{x} - \mu)' \, \Sigma^{-1} (\bar{x} - \mu) + tr \; \Sigma^{-1} A$$

Using the above equation, we get

$$\prod_{\alpha=1}^{N} f\,(x|\mu, \Sigma)$$

$$= \frac{1}{(2\pi)^{Np/2} |\Sigma|^{\frac{N}{2}}} exp\left[-\frac{1}{2} \{N(\bar{x} - \mu)' \, \Sigma^{-1} (\bar{x} - \mu) + tr \; \Sigma^{-1} A\} \right]$$

$$= \frac{1}{(2\pi)^{Np/2} |\Sigma|^{\frac{N}{2}}} \; exp\left[-\frac{N}{2} (\bar{x} - \mu)' \, \Sigma^{-1} (\bar{x} - \mu) \right] \; exp\left[-\frac{1}{2} \, tr \, (\Sigma^{-1} A) \right]$$

Using the factorization theorem, we observe that $\bar{x}$ and $\left(\frac{1}{N}\right)A$ form the sufficient set of statistics for μ and Σ. However, if μ is known, then $\frac{1}{N}\sum_{\alpha=1}^{N}(x_\alpha - \mu)(x_\alpha - \mu)'$ is a sufficient statistic for Σ. If Σ is known, then $\bar{x}$ is a sufficient statistic for μ.

2.5 Sample Multiple Correlation Coefficients

Sample Multiple Correlation Coefficient in multivariate analysis measures the strength of the relationship between multiple independent variables X and a single dependent variable Y. It indicates how well the independent variables collectively predict the dependent variable. In essence, it assesses the combined effect of multiple variables on a single outcome variable.

2.5.1 Applications

1. Predictive Modelling: It helps to evaluate the collective predictive power of multiple independent variables on a dependent variable.

2. Feature Selection: In multiple regression, it is applied to identify the most important independent variables contributing to the dependent variable.

3. Multivariate Analysis: It is used in techniques like multiple linear regression, principal component regression, and canonical correlation analysis.

4. Data Reduction: It helps to select a subset of variables that capture most of the variation in the data.

5. Inference: Make inferences about the population multiple correlation coefficient (ρ) based on the sample estimate (sample multiple correlation coefficient).

6. Comparison: Compare the strength of relationships between different sets of independent variables and a dependent variable.

7. Identification of Relationships: Detect and quantify the relationships between multiple variables in various fields, such as social sciences, healthcare, and finance.

8. Evaluation of Model Performance: Assess the goodness of fit of a model and compare the performance of different models.

2.5.2 Advantages

1. Measures Collective Impact: It evaluates the combined effect of multiple independent variables on a dependent variable.

2. Identifies Strong Relationships: Helps to identify the independent variables that have a strong relationship with the dependent variable.

3. Useful in Predictive Modelling: It is essential in building predictive models, such as multiple linear regression.

4. Quantifies Correlation: Provides a numerical measure to quantify the strength of the relationship.

5. Wide Applicability: Can be applied in various fields, including social sciences, healthcare, finance, and more.

2.5.3 Disadvantages

1. Assumes Linear Relationships: It assumes linear relationships between variables, which may not always be the case.

2. Sensitive to Outliers: It can be affected by outliers or extreme values in the data.

3. Does not Imply Causality: A high magnitude of multiple correlation coefficient does not imply causality between variables, but only correlation.

4. Can be Influenced by Multicollinearity: It can be affected by multicollinearity among independent variables.

5. Requires Large Samples: It requires a large sample size to produce reliable estimates.

2.6 Sample Partial Correlation Coefficients

Sample Partial Correlation Coefficients in multivariate analysis measure the association between two variables while controlling for the effect of a set of controlling random variables.

In the case of variables X_i, X_j and $X^{(2)} = \left(X_{q+1}, \dots, X_p\right)'$, it examines the extent of the relationship between variables X_i and X_j, after removing the effect of $X^{(2)}$.

The partial correlation coefficient between X_i and X_j $(1 \le i, j \le q)$ given $X^{(2)}$ is given by

$$\rho_{ij.q+1,\dots,p} = \frac{\sigma_{ij.q+1,\dots,p}}{\left(\sigma_{ii.q+1,\dots,p}\right)^{1/2}\left(\sigma_{jj.q+1,\dots,p}\right)^{1/2}}$$

where, $\Sigma_{11.2} = \left(\left(\sigma_{ij.q+1,\dots,p}\right)\right)$.

The maximum likelihood estimator of Σ is

$$\hat{\Sigma} = \frac{1}{N}\sum_{\alpha}\left(X_\alpha - \bar{X}\right)\left(X_\alpha - \bar{X}\right)'$$

where, $\overline{X} = \frac{1}{N}\sum_{\alpha=1}^{N} X_\alpha$.

The correspondence between $\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$ and $(\Sigma_{11.2}, B, \Sigma_{22})$ is

$\Sigma_{12} = B\Sigma_{22}$

$\Sigma_{11} = \Sigma_{11.2} + B\Sigma_{22}B'$

It follows that the maximum likelihood estimates of $\Sigma_{11.2}$, B and Σ_{22} are

$\hat{\Sigma}_{11.2} = \hat{\Sigma}_{11} - \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}\hat{\Sigma}_{21}$;

$\hat{B} = \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}$

and $\hat{\Sigma}_{22}$.

Hence MLE of $\rho_{ij.q+1,\ldots,p}$ is

$$\hat{\rho}_{ij.q+1,\ldots,p} = \frac{\hat{\sigma}_{ij.q+1,\ldots,p}}{\sqrt{\hat{\sigma}_{ii.q+1,\ldots,p}.\hat{\sigma}_{jj.q+1,\ldots,p}}} (i,j = 1,2,\ldots,q)$$

where $\hat{\sigma}_{ij.q+1,\ldots,p}$ is the $(i,j)^{th}$ element of $\hat{\Sigma}_{11.2}$. Let

$$\text{A}=\sum_{\alpha=1}^{N}\left(X_\alpha - \overline{X}\right)\left(X_\alpha - \overline{X}\right)' = \begin{pmatrix} \text{A}_{11} & \text{A}_{12} \\ \text{A}_{21} & \text{A}_{22} \end{pmatrix}$$

and

$$A_{11.2} = \left(\left(a_{ij.q+1\ldots p}\right)\right)$$

$$= A_{11} - A_{12}A_{22}^{-1}A_{21}$$

Then

$$\hat{\rho}_{ij.q+1,\ldots,p} = \frac{a_{ij.q+1,\ldots,p}}{\sqrt{a_{ii.q+1,\ldots,p}.a_{jj.q+1,\ldots,p}}}$$

$$= r_{ij.q+1\ldots p} \text{ (say)}$$

The estimate $r_{ij.q+1\ldots p}$ is called the sample partial correlation coefficient between $X_i \; and \; X_j$ holding $X_{q+1}, \ldots, X_p$ fixed.

2.6.1 Applications

1. Control for Confounding Variables: In observational studies, it helps to control for the effect of confounding variables, allowing researchers to examine the relationship between two variables of interest while accounting for the impact of additional variables.

2. Mediation Analysis: It is used to examine the relationship between two variables while adjusting for the effect of a mediating variable, helping to understand the underlying mechanisms and pathways.

3. Identification of Unique Relationships: It helps identify the unique relationship between two variables, independent of the effect of other variables, which is essential in understanding complex systems and relationships.

4. Data Reduction: It can be used to reduce data dimensionality by identifying the most important variables related to an outcome variable, while controlling for the effect of other variables.

5. Path Analysis: It is used in path analysis to examine the direct and indirect relationships between variables, helping to understand the causal relationships and pathways.

6. Structural Equation Modelling: The partial correlation coefficient is used in structural equation modelling to examine the relationships between latent variables while controlling for the effect of other variables.

7. Feature Selection: It can be used to select the most relevant features (variables) related to an outcome variable while controlling for the effect of other variables.

8. Causal Inference: It can be used to make causal inferences about the relationships between variables while controlling for the effect of other variables.

2.6.2 Advantages

1. Controls for Confounding Variables: It helps to isolate the relationship between two variables, controlling for the effect of additional variables.

2. Identifies Unique Relationships: It reveals the unique relationship between two variables, independent of the effect of other variables.

3. Handles Multiple Variables: It can handle multiple variables, allowing for the examination of complex relationships.

4. Robust to Noise: It is robust to noise and outliers in the data.

5. Wide Applicability: It can be applied in various fields, including social sciences, healthcare, finance, and more.

2.6.3 Disadvantages

1. Assumes Linear Relationships: It assumes linear relationships between variables, which may not always be the case.

2. Sensitive to Multicollinearity: It can be affected by multicollinearity among independent variables.

3. Requires Large Samples: It requires large sample sizes to produce reliable estimates.

4. Difficult to Interpret: It can be challenging to interpret, especially for non-experts.

5. Computer-Intensive: Its calculations can be computationally intensive, especially for large datasets.

2.7 Regression Coefficient

In multivariate analysis, a regression coefficient (also known as a beta coefficient or parameter estimate) is a numerical value that represents the change in the dependent variable (outcome variable) in response to a unit change in an independent variable (predictor variable), while holding all other independent variables constant.

Let

$$X = \begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}_{p-q}^{q} \sim N\left(\begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right)$$

Then

$$E\left(X^{(1)} \middle| X^{(2)}\right) = \mu^{(1)} + B\left(X^{(2)} - \mu^{(2)}\right)$$

with $B = \Sigma_{12}\Sigma_{22}^{-1}$. B is the matrix of regression coefficients of $X^{(1)}$ on $X^{(2)}$ and its MLE is

$$\hat{B} = \hat{\Sigma}_{12}\hat{\Sigma}_{22}^{-1}.$$

The regression coefficients describe the relationship between each independent variable and the dependent variable, while controlling for the effects of all other independent variables in the model.

For example, in a multiple linear regression model predicting stock prices (dependent variable) based on economic indicators (independent variables), a regression coefficient of 0.5 for the "GDP" variable means that for every one-unit increase in GDP, the stock price is expected to increase by 0.5 units, holding all other independent variables constant.

Regression coefficients are essential in multivariate analysis because they help to

1. Understand the relationships between variables

2. Predict outcomes

3. Identify important predictors

4. Control for confounding variables

2.7.1 Applications

1. Prediction: Use regression coefficients to predict dependent variable values for new observations.

2. Variable Selection: Identify significant predictors by examining the magnitude and significance of regression coefficients.

3. Relationship Interpretation: Understand the direction and strength of relationships between independent variables and the dependent variable.

4. Control for Confounding: Hold constant the effects of confounding variables to isolate the relationship between a specific independent variable and the dependent variable.

5. Mediation Analysis: Examine the indirect effects of independent variables on the dependent variable through a mediator variable.

6. Moderation Analysis: Investigate how the relationship between an independent variable and the dependent variable changes based on the level of a moderator variable.

7. Path Analysis: Examine complex relationships between variables using regression coefficients to estimate path coefficients.

8. Structural Equation Modelling: Use regression coefficients to estimate the relationships between latent variables and observed variables.

9. Feature Selection: Select the most important independent variables based on the magnitude of their regression coefficients.

10. Model Building: Use regression coefficients to develop and refine multivariate models that explain the relationships between variables.

Regression Coefficients are Essential in Various Fields, Including:

1. Business: Predict stock prices, sales, or customer behaviour.

2. Healthcare: Analyse the relationship between health outcomes and risk factors.

3. Social Sciences: Examine the relationships between social variables, such as crime rates and economic indicators.

4. Marketing: Understand the impact of advertising on sales.

5. Finance: Predict credit risk or stock performance.

2.7.2 Advantages

1. Quantifies Relationships: Regression coefficients provide a numerical measure of the strength and direction of relationships between variables.

2. Controls for Confounding: Regression coefficients help in controlling the effects of confounding variables, allowing for a more accurate understanding of relationships.

3. Predictive Power: Regression coefficients enable predictions of dependent variable values for new observations.

4. Identifies Important Predictors: Regression coefficients help identify the most important independent variables contributing to the dependent variable.

5. Flexibility: Regression coefficients can be used in various multivariate techniques, such as path analysis and structural equation modelling.

6. Interpretability: Regression coefficients are easily interpretable, allowing for a clear understanding of relationships.

7. Wide Applicability: Regression coefficients are widely used in various fields, including business, healthcare, social sciences, and more.

2.7.3 Disadvantages

1. Assumes Linear Relationships: Regression coefficients assume linear relationships between variables, which may not always be the case.

2. Sensitive to Multicollinearity: Regression coefficients can be affected by multicollinearity among independent variables.

3. Requires Large Samples: Regression coefficients require large sample sizes to produce reliable estimates.

4. Can be Misleading: Regression coefficients can be misleading if the underlying assumptions are not met or if the model is poorly specified.

5. Difficult to Interpret: Regression coefficients can be challenging to interpret in complex models or with non-significant results.

6. Computer-Intensive: Regression coefficient calculations can be computationally intensive for large datasets.

7. Assumes no Measurement Error: Regression coefficients assume no measurement error in the variables, which may not always be the case.

2.8 Summary

In this unit, we have covered the estimation of parameters of multivariate normal distribution using maximum likelihood estimation. The maximum likelihood estimation gives a unique and easy-to-determine solution in the case of multivariate normal distribution.

2.9 Self-Assessment Exercises

1. Let $X_\alpha; \alpha = 1,2, \dots, N$ be p component random sample from $N(\mu, \Sigma)$. Obtain of the maximum likelihood estimate of μ and Σ.
2. Let $X_\alpha; (\alpha = 1,2, \dots, N)$ be a i.i.d. random sample from $N_p(\mu, \Sigma)$. Obtain the M.L.E. of μ when Σ is known.
3. Let $X_\alpha; (\alpha = 1,2, \dots, N)$ be a i.i.d. random sample from $N_p(\mu, \Sigma)$. Obtain the M.L.E. of Σ when μ is known.
4. If $a_1X_1 + a_2X_2 + \cdots + a_pX_p = k$ (constant), then find $\rho_{12.3,4,\dots,p}$ and $r_{1.2,3,\dots,p}$
5. Let X be a random vector with the mean vector μ and dispersion matrix Σ. Using Markov's inequality, show that

$$P[(X-\mu)'\Sigma^{-1}(X-\mu) > \lambda] < \frac{p}{\lambda} \quad \text{for } \lambda > 0$$

2.10 References

- Johnson, R. A., Wichern, D. W. (2019): Applied Multivariate Statistical Analysis. United Kingdom: Pearson
- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Giri Narayan C. (1995): Multivariate Statistical Analysis.
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.
- Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York.

2.11 Further Readings

- Khatri C. G.: Multivariate Analysis.
- Mardia K. V.: Multivariate Analysis
- Seber, G.A.F.: *Multivariate Observations.* Wiley, New York.

UNIT 3: SAMPLING DISTRIBUTIONS

Structure

3.1 Introduction

Any function $T(X)$ of a random sample X that does not depend on any unknown parameter, is called a Statistic.

Since a statistic $T(X)$ is a function of X, it is a random variable with an associated probability distribution.

The probability distribution of a statistic is called its sampling distribution.

3.2 Objectives

After studying this unit, you should be able to:

- Define the correlation coefficients.
- Calculate the multiple and partial correlation coefficients.

- Describe the distribution of multiple correlation coefficients.
- Describe the distribution of partial correlation coefficients.

3.3 Distribution of the Matrix of Sample Regression Coefficients

Theorem 3.3.1.: Let $X_1, X_2, \dots, X_N$ be independently distributed with $X_\alpha \sim N_p(\mu_\alpha, \Sigma)$. Let $C = (C_{\alpha\beta})$ be $N \times N$ orthogonal matrix. Then $y_\alpha = \sum_{\beta=1}^{N} C_{\alpha\beta} X_\beta$ is distributed according to $N(\theta_\alpha, \Sigma)$, where $\theta_\alpha = \sum_{\beta=1}^{N} C_{\alpha\beta}\mu_\beta$ and $y_1, y_2, \dots, y_N$ are independently distributed.

Proof: Since y_α is a set of linear combinations of the components of $\{X_\alpha\}$, which have a joint normal distribution, the set of vectors $\{y_\alpha\}$ have a joint normal distribution. The expected value of y_α is

$$E(y_\alpha) = E\left[\sum_{\beta=1}^{N} C_{\alpha\beta} X_\beta\right] = \theta_\alpha$$

$$cov\,(y_\alpha, y_\gamma) = E(y_\alpha - \theta_\alpha)(y_\gamma - \theta_\gamma)'$$

$$= E\left[\sum_{\beta=1}^{N} C_{\alpha\beta}(X_\beta - \mu_\beta)\right]\left[\sum_{\beta^*=1}^{N} C_{\gamma\beta^*}(X_{\beta^*} - \mu_{\beta^*})'\right]'$$

$$= \sum_{\beta=1}^{N}\sum_{\beta^*=1}^{N} C_{\alpha\beta} C_{\gamma\beta^*} E(X_\beta - \mu_\beta)(X_{\beta^*} - \mu_{\beta^*})'$$

Notice that

$$E(X_\beta - \mu_\beta)(X_{\beta^*} - \mu_{\beta^*})' = \begin{cases} \Sigma, \text{if } \beta = \beta^* \\ 0, \quad \text{if } \beta \neq \beta^* \end{cases}$$

$$= \delta_{\beta\beta^*}\Sigma$$

where $\delta_{\beta\beta^*} = 1$ if $\beta = \beta^*$ and 0 otherwise.

Hence

$$cov\left(y_\alpha, y_\gamma\right) = \sum_{\beta=1}^{N}\sum_{\beta^*=1}^{N} C_{\alpha\beta}C_{\gamma\beta^*}\delta_{\beta\beta^*}\Sigma$$

$$= \left(\sum_{\beta=1}^{N} C_{\alpha\beta}C_{\gamma\beta}\right)\Sigma$$

Since C is an orthogonal matrix, $\sum_{\beta=1}^{N} C_{\alpha\beta}C_{\gamma\beta} = 1$ if $\alpha = \gamma$ and 0 if $\alpha \neq \gamma$. Therefore

$$cov\left(y_\alpha, y_\gamma\right) = \delta_{\alpha\gamma}\Sigma$$

Hence y_α and y_γ are independent for $\alpha \neq \gamma$ and covariance matrix of y_α is Σ. This implies that

$$y_\alpha \sim N(\theta_\alpha, \Sigma)$$

Let $C = \left(\left(C_{\alpha\beta}\right)\right)$ be orthogonal and $y_\alpha = \sum_{\beta=1}^{N} C_{\alpha\beta}X_\beta$, then

$$\sum_{\alpha=1}^{N} y_\alpha y_\alpha' = \sum_{\alpha}\left[\sum_{\beta} C_{\alpha\beta}X_\beta \sum_{\gamma} C_{\alpha\gamma}X_\gamma'\right] = \sum_{\beta,\gamma}\left(\sum_{\alpha} C_{\alpha\beta}C_{\alpha\gamma}\right)X_\beta X_\gamma'$$

$$= \sum_{\beta,\gamma}\delta_{\beta\gamma}X_\beta X_\gamma' = \sum_{\beta} X_\beta X_\beta'$$

Hence $\sum_{\alpha=1}^{N} y_\alpha y_\alpha' = \sum_{\alpha=1}^{N} X_\alpha X_\alpha'$

Theorem 3.3.2.: The mean of a sample of size N from $N(\mu, \Sigma)$ is distributed according to $N\left(\mu, \frac{1}{N}\Sigma\right)$ and independent of $\hat{\Sigma}$. Further, N$\hat{\Sigma}$ is distributed as $\sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$ where Z_α is distributed according to $N(0, \Sigma)$ independently of $Z_\beta (\alpha \neq \beta)$.

Proof: There exists as $N \times N$ orthogonal matrix $C = \left(\left(C_{\alpha\beta}\right)\right)$ with the last row $\left(\frac{1}{\sqrt{N}}, \frac{1}{\sqrt{N}}, \ldots, \frac{1}{\sqrt{N}}\right)$.

Define

$$Z_\alpha = \sum_\beta C_{\alpha\beta} X_\beta$$

Then

$$Z_N = \sum_\beta \frac{1}{\sqrt{N}} X_\beta$$

$$= \sqrt{N}\overline{X}$$

$$A = \sum_{\alpha=1}^{N} X_\alpha X_\alpha' - N\,\overline{X}\,\overline{X}'$$

$$= \sum_{\alpha=1}^{N} Z_\alpha Z_\alpha' - Z_N Z_N'$$

$$= \sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$$

Since, Z_N is independent of $Z_1, \ldots, Z_{N-1}$, this implies that $\overline{X}$ is independent of A.

$$EZ_N = \sqrt{N}\mu$$

$$E\left(Z_N - \sqrt{N}\mu\right)\left(Z_N - \sqrt{N}\mu\right)' = \Sigma$$

Hence $Z_N \sim N\left(\sqrt{N}\mu, \Sigma\right)$ and $\overline{X} = \left(\frac{1}{\sqrt{N}}\right) Z_N \sim N\left(\mu, \frac{1}{N}\Sigma\right)$

Further $\forall \alpha \neq N$

$$EZ_\alpha = \sum_\beta C_{\alpha\beta} EX_\beta$$

$$= \sum_\beta C_{\alpha\beta}\mu$$

$$= \left(\sum_{\beta} C_{\alpha\beta} C_{N\beta} \right) \sqrt{N}\mu$$

$$= 0$$

Hence A is distributed according to $\sum_{\alpha=1}^{N-1} Z_\alpha Z'_\alpha$, where $Z_\alpha \sim N(0, \Sigma)$. Then

$$E(\hat{\Sigma}) = \frac{1}{N} E\left(\sum_{\alpha=1}^{N-1} Z_\alpha Z'_\alpha \right)$$

$$= \frac{N-1}{N} \Sigma$$

Hence $\hat{\Sigma}$ is a biased estimator of Σ. We shall, therefore, define

$$S = \frac{1}{N-1} A$$

$$= \frac{1}{N-1} \sum_{\alpha=1}^{N} (X_\alpha - \overline{X})(X_\alpha - \overline{X})'$$

as the sample covariance matrix. It is an unbiased estimator of Σ.

Now, the likelihood function of $X_1, \ldots, X_N$ can be written as

$$\frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{N}{2}}} \; exp\left[-\frac{1}{2} \sum_{\alpha=1}^{N} (X_\alpha - \mu)' \Sigma^{-1} (X_\alpha - \mu) \right]$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{N}{2}}} \; exp\left[-\frac{1}{2} \left\{ N(\overline{X} - \mu)' \Sigma^{-1} (\overline{X} - \mu) + tr(\Sigma^{-1} A) \right\} \right]$$

Using the factorization theorem, we observe that $\overline{X}$ and $\left(\frac{1}{N}\right) A$ form a sufficient set of statistics for μ and Σ. However, if μ is known $\left(\frac{1}{N}\right) \sum_{\alpha=1}^{N} (X_\alpha - \mu)(X_\alpha - \mu)'$ is a sufficient statistic for Σ.

3.4 Distribution of Sample Correlation Coefficients

If $X \sim N_p(X|\mu, \Sigma)$ then the conditional distribution of sub-vector $X^{(1)}$ given $X^{(2)}$ is $N\left(\mu^{(1)} + B\left(X^{(2)} - \mu^{(2)}\right), \Sigma_{11.2}\right)$, with

$B = \Sigma_{12}\Sigma_{22}^{-1}$

$\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$

Multiple Correlation Coefficients

For the sake of convenience, we will treat the case of multiple correlation coefficients between X_1 and the set $X_2, \ldots, X_p$. Let $\overline{R}$ denotes the population multiple correlation coefficient between X_1 and $X_2, \ldots, X_p$. Then

$$\overline{R} = \rho_{1.2,3\ldots,p} = \frac{\beta\Sigma_{22}\beta'}{\sqrt{\sigma_{11}\beta\Sigma_{22}\beta'}} = \sqrt{\frac{\beta\Sigma_{22}\beta'}{\sigma_{11}}} = \sqrt{\frac{\sigma_{(1)}\Sigma_{22}^{-1}\sigma'_{(1)}}{\sigma_{11}}}$$

where

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{(1)} \\ \sigma'_{(1)} & \Sigma_{22} \end{pmatrix}$$

The sample multiple correlation coefficients between X_1 and the set $X_2, \ldots, X_p$ is

$$R = \sqrt{\frac{\hat{\beta}\hat{\Sigma}_{22}\hat{\beta}'}{\hat{\sigma}_{11}}} = \sqrt{\frac{\hat{\sigma}_{(1)}\hat{\Sigma}_{22}^{-1}\hat{\sigma}'_{(1)}}{\hat{\sigma}_{11}}} = \sqrt{\frac{a_{(1)}A_{22}^{-1}a'_{(1)}}{a_{11}}}$$

where

$\hat{\beta} = \hat{\sigma}_{(1)}\hat{\Sigma}_{22}^{-1}$

$= \frac{a_{(1)}}{N}\left(\frac{A_{22}}{N}\right)^{-1}$

$= a_{(1)}A_{22}^{-1}$

$$\hat{\Sigma} = \frac{1}{N}A = \begin{pmatrix} \hat{\sigma}_{11} & \hat{\sigma}_{(1)} \\ \hat{\sigma}'_{(1)} & \hat{\Sigma}_{22} \end{pmatrix}$$

$$A = \begin{pmatrix} a_{11} & a_{(1)} \\ a'_{(1)} & A_{22} \end{pmatrix}$$

Since we can define $\overline{R}, \sigma_{(1)}, \Sigma_{22}$ as a one-to-one transformation of Σ, R is the MLE of $\overline{R}$.

Further, we obtain

$$1 - R^2 = 1 - \frac{a_{(1)}A_{22}^{-1}a'_{(1)}}{a_{11}}$$

$$= \frac{a_{11} - a_{(1)}A_{22}^{-1}a'_{(1)}}{a_{11}}$$

$$= \frac{\left|a_{11} - a_{(1)}A_{22}^{-1}a'_{(1)}\right||A_{22}|}{a_{11}|A_{22}|}$$

$$= \frac{|A|}{a_{11}|A_{22}|}$$

Result: Determinant of partitioned matrix:

$$|A| = \begin{vmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{vmatrix}$$

$$= |A_{11}||A_{22} - A_{21}A_{11}^{-1}A_{12}|$$

$$= |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{21}|$$

Proof: We have

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11} & 0 \\ A_{21} & I \end{bmatrix}\begin{bmatrix} I & A_{11}^{-1}A_{12} \\ 0 & A_{22} - A_{21}A_{11}^{-1}A_{12} \end{bmatrix}$$

$$= \begin{bmatrix} I & A_{21} \\ 0 & A_{22} \end{bmatrix} \begin{bmatrix} A_{11} - A_{12}A_{22}^{-1}A_{21} & 0 \\ A_{22}^{-1}A_{21} & I \end{bmatrix}$$

$$= [A_{22}][A_{11} - A_{12}A_{22}^{-1}A_{21}]$$

Taking determinant, we have

$$|A| = |A_{22}||A_{11} - A_{12}A_{22}^{-1}A_{21}|$$

Note: R and $\hat{\beta}$ have the same properties in the sample that $\overline{R}$ and β have in the population.

Result: Among all $(p-1)$ component row vectors d defining linear combination $dX_{\alpha}^{(2)}$ of the components of $X_{\alpha}^{(2)}$ the vector $d = \hat{\beta}$ is the one that minimizes

$$\sum_{\alpha=1}^{N} \left[\left(X_{1\alpha} - \overline{X}_1\right) - d\left(X_{\alpha}^{(2)} - \overline{X}^{(2)}\right) \right]^2$$

Proof: Since $\hat{\beta} = a_{(1)}A_{22}^{-1}$

$$\sum_{\alpha=1}^{N} \left[\left(X_{1\alpha} - \overline{X}_1\right) - d\left(X_{\alpha}^{(2)} - \overline{X}^{(2)}\right) \right]^2$$

$$= \sum \left[\left(X_{1\alpha} - \overline{X}_1\right) - \hat{\beta}\left(X_{\alpha}^{(2)} - \overline{X}^{(2)}\right) \right]^2 + \left(\hat{\beta} - d\right)A_{22}\left(\hat{\beta} - d\right)$$

$$\geq \sum \left[\left(X_{1\alpha} - \overline{X}_1\right) - \hat{\beta}\left(X_{\alpha}^{(2)} - \overline{X}^{(2)}\right) \right]^2$$

The equality occurs when $\left(\hat{\beta} - d\right)A_{22}\left(\hat{\beta} - d\right)$, i.e., when $d = \hat{\beta} = a_{(1)}A_{22}^{-1}$.

Since

$$\sum_{\alpha=1}^{N} \left[\left(X_{1\alpha} - \overline{X}_1\right) - \hat{\beta}\left(X_{\alpha}^{(2)} - \overline{X}^{(2)}\right) \right] \left(X_{\alpha}^{(2)} - \overline{X}^{(2)}\right)'$$

$$= a_{(1)} - \hat{\beta}A_{22} = 0$$

Hence the minimum value is

$$\sum \left[\left(X_{1\alpha} - \overline{X}_1 \right) - \hat{\beta} \left(X_{\alpha}^{(2)} - \overline{X}^{(2)} \right) \right]^2$$

$$= a_{11} - a_{(1)} A_{22}^{-1} a'_{(1)} = a_{11.2}.$$

3.5 Distribution of the Sample Multiple Correlation Coefficient

The Distribution of the Sample Multiple Correlation Coefficient when the population multiple correlation coefficient is zero.

The sample multiple correlation coefficients between X_1 and the set $X_2, \dots, X_p$ is

$$R^2 = \frac{a_{(1)} A_{22}^{-1} a'_{(1)}}{a_{11}}$$

and

$$1 - R^2 = \frac{a_{11} - a_{(1)} A_{22}^{-1} a'_{(1)}}{a_{11}}$$

$$= \frac{a_{11.2}}{a_{11}}$$

Therefore

$$\frac{R^2}{1 - R^2}$$

$$= \frac{a_{(1)} A_{22}^{-1} a'_{(1)}}{a_{11}} \left(\frac{a_{11} - a_{(1)} A_{22}^{-1} a'_{(1)}}{a_{11}} \right)^{-1}$$

$$= \frac{a_{(1)} A_{22}^{-1} a'_{(1)}}{a_{11} - a_{(1)} A_{22}^{-1} a'_{(1)}}$$

$$= \frac{a_{(1)} A_{22}^{-1} a'_{(1)}}{a_{11.2}}$$

Theorem 3.3.3: If $\overline{R} = 0$, $i.e.$, $\left(\sigma_{12}, \dots, \sigma_{1p} \right)' = \sigma_{(1)} = 0 = \beta$, then $\frac{(N-p)R^2}{(p-1)(1-R^2)}$ is distributed as F with $(p-1)$ and $(N-p)$ degrees of freedom.

Proof: If $y_\alpha \sim N(\Gamma w_\alpha, \Phi), G = YH^{-1}, H = \sum_{\alpha=1}^{m} w_\alpha w'_\alpha$, then $\sum_{\alpha=1}^{m} y_\alpha y'_\alpha - GHG'$ is distributed as $\sum_{\alpha=1}^{m-r} U_\alpha U'_\alpha$ where $U_\alpha \sim N(0, \Phi)$.

If $\Gamma = 0$, GHG' is distributed as $\sum_{\alpha=m-r+1}^{m} U_\alpha U'_\alpha$.

When $\beta = 0$, i.e., $\overline{R} = 0$, $a_{11.2} = a_{11} - a_{(1)}A_{22}^{-1}a'_{(1)}$ is distributed as $\sum_{\alpha=1}^{N-p} V_\alpha^2$ and $a_{(1)}A_{22}^{-1}a'_{(1)}$ is distributed as $\sum_{\alpha=N-p+1}^{N-1} V_\alpha^2$, where V_α are independent, each with distribution $N(0, \sigma_{11})$.

Then

$$\frac{a_{11.2}}{\sigma_{11}} \sim \chi^2_{N-p}$$

Consider

$$\frac{a_{11}}{\sigma_{11}} = \frac{a_{11} - a_{(1)}A_{22}^{-1}a'_{(1)}}{\sigma_{11}} + \frac{a_{(1)}A_{22}^{-1}a'_{(1)}}{\sigma_{11}}$$

or

$$Q = Q_1 + Q_2$$

where

$$Q \sim \chi^2_{N-1}$$

and

$$Q_1 \sim \chi^2_{N-p}$$

From the Fisher Cochran Theorem Q_2 is independently distributed of Q_1 and follows $\chi^2_{N-1-N+p}$.

Hence

$$\frac{a_{(1)}A_{22}^{-1}a'_{(1)}}{\sigma_{11}} \sim \chi^2_{p-1}$$

Since $\frac{a_{11.2}}{\sigma_{11}}$ and $\frac{a_{(1)}A_{22}^{-1}a'_{(1)}}{\sigma_{11}}$ are independently distributed, we have

$$F = \frac{R^2}{1-R^2} \times \frac{N-p}{p-1}$$

$$= \frac{\frac{a_{(1)}A_{22}^{-1}a'_{(1)}}{\sigma_{11}}}{\frac{a_{11.2}}{\sigma_{11}}} \times \frac{N-p}{p-1}$$

$$= \frac{\chi^2_{p-1}}{\chi^2_{N-p}} \times \frac{N-p}{p-1}$$

$$\sim F_{p-1,N-p}$$

The p.d.f. of the statistic F is

$$f(F) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} F^{\frac{\nu_1}{2}-1}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1}{\nu_2}F\right)^{\frac{\nu_1+\nu_2}{2}}}$$

Substituting

$\nu_1 = p-1,$

$\nu_2 = N-p,$

$$F = \frac{R^2}{1-R^2}\left(\frac{\nu_2}{\nu_1}\right)$$

and observing that

$$\frac{dF}{dR^2} = \frac{1}{(1-R^2)^2}\left(\frac{\nu_2}{\nu_1}\right)$$

we get the pdf of R^2 as

$$g_{R^2}(r^2) = \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} \left\{\frac{r^2}{1-r^2}\left(\frac{\nu_2}{\nu_1}\right)\right\}^{\frac{\nu_1}{2}-1}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)\left(1+\frac{r^2}{1-r^2}\right)^{\frac{\nu_1+\nu_2}{2}}} \times \frac{1}{(1-r^2)^2}\left(\frac{\nu_2}{\nu_1}\right)$$

$$= \frac{\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}-\frac{\nu_1}{2}+1-1} \left(\frac{R^2}{1-R^2}\right)^{\frac{\nu_1}{2}-1}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)\left(\frac{1}{1-R^2}\right)^{\frac{\nu_1+\nu_2}{2}}} \times \frac{1}{(1-R^2)^2}$$

$$= \frac{(R^2)^{\frac{\nu_1}{2}-1}(1-R^2)^{\frac{\nu_1+\nu_2}{2}-\frac{\nu_1}{2}+1-2}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}$$

$$= \frac{(R^2)^{\frac{\nu_1}{2}-1}(1-R^2)^{\frac{\nu_2}{2}-1}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}$$

Putting $\frac{dR^2}{dR} = 2R$, we get the pdf of R as

$$g_R(r) = \frac{2(r)^{\nu_1-1}(1-r^2)^{\frac{\nu_2}{2}-1}}{B\left(\frac{\nu_1}{2},\frac{\nu_2}{2}\right)}$$

or

$$g_R(r) = \frac{2(r)^{p-2}(1-r^2)^{\frac{N-p-2}{2}}}{B\left(\frac{p-1}{2},\frac{N-p}{2}\right)}.$$

Note: If $\overline{R} \neq 0$, the distribution of R is much more difficult to derive.

Example: Derive the coefficient of correlation of bivariate case from the multiple correlation coefficient. Solution: Let the predictor variable be X and the response variable be Y. Then, we have

$$\Sigma = \begin{bmatrix} \sigma_X^2 & \rho\ \sigma_X\sigma_Y \\ \rho\ \sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

where σ_X and σ_Y be variances of X and Y and the correlation coefficient between them be ρ.

Then

$Var(Y|X) = \sigma_Y^2(1-\rho^2)$,

$$R_{XY}^2 = \frac{\sigma_Y^2 - \sigma_Y^2(1-\rho^2)}{\sigma_Y^2}$$

$$= \rho^2$$

and hence

$R = |\rho|$

Example: Consider the mean vector be $\mu_X = \begin{bmatrix} 3 \\ -2 \end{bmatrix}$ and $\mu_Y = 4$ and the variance-covariance matrices are

$\Sigma_{XX} = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}, \sigma_{YY} = 9\,, \sigma_{XY} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$

i) Fit the equation $Y = b_0 + b_1X_1 + b_2X_2$ as best linear equation

ii) Find the multiple correlation coefficient

iii) Also obtain the mhe mean square error.

Solution: (i)

$b = \Sigma_{XX}^{-1}\,\sigma_{XY}$

$$= \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

$b_0 = \mu_Y - b'\mu_X$

$$= 4 - [2 \quad -1] \begin{bmatrix} 3 \\ -2 \end{bmatrix}$$

$$= 4 - 8$$

$$= -4$$

Therefore,

$$Y = b_0 + b'X$$

$$= -4 + 2X_1\_X_2$$

(ii) The multiple correlation coefficient is

$$R^2 = \frac{\sigma'_{XY}\Sigma_{XX}^{-1}\,\sigma_{XY}}{\sigma_{YY}}$$

$$= \frac{[3 \quad 1]\begin{bmatrix}2 & 1\\ 1 & 1\end{bmatrix}^{-1}\begin{bmatrix}3\\ 1\end{bmatrix}}{9}$$

$$= \frac{1}{9}\left[[3 \quad 1]\begin{bmatrix}2 & 1\\ 1 & 1\end{bmatrix}^{-1}\begin{bmatrix}3\\ 1\end{bmatrix}\right]$$

$$= \frac{1}{9}\left[[3 \quad 1]\begin{bmatrix}2\\ -1\end{bmatrix}\right]$$

$$= \frac{5}{9}$$

$$\Rightarrow R = \sqrt{\frac{5}{9}} = 0.75$$

(iii) The mean square error is

$$M.S.E. = \sigma_{YY}(1 - R^2)$$

$$= 9\left(1 - \frac{5}{9}\right)$$

$$= 4$$

$$\Rightarrow M.S.E. = 4$$

Likelihood Ratio Criteria

Now we derive the likelihood ratio test of testing the hypothesis $H_0: \overline{R} = 0$. Since $\overline{R} \geq 0$, the alternative hypothesis is $H_1: \overline{R} > 0$.

$$\overline{R} = \sqrt{\frac{\sigma_{(1)}\Sigma_{22}^{-1}\sigma'_{(1)}}{\sigma_{11}}} = 0$$

$$\Leftrightarrow \sigma_{(1)}\Sigma_{22}^{-1}\sigma'_{(1)} = 0$$

$$\Leftrightarrow \sigma_{(1)} = 0$$

Ω: parameter space

ω: region in the parameter space specified by H_0

The likelihood function is

$$L(\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{1}{2}pN}|\Sigma|^{\frac{1}{2}N}} \exp\left[-\frac{1}{2}\sum_{\alpha}(X_\alpha - \mu)'\,\Sigma^{-1}(X_\alpha - \mu)\right]$$

We compute

$$\lambda = \frac{max_{(\mu,\Sigma)\in\omega} L(\mu, \Sigma)}{max_{(\mu,\Sigma)\in\Omega} L(\mu, \Sigma)}$$

The likelihood ratio criterion is that if $\lambda \leq \lambda_0$, λ_0 is some specified value, then we reject the hypothesis H_0.

Now, Maximum of $L(\mu, \Sigma)$ over Ω occurs is

$$L(\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{1}{2}pN}\left|\frac{\mathrm{A}}{\mathrm{N}}\right|^{\frac{1}{2}N}} \exp\left[-\frac{1}{2}\sum_{\alpha}(X_\alpha - \bar{X})'\left(\frac{\mathrm{A}}{\mathrm{N}}\right)^{-1}(X_\alpha - \bar{X})\right]$$

$$L(\mu, \Sigma) = \frac{|\mathrm{N}|^{\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}pN}|\mathrm{A}|^{\frac{1}{2}N}} exp\left[-\frac{1}{2}tr\left(\frac{\mathrm{A}}{\mathrm{N}}\right)^{-1}\sum_{\alpha}(X_\alpha - \bar{X})'(X_\alpha - \bar{X})\right]$$

$\mu^* = \overline{X}$,

$$\Sigma^* = \left(\frac{1}{N}\right) A$$

$$= \frac{1}{N} \sum_{\alpha} (X_\alpha - \bar{X})\,(X_\alpha - \bar{X})'$$

and

$$max_{\mu^*, \Sigma^* \in \Omega} L(\mu^*, \Sigma^*) = \frac{N^{\frac{1}{2}pN} e^{-\frac{1}{2}pN}}{(2\pi)^{\frac{1}{2}pN} |A|^{\frac{1}{2}N}}$$

In region $\omega, \Sigma = \begin{pmatrix} \sigma_{11} & 0 \\ 0 & \Sigma_{22} \end{pmatrix}$ and the likelihood function is

$$L\left(\mu^*, \Sigma^* \middle| \sigma^*_{(1)} = 0\right)$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}N} {\sigma^*_{11}}^{\frac{1}{2}N}} \exp\left[-\frac{1}{2\sigma^*_{11} \sum (X_{1\alpha} - \mu^*_1)^2}\right]$$

$$\times \frac{1}{(2\pi)^{\frac{1}{2}(p-1)N} |\Sigma_{22}|^{\frac{1}{2}N}} \exp\left[-\frac{1}{2} \sum_{\alpha} \left(X^{(2)}_\alpha - \mu^{(2)}\right)' \Sigma^{-1}_{22} \left(X^{(2)}_\alpha - \mu^{(2)}\right)\right]$$

The first factor is maximized at $\mu^*_1 = \overline{X}_1 \; and \; \sigma^*_{11} = \frac{1}{N} a_{11}$, and the second factor is maximised when $\mu^{(2)} = \overline{X}^{(2)}, \Sigma_{22} = \frac{1}{N} A_{22}$. The maximized function is

$$max_{\mu^*, \Sigma^* \in \omega} L(\mu^*, \Sigma^*) = \frac{N^{\frac{1}{2}N} e^{-\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}N} {a_{11}}^{\frac{1}{2}N}} \times \frac{N^{\frac{1}{2}(p-1)N} e^{-\frac{1}{2}(p-1)N}}{(2\pi)^{\frac{1}{2}(p-1)N} |A_{22}|^{\frac{1}{2}N}}$$

The ratio

$$\lambda = \frac{|A|^{\frac{1}{2}N}}{|a_{11}|^{\frac{1}{2}N} |A_{22}|^{\frac{1}{2}N}}$$

$$\Rightarrow \lambda^{\left(\frac{2}{N}\right)} = \frac{|A|}{|a_{11}||A_{22}|} = 1 - R^2$$

The Critical region is $\lambda < \lambda_0$ where λ_0 is chosen so that

$P(\lambda < \lambda_0 | H_0) = \alpha.$

It is clear that an equivalent test is

$\lambda^{\frac{2}{N}} < \lambda_0^{\frac{2}{N}}$

$\Rightarrow \ \ 1 - \lambda^{\frac{2}{N}} = R^2 > 1 - \lambda_0^{\frac{2}{N}} = R_0^2$

Since

$F = \dfrac{R^2}{1 - R^2} \times \dfrac{N - p}{p - 1},$

we have

$$\frac{R^2}{1 - R^2} \times \frac{N - p}{p - 1} > \frac{R_0^2}{1 - R_0^2} \times \frac{N - p}{p - 1}$$

$\Rightarrow F > F_{p-1,N-p}(\alpha)$

F is a monotone increasing function of R^2, an equivalent test is

If $F > F_{p-1,N-p}(\alpha)$ reject H_0

where $F_{p-1,N-p}(\alpha)$ is chosen so that $\alpha = P(F > F_{p-1,N-p}(\alpha) | H_0)$.

3.6 Distribution of Sample Partial Correlation Coefficient

For finding the distribution of sample partial correlation coefficient, we shall use the following result.

Theorem 3.3.4.: Suppose $y_1, y_2, \ldots, y_m$ are independent with $y_\alpha \sim N(\Gamma w_\alpha, \Phi)$, where w_α is an $r \times 1$ vector. Let

$$G = \sum_{\alpha} y_\alpha w'_\alpha H^{-1}$$

$$= YW'H^{-1}$$

where

$H = \sum_{\alpha=1}^{m} w_\alpha w'_\alpha \qquad ; \{Y = (y_1, \dots, y_m) : p \times m, W = (w_1, \dots, w_m) : r \times m\}$

is non-singular. Then $\sum_{\alpha=1}^{m} y_\alpha y'_\alpha - GHG'$ is distributed as $\sum_{\alpha=1}^{m-r} U_\alpha U'_\alpha$ where U_α are independently distributed each according to $N(0, \Phi)$ and independently of G.

Proof: Let $W = (w_1, w_2, \dots, w_m)$. There exists a non-singular square matrix F such that

$FHF' = I$

or $F'^{-1}H^{-1}F^{-1} = I$

Further, let $P_2 = FW$, then

$$P_2 P'_2 = FWW'F'$$

$$= F \sum_{\alpha=1}^{m} w_\alpha w'_\alpha F'$$

$$= FHF' = I$$

Thus m component rows of P_2 are orthogonal and of unit length. It is possible to find a $(m - r) \times m$ matrix P_1 such that

$$P = \begin{pmatrix} P_1 \\ P_2 \end{pmatrix}$$

is orthogonal. Let $U = YP'$ or $Y = UP$. Then columns of U, say U_α, are independently and normally distributed, each with covariance matrix Φ. The means are given by

$$EU = E(YP')$$

$$= \Gamma WP'$$

$$= \Gamma F^{-1} P_2 (P'_1 \quad P'_2)$$

$$= (0 \quad \Gamma F^{-1})$$

Now

$$\sum_{\alpha=1}^{m} y_\alpha y_\alpha' = \sum_{\alpha=1}^{m} U_\alpha U_\alpha'$$

$$GHG' = (YW'H^{-1})H(YW'H^{-1})'$$

$$= YW'H^{-1}WY'$$

$$= UPP_2'(F^{-1})'H^{-1}F^{-1}P_2P'$$

$$= U\begin{pmatrix} P_1 \\ P_2 \end{pmatrix} P_2'(F^{-1})'H^{-1}F^{-1}P_2(P_1' \quad P_2')U'$$

$$= U\begin{pmatrix} 0 \\ I \end{pmatrix} (F^{-1})'H^{-1}F^{-1}(0 \quad I)U'$$

Since $FHF' = I$, we have $\{(F^{-1})'H^{-1}F^{-1} = I\}$. Hence

$$GHG' = U\begin{pmatrix} 0 \\ I \end{pmatrix} (0 \quad I)U'$$

$$= \sum_{\alpha=m-r+1}^{m} U_\alpha U_\alpha'$$

Thus

$$\sum_{\alpha=1}^{m} y_\alpha y_\alpha' - GHG' = \sum_{\alpha=1}^{m-r} U_\alpha U_\alpha'$$

$U_\alpha \sim N(0, \Phi) \qquad \forall \alpha = 1,2, \dots, m-r$

Hence $\sum_{\alpha=1}^{m} y_\alpha y_\alpha' - GHG'$ is distributed as $\sum_{\alpha=1}^{m-r} U_\alpha U_\alpha'$ where $U_\alpha \sim N(0, \Phi)$.

If $\Gamma = 0$, then $EU = 0$. Hence the matrix GHG' is distributed as $\sum_{\alpha=m-r+1}^{m} U_\alpha U_\alpha'$, where the U_α are independently distributed, each according to $N(0, \Phi)$.

We obtain the distribution of $A_{11.2}$ in the following theorem.

Theorem 3.3.5.: The matrix $A_{11.2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$ is distributed as $\sum_{\alpha=1}^{N-1-(p-q)} U_\alpha U_\alpha'$, where U_α are independently distributed, each according to $N(0, \Sigma_{11.2})$. If $\Sigma_{12} = 0$ then $A_{12}A_{22}^{-1}A_{21}$ is distributed as $\sum_{\alpha=N-(p-q)}^{N-1} U_\alpha U_\alpha'$, .

Proof: Now A is distributed as $\sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha'$, where Z_α are independent, each with distribution $N(0, \Sigma)$.

Let

$$Z_\alpha = \begin{pmatrix} Z_\alpha^{(1)} \\ Z_\alpha^{(2)} \end{pmatrix} \begin{matrix} q \\ p-q \end{matrix}$$

The conditional density of $Z_\alpha^{(1)}$ given $Z_\alpha^{(2)}$ is $N\left(BZ_\alpha^{(2)}, \Sigma_{11.2}\right)$.

Now we apply the previous theorem with

$Z_\alpha^{(1)} = y_\alpha, Z_\alpha^{(2)} = w_\alpha,$

$N - 1 = m, p - q = r$

$B = \Gamma,$

$\Sigma_{11.2} = \Phi,$

$A_{11} = \sum y_\alpha y_\alpha',$

$A_{12}A_{22}^{-1} = G,$

$A_{22} = H.$

We find the conditional distribution of $A_{11} - (A_{12}A_{22}^{-1})A_{22}(A_{22}^{-1}A_{12}') = A_{11.2}$ given $Z_\alpha^{(2)}$ as that of $\sum_{\alpha=1}^{N-1-(p-q)} U_\alpha U_\alpha'$ where U_α are independent, each with distribution $N(0, \Sigma_{11.2})$.

If $\Sigma_{12} = 0 \Rightarrow B = 0$, then $A_{11.2}$ is distributed as $\sum_{\alpha=1}^{N-1-(p-q)} U_\alpha U_\alpha'$ and $A_{12}A_{22}^{-1}A_{21}$ is distributed as $\sum_{\alpha=N-(p-q)}^{N-1} U_\alpha U_\alpha'$.

It follows that the distribution of $r_{ij.q+1,\ldots,p}$ based on N observations is the same as an ordinary correlation coefficient based on $N-(p-q)$ observations with a corresponding population correlation value $\rho_{ij.q+1,\ldots,p}$.

If the cdf of r_{ij} is $F(r|N,\rho_{ij})$ then the cdf of $r_{ij.q+1,\ldots,p}(1\leq i,j\leq q)$ is

$F(r|N-(p-q),\rho_{ij.q+1,\ldots,p})$.

Here r_{ij} and $r_{ij.q+1,\ldots,p}$ are based on sample of size N from a normal distribution.

Example: If all the correlation coefficients in a p-variate normal distribution are equal to $\rho\neq 0$, show that

$$\rho\geq -\frac{1}{p-1}$$

Solution: Given that

$\rho_{ij}=\rho\,,\quad i,j=1,2,\ldots,p\,;i\neq j,$

we have

$$\rho_{ij.k}=\frac{\sigma_{ij}-\sigma_{ik}\sigma_{kk}^{-1}\sigma_{jk}}{(\sigma_{ii}-\sigma_{ik}\sigma_{kk}^{-1}\sigma_{ik})^{1/2}\left(\sigma_{jj}-\sigma_{jk}\sigma_{kk}^{-1}\sigma_{jk}\right)^{1/2}}$$

$$=\frac{\sigma_i\sigma_j\rho_{ij}-\sigma_i\sigma_k\rho_{ik}(\sigma_k\sigma_k)^{-1}\sigma_j\sigma_k\rho_{jk}}{\{\sigma_{ii}-\sigma_i\sigma_k\rho_{ik}(\sigma_k\sigma_k)^{-1}\sigma_i\sigma_k\rho_{ik}\}^{1/2}\left\{\sigma_{jj}-\sigma_j\sigma_k\rho_{jk}(\sigma_k\sigma_k)^{-1}\sigma_j\sigma_k\rho_{jk}\right\}^{1/2}}$$

$$=\frac{\sigma_i\sigma_j\left(\rho_{ij}-\rho_{ik}\rho_{jk}\right)}{\sigma_i\sigma_j\{1-\rho_{ik}^2\}^{1/2}\left\{1-\rho_{jk}^2\right\}^{1/2}}$$

$$=\frac{\rho-\rho^2}{(1-\rho^2)^{\frac{1}{2}}(1-\rho^2)^{\frac{1}{2}}}$$

$$=\frac{\rho}{1+\rho}$$

Thus, every partial correlation coefficient of order 1 is $\frac{\rho}{1+\rho}$. Similarly,

$$\rho_{ij.k1} = \frac{\left(\rho_{ij.1} - \rho_{ik.1}\rho_{jk.1}\right)}{\{1-\rho_{ik.1}^2\}^{1/2}\{1-\rho_{jk.1}^2\}^{1/2}}$$

$$= \frac{\left(\frac{\rho}{1+\rho}\right) - \left(\frac{\rho}{1+\rho}\right)^2}{1-\left(\frac{\rho}{1+\rho}\right)^2}$$

$$= \frac{\rho}{1+2\rho}$$

Thus, every partial correlation coefficient of order 1 is $\frac{\rho}{1+2\rho}$.

The partial correlation coefficient of the highest order in p variate distribution is $(p-2)$. By the method of induction, every partial correlation coefficient of order $(p-2)$ is

$$\frac{\rho}{1+(p-2)\rho}$$

Since $\left|\rho_{ij.(p-1)\ components}\right| \leq 1$, we have

$$-1 \leq \frac{\rho}{1+(p-2)\rho} \leq 1$$

Consider lower limit

$$-1 \leq \frac{\rho}{1+(p-2)\rho}$$

or

$$-1-(p-2)\rho \leq \rho$$

or $\quad -1 \leq \rho - 2\rho + p\rho$

or $\quad -1 \leq (p-1)\rho$

$$\Rightarrow \frac{-1}{(p-1)} \leq \rho$$

3.7 The Matrix of Residual Sum of Squares and Cross Products

The observational equivalent of the effects model is

$$x_{ij} = \bar{x} + (x_i + \bar{x}) + \left(x_{ij} - x_i\right)$$

=overall sample mean + treatment effect + residual (under univariate ANOVA)

After manipulation

$$\sum_{i=1}^{h}\sum_{j=1}^{N_i}(x_{ij}-\bar{x})\,(x_{ij}-\bar{x})' = \sum_{i=1}^{h} n_i\,(x_i-\bar{x})(x_i-\bar{x})' + \sum_{i=1}^{h}\sum_{j=1}^{N_i}(x_{ij}-\bar{x})\,(x_{ij}-x_i)'$$

Where

$\sum_{i=1}^{h}\sum_{j=1}^{N_i}(x_{ij}-\bar{x})\,(x_{ij}-\bar{x})'$ is the total corrected sums of squares and cross products matrix,

$\sum_{i=1}^{h} n_i\,(x_i-\bar{x})(x_i-\bar{x})'$ is the treatment sums of squares and cross products matrix,

$\sum_{i=1}^{h}\sum_{j=1}^{N_i}(x_{ij}-\bar{x})\,(x_{ij}-x_i)'$ is the residual sums of squares and cross products matrix

3.8 Summary

In this unit, we have covered the following points:

1. Define the correlation coefficients.
2. Describe the sample multiple and partial correlations.
3. Derive the central distributions of multiple correlation coefficients.
4. Derive the central distribution of partial correlation coefficients.

3.9 Self-Assessment Exercises

1. Let $X_1, X_2, \dots, X_N$ be independently distributed with $X_\alpha \sim N_p(\mu_\alpha, \Sigma)$. Let $C = (C_{\alpha\beta})$ be $N \times N$ orthogonal matrix. Then prove that $y_\alpha = \sum_{\beta=1}^{N} C_{\alpha\beta} X_\beta$ is distributed according to $N(\theta_\alpha, \Sigma)$, where $\theta_\alpha = \sum_{\beta=1}^{N} C_{\alpha\beta}\mu_\beta$ and $Y_1, Y_2, \dots, Y_N$ are independently distributed. Also, find the sufficient statistic for Σ if μ is known.
2. Derive the null distribution of sample multiple correlation coefficients.
3. Derive the null distribution of sample partial correlation coefficients.
4. If all the correlation coefficients in a p-variate normal distribution are equal to $\rho \neq 0$, show that

$$\rho^2_{1(2,3,\dots,p)} = \frac{(p-1)\rho^2}{1+(p-2)\rho}$$

3.10 References

- Johnson, R. A., Wichern, D. W. (2019): Applied Multivariate Statistical Analysis. United Kingdom: Pearson.
- Muirhead, R. J. (2009): Aspects of Multivariate Statistical Theory. Germany: Wiley.
- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Giri Narayan C. (1995): Multivariate Statistical Analysis
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.
- Mardia, K. V., Bibby, J. M., Kent, J. T. (1979): Multivariate Analysis. United Kingdom: Academic Press.
- Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York.

3.11 Further Readings

- Kotz, S., Balakrishnan, N. and Johnson, N.L.: Continuous Multivariate Distribution Models and Applications (Second Edition). Volume 1, Wiley - Inter science, New York.
- Khatri, C. G.: Multivariate Analysis.

- Mardia, K. V.: Multivariate Analysis.
- Seber, G.A.F.: *Multivariate Observations.* Wiley, New York.
- Rencher, Alvin C.: Multivariate Statistical Inference and Applications. John Wiley. New York, New York.

UNIT: 4 WISHART DISTRIBUTION

Structure

4.1 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).

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.

4.2 Objectives

After studying this unit, you should be able to:

- Discuss the Wishart distribution.
- Define the distribution of mean and sample covariance.
- Derive the conditional distribution and marginal distribution of Wishart distribution.
- Discuss the Cochran theorem.
- Also discuss the distribution of characteristic roots and vectors of the Wishart matrix.

4.3 WISHART DISTRIBUTION

Theorem 4.3.1.: Suppose p-components $Z_1, Z_2, \dots, Z_n$ $(n \geq p)$ are independent, each distributed according to $N(0, \Sigma)$. Then the p.d.f. of $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ is

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

for A positive definite and zero otherwise.

This distribution is called Wishart distribution and is a generalization of χ^2- distribution.

By pdf of $\mathrm{A} = \sum_{\alpha=1}^{\mathrm{n}} \mathrm{Z}_\alpha \mathrm{Z}_\alpha'$ we mean the joint distribution of $p(p+1)/2$ different elements of A.

Proof: We shall use the following results:

If scalars U_α are independently distributed and $U_\alpha \sim N(\Gamma w_\alpha, \phi)$ where, w_α is the r component vectors and $G = \sum_{\alpha=1}^{m} Y_\alpha w_\alpha H^{-1}$ where, $H = \sum_{\alpha=1}^{m} w_\alpha\, w_\alpha'$ is a non-singular matrix then $\sum_{\alpha=1}^{m} U_\alpha\, U_\alpha' - GHG' \sim \sum_{\alpha=1}^{m-r} U_\alpha\, U_\alpha'$ where, U_α are independently distributed according to $N(0, \phi)$ and independent of G.

Suppose scalar U_α are independently distributed and it is follows $U_\alpha \sim N(\Gamma w_\alpha, \phi)$ then $\sum_{\alpha=1}^{n} U_\alpha^2 - \sum_{\alpha=1}^{n} U_\alpha w_\alpha'(\sum w_\alpha w_\alpha')^{-1} \sum_{\alpha=1}^{n} U_\alpha w_\alpha \sim \sum_{\alpha=1}^{n-q} V_\alpha^2$ where, q is the number of components of w_α, V_α are independent and $V_\alpha \sim N(0, \phi)$ and V_α are independent of $\sum_{\alpha=1}^{n} U_\alpha w_\alpha$ In particular if $\phi = 1$, then $t = \sum_\alpha U_\alpha^2 - \sum_\alpha U_\alpha w_\alpha'(\sum w_\alpha w_\alpha')^{-1} \sum U_\alpha w_\alpha$ has the χ^2- distribution with $(n-q)$ degrees of freedom and pdf is

$$f(t) = \frac{e^{-t/2} t^{\{(n-q)/2\}-1}}{2^{(n-q)/2}\Gamma\left(\frac{n-q}{2}\right)} \;\; ; \;\; t \geq 0$$

Further $\sum U_\alpha w_\alpha \sim$ Normal distribution.

If $\Gamma = 0$ then $E(\sum U_\alpha w_\alpha) = 0$ and variance-covariance matrix is

$$E\left(\sum_{\alpha=1}^{n} U_\alpha\, w_\alpha \sum_{\beta=1}^{n} U_\beta\, w_\beta'\right) = \sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} w_\alpha\, w_\beta'\, \delta_{\alpha\beta} = \sum_{\alpha=1}^{n} w_\alpha\, w_\alpha'$$

$\delta_{\alpha\beta} = \begin{cases} 1 & if\ \alpha = \beta \\ 0 & if\ \alpha \neq \beta \end{cases}$: Kronecker delta

Hence

$$\sum U_\alpha w_\alpha \sim N(0, \sum_\alpha w_\alpha w_\alpha')$$

Now

$$A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha', \qquad Z_\alpha \sim N(0, \Sigma)$$

Since Σ is positive definite there exist a non-singular triangular matrix C such that

$$C\Sigma C' = I$$

$$\Rightarrow CC' = \Sigma^{-1}$$

Now we use the transformation

$$B = CAC'$$

$$= C\left(\sum Z_\alpha Z'_\alpha\right)C'$$

$$= \sum U_\alpha U'_\alpha$$

Here

$$U_\alpha = CZ_\alpha \sim N(0, I)$$

Jacobean of the transformation is

$$|C|^{p+1} = |\Sigma|^{-\frac{1}{2}(p+1)}$$

$$b_{ij} = \sum_{k,l} c_{ik} a_{kl} c_{jl} \qquad \left(c_{ij} = 0, \text{if } i > j\right)$$

Since C is a triangular matrix, its determinant is equal to the product of its diagonal elements, i.e.,

$$|C| = \prod_{i=1}^{p} c_{ii}$$

Further, the partial derivatives of $b_{11}, \dots, b_{1p}, b_{22}, \dots, b_{2p}, \dots, b_{pp}$ with respect to $a_{11}, \dots, a_{1p}$, $a_{22}, \dots, a_{2p}, \dots, a_{pp}$ are given by

$$\frac{\partial b_{ij}}{\partial a_{kk}} = c_{ik}c_{jk}; \frac{\partial b_{ij}}{\partial a_{kl}} = c_{ik}c_{jl} + c_{il}c_{jk} \qquad l \neq k$$

Hence, the Jacobian of the transformation $J(B \to A)$ is obtained as

$$J(B \to A) = \begin{bmatrix} \frac{\partial b_{11}}{\partial a_{11}} & \cdots & \frac{\partial b_{11}}{\partial a_{1p}} & \frac{\partial b_{11}}{\partial a_{22}} & \cdots & \frac{\partial b_{11}}{\partial a_{2p}} & \frac{\partial b_{11}}{\partial a_{33}} & \cdots & \frac{\partial b_{11}}{\partial a_{3p}} & \cdots & \frac{\partial b_{11}}{\partial a_{pp}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial b_{1p}}{\partial a_{11}} & \cdots & \frac{\partial b_{1p}}{\partial a_{1p}} & \frac{\partial b_{1p}}{\partial a_{22}} & \cdots & \frac{\partial b_{1p}}{\partial a_{2p}} & \frac{\partial b_{1p}}{\partial a_{33}} & \cdots & \frac{\partial b_{1p}}{\partial a_{3p}} & \cdots & \frac{\partial b_{1p}}{\partial a_{pp}} \\ \frac{\partial b_{22}}{\partial a_{11}} & \cdots & \frac{\partial b_{22}}{\partial a_{1p}} & \frac{\partial b_{22}}{\partial a_{22}} & \cdots & \frac{\partial b_{22}}{\partial a_{2p}} & \frac{\partial b_{22}}{\partial a_{33}} & \cdots & \frac{\partial b_{22}}{\partial a_{3p}} & \cdots & \frac{\partial b_{22}}{\partial a_{pp}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial b_{2p}}{\partial a_{11}} & \cdots & \frac{\partial b_{2p}}{\partial a_{1p}} & \frac{\partial b_{2p}}{\partial a_{22}} & \cdots & \frac{\partial b_{2p}}{\partial a_{2p}} & \frac{\partial b_{2p}}{\partial a_{33}} & \cdots & \frac{\partial b_{2p}}{\partial a_{3p}} & \cdots & \frac{\partial b_{2p}}{\partial a_{pp}} \\ \frac{\partial b_{33}}{\partial a_{11}} & \cdots & \frac{\partial b_{33}}{\partial a_{1p}} & \frac{\partial b_{33}}{\partial a_{22}} & \cdots & \frac{\partial b_{33}}{\partial a_{2p}} & \frac{\partial b_{33}}{\partial a_{33}} & \cdots & \frac{\partial b_{33}}{\partial a_{3p}} & \cdots & \frac{\partial b_{33}}{\partial a_{pp}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial b_{3p}}{\partial a_{11}} & \cdots & \frac{\partial b_{3p}}{\partial a_{1p}} & \frac{\partial b_{3p}}{\partial a_{22}} & \cdots & \frac{\partial b_{3p}}{\partial a_{2p}} & \frac{\partial b_{3p}}{\partial a_{33}} & \cdots & \frac{\partial b_{3p}}{\partial a_{3p}} & \cdots & \frac{\partial b_{3p}}{\partial a_{pp}} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{\partial b_{pp}}{\partial a_{11}} & \cdots & \frac{\partial b_{pp}}{\partial a_{1p}} & \frac{\partial b_{pp}}{\partial a_{22}} & \cdots & \frac{\partial b_{pp}}{\partial a_{2p}} & \frac{\partial b_{pp}}{\partial a_{33}} & \cdots & \frac{\partial b_{pp}}{\partial a_{3p}} & \cdots & \frac{\partial b_{pp}}{\partial a_{pp}} \end{bmatrix}$$

$$= \begin{bmatrix} c_{11}^2 & 2c_{11}c_{12} & \cdots & 2c_{12}c_{1p} & c_{12}^2 & \cdots & c_{1p}^2 \\ 0 & c_{11}c_{22} & \cdots & c_{11}c_{2p} & c_{12}c_{22} & \cdots & c_{1p}c_{2p} \\ 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & c_{11}c_{pp} & 0 & \cdots & c_{1p}c_{pp} \\ 0 & 0 & \cdots & 0 & c_{22}^2 & \cdots & c_{2p}^2 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & c_{pp}^2 \end{bmatrix}$$

$$= \prod_{j=1}^{p} c_{jj}^{p+1}$$

$$= |C|^{p+1}$$

$$= |\Sigma|^{-\frac{1}{2}(p+1)}$$

Let

$$B_{ii} = \left(\left(b_{jk}\right)\right) \quad j,k = i,\dots,p$$

$$b_{(i)} = (b_{i,i+1}, b_{i,i+2}, \dots, b_{i,i+p})$$

$$B_{ii} = \begin{pmatrix} b_{ii} & b_{(i)} \\ b'_{(i)} & B_{i+1,i+1} \end{pmatrix}$$

$$b_{ii.i+1,\dots,p} = b_{ii} - b_{(i)} B^{-1}_{i+1,i+1} b'_{(i)}$$

$$U^{(j)\prime}_{\alpha} = \left(U_{j\alpha}, U_{j+1,\alpha}, \dots, U_{p\alpha}\right)$$

Then

$$B_{ii} = \sum_{\alpha} U^{(i)}_{\alpha} U^{(i)\prime}_{\alpha} = \begin{pmatrix} \sum_{\alpha} U^2_{i\alpha} & \sum_{\alpha} U_{i\alpha} U^{(i+1)\prime} \\ \sum_{\alpha} U_{i\alpha}\, U^{(i+1)} & \sum U^{(i+1)}\, U^{(i+1)\prime} \end{pmatrix}$$

$$b_{ii} = \sum_{\alpha} U^2_{i\alpha}$$

$$b_{(i)} = \sum_{\alpha} U_{i\alpha} U^{(i+1)\prime}$$

$$B_{i+1,i+1} = \sum U^{(i+1)}\, U^{(i+1)\prime}$$

Hence

$$b_{ii.i+1,\dots,p} = \sum_{\alpha} U^2_{i\alpha} - \left\{\sum_{\alpha} U_{i\alpha} U^{(i+1)\prime}\right\}\left\{\sum U^{(i+1)}\, U^{(i+1)\prime}\right\}^{-1}\left\{\sum_{\alpha} U_{i\alpha}\, U^{(r+1)}\right\}$$

Since $\Sigma = I$, the set $(U_{i1}, U_{i2}, \dots, U_{in})$ is distributed independently of $\left(U_{j1}, U_{j2}, \dots, U_{jn}\right), j \neq i$. Therefore, conditional on $U^{(i+1)}$, each element $U_{j\alpha}$ $(j \neq \alpha, \alpha = 1,2, \dots, n)$, is distributed according to $N(0,1)$.

Hence given $U^{(i+1)}, b_{ii.i+1,\dots,p}$ has a $\chi^2 -$ distribution with $n-(p-i)$ degrees of freedom and distributed independently of $b_{(i)}$ $(= \sum_\alpha U_{i\alpha} U^{(i+1)'})$.

Given $U^{(i+1)}$, $b_{(i)}$ is conditionally distributed according to $N(0, B_{i+1,i+1})$.

Observe that the conditional distribution depends on $U_\alpha^{(i+1)}$ only through $B_{i+1,i+1}$, i.e., the conditional density is of the form $f_i\left(b_{ii.i+1,\dots,p}, b_{(i)} | U_\alpha^{(i+1)}\right) = f_i\left(b_{ii.i+1,\dots,p}, b_{(i)} |\, B_{i+1,i+1}\right)$. Thus, the joint density of $b_{11.2,\dots,p}, b_{(1)}, b_{22.3,\dots,p}, b_{(2)}, \dots, b_{p-1,p-1.p}, b_{(p-1)}, b_{pp}$ is

$$\frac{b_{pp}^{\frac{1}{2}n-1} e^{-\frac{1}{2}b_{pp}}}{2^{\frac{1}{2}n}\Gamma\left(\frac{1}{2}n\right)} \prod_{i=1}^{p-1} \left\{ \frac{b_{ii.i+1,\dots,p}^{\frac{1}{2}[n-(p-i)]-1} e^{-\frac{1}{2}b_{ii.i+1,\dots,p}}}{2^{\frac{1}{2}[n-(p-i)]}\Gamma\left\{\frac{1}{2}[n-(p-i)]\right\}} \times \frac{e^{-\frac{1}{2}b_{(i)}B_{i+1,i+1}^{-1}b_{(i)}'}}{(2\pi)^{\frac{1}{2}(p-i)}\left|\, B_{i+1,i+1}\right|^{\frac{1}{2}}} \right\} \tag{4.1}$$

We find the density of $b_{11}, b_{(1)}, b_{22}, b_{(2)}, \dots, b_{pp}$ by substituting in (4.1)

$b_{ii.i+1,\dots,p} = b_{ii} - b_{(i)}B_{i+1,i+1}^{-1}b_{(i)}'$.

Jacobian of the transformation is one. The exponent of e in (4.1) is

$$-\frac{1}{2}\left[b_{pp} + \sum_{i=1}^{p-1} b_{ii.i+1,\dots,p} + \sum_{i=1}^{p-1} b_{(i)}B_{i+1,i+1}^{-1}b_{(i)}'\right]$$

$$= -\frac{1}{2}\left[b_{pp} + \sum_{i=1}^{p-1} b_{ii} - \sum_{i=1}^{p-1} b_{(i)}B_{i+1,i+1}^{-1}b_{(i)}' + \sum_{i=1}^{p-1} b_{(i)}B_{i+1,i+1}^{-1}b_{(i)}'\right]$$

$$= -\frac{1}{2}\left[b_{pp} + \sum_{i=1}^{p-1} b_{ii}\right]$$

$$= -\frac{1}{2}trB$$

Again

$$|A| = |A_{11} - A_{12}A_{22}^{-1}A_{21}||A_{22}|$$

$$B_{ii} = \begin{vmatrix} b_{ii} & b_{(i)} \\ b'_{(i)} & B_{i+1,i+1} \end{vmatrix}, \text{ where } B_{ii} = A, b_{ii} = A_{11}, b_{(i)} = A_{12}, B_{i+1,i+1} = A_{22}$$

$$b_{ii.i+1,\dots,p} = b_{ii} - b_{(i)}B_{i+1,i+1}^{-1}b'_{(i)} = \frac{\begin{vmatrix} b_{ii} & b_{(i)} \\ b'_{(i)} & B_{i+1,i+1} \end{vmatrix}}{\left|B_{i+1,i+1}\right|} = \frac{|B_{ii}|}{\left|B_{i+1,i+1}\right|}$$

We find

$$b_{pp}\prod_{i=1}^{p-1} b_{ii.i+1,\dots,p} = b_{pp}\prod_{i=1}^{p-1}\frac{|B_{ii}|}{\left|B_{i+1,i+1}\right|}$$

$$= b_{pp}\left[\frac{|B_{11}|}{|B_{22}|}\frac{|B_{22}|}{|B_{33}|}\dots\frac{\left|B_{p-1,p-1}\right|}{\left|B_{pp}\right|}\right]$$

$$= b_{pp}\left[\frac{|B_{11}|}{\left|B_{pp}\right|}\right] \qquad \left(\text{Since } B_{pp} = b_{pp}\right)$$

$$= |B_{11}| = |B|$$

Then

$$b_{pp}^{\frac{1}{2}n-1}\prod_{i=1}^{p-1}\frac{b_{ii.i+1,\dots,p}^{\frac{1}{2}[n-(p-i)]-1}}{\left|B_{i+1,i+1}\right|^{\frac{1}{2}}}$$

$$= \left[b_{pp}\prod_{i=1}^{p-1} b_{ii.i+1,\dots,p}\right]^{\frac{1}{2}(n-p-1)}\left[b_{pp}^{\frac{1}{2}(p-1)}\prod_{i=1}^{p-1}\frac{b_{ii.i+1,\dots,p}^{\frac{1}{2}(i-1)}}{\left|B_{i+1,i+1}\right|^{\frac{1}{2}}}\right]$$

$$= |B|^{\frac{1}{2}(n-p-1)}b_{pp}^{\frac{1}{2}(p-1)}\prod_{i=1}^{p-1}\frac{|B_{ii}|^{\frac{1}{2}(i-1)}}{\left|B_{i+1,i+1}\right|^{\frac{i}{2}}}$$

$$= |B|^{\frac{1}{2}(n-p-1)} b_{pp}^{\frac{1}{2}(p-1)} \left[\frac{1}{|B_{22}|^{\frac{1}{2}}} \frac{|B_{22}|^{\frac{1}{2}}}{|B_{33}|^{\frac{1}{2}}} \cdots \frac{\left|B_{p-1,p-1}\right|^{\frac{1}{2}(p-1+1)}}{\left|B_{pp}\right|^{\frac{p-1}{2}}} \right]$$

$$= |B|^{\frac{1}{2}(n-p-1)} b_{pp}^{\frac{1}{2}(p-1)} \left[\frac{|1|^{\frac{1}{2}(p)}}{\left|B_{pp}\right|^{\frac{(p-1)}{2}}} \right]$$

$$= |B|^{\frac{1}{2}(n-p-1)}$$

The power of π is

$$\frac{\sum_{i=1}^{p-1}(p-i)}{2} = \frac{1}{2}[(p-1)+(p-2)+\cdots+1] = \frac{p(p-1)}{4}$$

The power of 2 is

$$\frac{1}{2}\left[n + \sum_{i=1}^{p-1}[n-(p-i)] + \sum_{i=1}^{p-1}(p-i) \right] = \frac{1}{2}\left[n + \sum_{i=1}^{p-1} n \right] = \frac{1}{2}[n+(p-1)n] = \frac{np}{2}$$

The power of Γ is

$$\Gamma\left(\frac{1}{2}n\right) \prod_{i=1}^{p-1} \Gamma\left\{\frac{1}{2}[n-(p-i)]\right\} = \Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{n-p+1}{2}\right) \Gamma\left(\frac{n-p+2}{2}\right) \ldots \Gamma\left(\frac{n-1}{2}\right)$$

$$= \prod_{i=1}^{p-1} \Gamma\left\{\frac{1}{2}[n-i+1]\right\}$$

Thus, the density of $b_{11}, \ldots, b_{1p}, b_{22}, \ldots, b_{2p}, \ldots, b_{pp}$ is

$$\frac{|B|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}trB}}{2^{\frac{1}{2}np} \pi^{p(p-1)/4} \prod_{i=1}^{p-1} \Gamma\left\{\frac{1}{2}[n-i+1]\right\}}$$

Therefore the p.d.f. of $A = B\Sigma$ is

$W(A|\Sigma, n)$

$$= \begin{cases} \dfrac{|A|^{\frac{1}{2}(n-p-1)} e^{-\frac{1}{2}trA\Sigma^{-1}}}{2^{\frac{1}{2}np}\prod^{p(p-1)/4}|\Sigma|^{\frac{1}{2}(n-p-1)}\prod_{i=1}^{p-1}\Gamma\left\{\frac{1}{2}[n-i+1]\right\}} & \text{if } A \text{ is positive definite} \\ 0 & \text{otherwise} \end{cases}$$

Hence, if the p-component vectors $X_1, \dots, X_p$ $(N > p)$ are independent, each with the distribution $N(\mu, \Sigma)$, then the density of

$A = \sum_{\alpha=1}^{N}(X_\alpha - \overline{X})(X_\alpha - \overline{X})'$ is $W(A|\Sigma, N-1)$.

Remark: Let the symmetric matrix B be transformed into the symmetric matrix A by $B = CAC'$, where C is a non-singular triangular matrix (i.e. $c_{ij} = 0$ for $i > j$). Then the Jacobian of the transformation is $mod|C|^{(p+1)}$.

Here $W(A|\Sigma, n)$ denotes the $p.d.f.$ and $W(\Sigma, n)$ denotes the associated distribution

Theorem 4.3.2.: Let $X_1, X_2, \dots, X_n (n \geq p+1)$ be distributed independently, each according to $N(\mu, \Sigma)$, then the distribution of $S = \frac{1}{N-1}A$ is $W\left(\frac{1}{N-1}\Sigma, N-1\right)$.

Proof: Let $n = N - 1$

$$S = \frac{1}{n}A = \frac{1}{n}\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha' = \sum_{\alpha=1}^{n}\left(\frac{1}{\sqrt{n}}Z_\alpha\right)\left(\frac{1}{\sqrt{n}}Z_\alpha\right)' = \sum_{\alpha=1}^{n} U_\alpha U_\alpha'$$

$U_\alpha \sim N(0, \frac{1}{n}\Sigma)$

$U_\alpha' s$ are independent.

Hence, $S = \sum_{\alpha=1}^{n} U_\alpha U_\alpha' \sim W\left(\frac{1}{n}\Sigma, n\right)$, with $n = N - 1$ and we follow the theorem.

4.4 Some Properties of the Wishart Distribution

There are some following properties:

4.4.1 Characteristic Function

Theorem 4.4.1.: Let $A = \sum_{\alpha=1}^{n} Z_\alpha Z'_\alpha = \left((a_{jk})\right)$ $\qquad a_{jk} = a_{kj}$

Here, $Z_\alpha \sim N(0, \Sigma)$. Then the characteristic function of $a_{11}, \ldots, a_{pp}, 2a_{12}, \ldots, 2a_{p-1,p}$ is given by

$$E\{exp[i\ tr(A\Theta)]\} = \frac{|\Sigma^{-1}|^{\frac{1}{2}n}}{|\Sigma^{-1} - 2i\Theta|^{\frac{1}{2}n}} = |1 - 2i\Theta\Sigma^{-1}|^{\frac{-n}{2}}$$

where Θ is a real symmetric matrix of order $p \times p$.

Proof: The characteristic function of A is given as

$$\phi_A(\Theta) = E\{exp[i\ tr(A\Theta)]\}$$

$$= E\left\{exp\left[i\ tr\left(\sum_{\alpha=1}^{n} Z_\alpha Z'_\alpha\ \Theta\right)\right]\right\}$$

$$= E\left\{exp\left(i\sum_{\alpha=1}^{n} Z_\alpha \Theta Z'_\alpha\right)\right\}$$

$$= \prod_{\alpha=1}^{n} E\ exp(iZ_\alpha \Theta Z'_\alpha) = [E\ exp(iZ_\alpha \Theta Z'_\alpha)]^n \qquad (4.2)$$

Notice that Z_α are independently and identically distributed $Z_\alpha \sim N(0, \Sigma)$.

There exists a non-singular matrix C such that

$$C'\Sigma^{-1}C = I$$

$$C'\Theta C = D$$

D is a diagonal matrix.

Let $Z = Cy$

Then

$$E[\exp(iZ'\Theta Z)] = E[\exp(iy'Dy)] = E\left[\exp\left(i\sum_{j=1}^{p} d_{jj}y_j^2\right)\right] = \prod_{j=1}^{p} E\left[\exp\left(id_{jj}y_j^2\right)\right]$$

Since $Z \sim N(0, \Sigma)$

$Y \sim N(0, (C'\Sigma^{-1}\mathrm{C})^{-1})$

$Y \sim N(0, I)$

Thus $y_j \sim N(0, I)$ for every $j = 1, \ldots, p$ and $y_j^2 \sim \chi^2$ with one d. f.

Therefore

$$E[\exp(iZ'\Theta Z)] = \prod_{j=1}^{p} \left(1 - 2id_{jj}\right)^{-\frac{1}{2}}$$

$$= |I - 2iD|^{-\frac{1}{2}} = |C'\Sigma^{-1}C - 2iC'\Theta C|^{-\frac{1}{2}}$$

$$= |C'|^{-\frac{1}{2}}|\Sigma^{-1} - 2i\Theta|^{-\frac{1}{2}}|C|^{-\frac{1}{2}} = |C'C|^{-\frac{2}{2}}|\Sigma^{-1} - 2i\Theta|^{-\frac{1}{2}}$$

$$= |C'C|^{-1}|\Sigma^{-1} - 2i\Theta|^{-\frac{1}{2}}$$

$$= |\Sigma|^{-1}|\Sigma^{-1} - 2i\Theta|^{-\frac{1}{2}}$$

Hence from (4.2)

$E \exp[i(trA\Theta)]$

$$= \frac{|\Sigma^{-1}|^{\frac{1}{2}n}}{|\Sigma^{-1} - 2i\Theta|^{\frac{1}{2}n}}$$

$$= |I - 2i\Theta\Sigma|^{-\frac{1}{2}n}$$

Note: We can obtain the moments of the elements of A either using the characteristic function or from the original normal distribution.

$$Ea_{ij} = E\sum_{\alpha=1}^{n} Z_{i\alpha}Z_{j\alpha}$$

$$= \sum_{\alpha=1}^{n} \sigma_{ij}$$

$$= n\sigma_{ij}$$

or $E(A) = n\Sigma$

$$Ea_{ij}a_{kl} = \sum_{\alpha,\beta=1}^{n} E\left(Z_{i\alpha}Z_{j\alpha}Z_{k\beta}Z_{l\beta}\right)$$

$$= \sum_{\alpha=1}^{n} E\left(Z_{i\alpha}Z_{j\alpha}Z_{k\alpha}Z_{l\alpha}\right) + \sum_{\alpha,\beta,\alpha\neq\beta} E\left(Z_{i\alpha}Z_{j\alpha}Z_{k\beta}Z_{l\beta}\right)$$

$$= n\left(\sigma_{ij}\sigma_{kl} + \sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{kj}\right) + n(n-1)\sigma_{ij}\sigma_{kl}$$

$$= n^2\sigma_{ij}\sigma_{kl} + n\sigma_{ik}\sigma_{jl} + n\sigma_{il}\sigma_{kj}$$

Hence

$$E\left(a_{ij} - Ea_{ij}\right)(a_{kl} - Ea_{kl})$$

$$= n(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk})$$

$$E\left(a_{ij} - Ea_{ij}\right)^2 = n\left(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}\right) \qquad \text{for } l = i, k = j$$

4.4.2 Reproductive Property of Wishart Distribution

Theorem 4.4.2.: If the A_1 and A_2 are two independent Wishart matrices and it follow $W_p(\Sigma, n_1)$ and $W_p(\Sigma, n_2)$, respectively, then sum of $(A_1 + A_2) \sim W_p(\Sigma, n_1 + n_2)$.

Proof: If $A_1 \sim W_p(\Sigma, n_1)$ then the characteristic function of A_1 is $\phi_{A_1}(U) = |I - 2iU\Sigma|^{-n_1/2}$.

Similarly, $A_2 \sim W_p(\Sigma, n_2)$ implies that $\phi_{A_2}(U) = |I - 2iU\Sigma|^{-n_2/2}$.

Further A_1 and A_2 are independently distributed. Therefore

$$\phi_{A_1+A_2}(U) = \phi_{A_1}(U)\phi_{A_2}(U)$$

$= |I - 2i\Theta\Sigma|^{-(n_1+n_2)/2}$

Hence $(A_1 + A_2) \sim W_p(\Sigma, n_1 + n_2)$.

Note: The Wishart distribution is a multivariate extension of χ^2-distribution.

If $M \sim W_1(n, \sigma^2)$ then $\frac{1}{\sigma^2} M = \chi_n^2$.

Theorem 4.4.3.: If the $A_i (i = 1, \dots, q)$ are independently distributed with $A_i \sim W(\Sigma, n_i), i = 1,2, \dots, q$, then

$$A = \sum_{i=1}^{q} A_i$$

is distributed according to $W(\Sigma, n)$, $n = \sum_{i=1}^{q} n_i$.

Proof: A_1 is distributed as $\sum_{\alpha=1}^{n_1} Z_\alpha Z_\alpha'$

A_2 is distributed as $\sum_{\alpha=n_1+1}^{n_1+n_2} Z_\alpha Z_\alpha'$

A_q is distributed as $\sum_{\alpha=n_1+\cdots+n_{q-1}}^{n_1+\cdots+n_q} Z_\alpha Z_\alpha'$

where $Z_\alpha \sim N(0, \Sigma)$ are Z_α are independently distributed.

Hence $A = \sum_{i=1}^{q} A_i$ is distributed according to $\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ where $n = \sum_{i=1}^{q} n_i$, $Z_\alpha \sim N(0, \Sigma)$ and Z_α are independent.

Theorem 4.4.4.: If $A \sim W(\Sigma, n)$ and L is any $(p \times 1)$ vector. Prove that $\frac{L'AL}{L'\Sigma L} \sim \chi_{(n)}^2$.

Proof: Let $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ and we write $v_\alpha = \sum_{\alpha=1}^{n} L' Z_\alpha$. Obviously v_α is a scalar so that $v_\alpha = v_\alpha'$. Then

$$L'AL = L' \left(\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha' \right) L$$

$$= \left(\sum_{\alpha=1}^{n} L' Z_\alpha \, (L' Z_\alpha)' \right)$$

$$= \sum_{\alpha=1}^{n} v_\alpha v_\alpha'$$

$$= \sum_{\alpha=1}^{n} v_\alpha^2$$

Where

Since $Z_\alpha \sim N(0, \Sigma)$, we have $v_\alpha \sim N(0, L'\Sigma L)$. Therefore

$$\frac{v_\alpha}{(L'\Sigma L)^{1/2}} \sim N(0, I)$$

or $\frac{\sum_{\alpha=1}^{n} v_\alpha^2}{(L'\Sigma L)} \sim \chi^2_{(n)}$

or $\frac{(L'\mathrm{A}L)}{(L'\Sigma L)} \sim \chi^2_{(n)}$.

Hence we follow the result.

4.4.3 Marginal Distribution

Let $A \sim W(\Sigma, n)$. Now we derive the marginal density function of any arbitrary set of elements of A using the following two theorems:

Theorem4.4.5.: Let

$$\mathrm{A} = \begin{matrix} \begin{pmatrix} \mathrm{A}_{11} & \mathrm{A}_{12} \\ \mathrm{A}_{21} & \mathrm{A}_{22} \end{pmatrix} & \begin{matrix} q \\ p-q \end{matrix} \\ \begin{matrix} q & p-q \end{matrix} & \end{matrix}$$

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

If $A \sim W(\Sigma, n)$ then $\mathrm{A}_{11} \sim W(\Sigma_{11}, n)$.

Proof: We can write

$$A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$$

where Z_α are independently distributed each according to $N(0, \Sigma)$. Let

$$Z_\alpha = \begin{pmatrix} Z_\alpha^{(1)} \\ Z_\alpha^{(2)} \end{pmatrix} p - q$$

$$\mathrm{A}_{11} = \sum_{\alpha=1}^{n} Z_\alpha^{(1)} Z_\alpha^{(1)\prime}$$

where $Z_\alpha^{(1)} \sim N(0, \Sigma_{11})$ and $Z_\alpha^{(1)}$ are mutually independently distributed.

Hence $\mathrm{A}_{11} \sim W(\Sigma_{11}, n)$.

Theorem 4.4.6.: Let A and Σ be partitioned into $p_1, \ldots, p_q$ rows and columns as

$$\mathrm{A} = \begin{bmatrix} \mathrm{A}_{11} & \mathrm{A}_{12} & \cdots & \mathrm{A}_{1q} \\ \mathrm{A}_{21} & \mathrm{A}_{22} & \cdots & \mathrm{A}_{2q} \\ \vdots & \vdots & \cdots & \vdots \\ \mathrm{A}_{q1} & \mathrm{A}_{q2} & \cdots & \mathrm{A}_{qq} \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1q} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2q} \\ \vdots & \vdots & \cdots & \vdots \\ \Sigma_{q1} & \Sigma_{q2} & \cdots & \Sigma_{qq} \end{bmatrix}$$

If $\Sigma_{ij} = 0$ for $i \neq j$ and if $A \sim W(\Sigma, n)$ then $\mathrm{A}_{11}, \mathrm{A}_{22}, \ldots, \mathrm{A}_{qq}$ are independent and $\mathrm{A}_{ii} \sim W(\Sigma_{ii}, n)$.

Proof: Let $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$

where Z_α are independent and each distributed according to $N(0, \Sigma)$. Let

$$Z_\alpha = \begin{bmatrix} Z_\alpha^{(1)} \\ Z_\alpha^{(2)} \\ \vdots \\ Z_\alpha^{(q)} \end{bmatrix}$$

Since $\Sigma_{ij} = 0$, hence $Z_\alpha^{(i)}$ are independent of $Z_\alpha^{(j)}$ for $i \neq j$. Then $A_{ii} = \sum_{\alpha=1}^{n} Z_\alpha^{(i)} Z_\alpha^{(i)\prime}$ are independent of $A_{jj} = \sum_{\alpha=1}^{n} Z_\alpha^{(j)} Z_\alpha^{(j)\prime}$, and distributed as $W(\Sigma_{jj}, n)$.

4.4.4 Conditional Distribution

If $A \sim W(\Sigma, n)$,

$$A = \begin{matrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{matrix} q \\ p-q \end{matrix} \\ q \quad p-q \end{matrix} \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Let $A_{11.2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$ and $\Sigma_{11.2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Then the conditional distribution of $A_{11.2}|A_{22}$.

$A_{11.2} \sim W(\Sigma_{11.2}, n - (p-q))$

$$A_{11.2} \sim \sum_{\alpha=1}^{n-(p-q)} U_\alpha U_\alpha'$$

where U_α are independent and each distributed according to $N(0, \Sigma_{11.2})$.

Proof: Let $A \sim W(\Sigma, n)$. We consider $A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$, where Z_α are independent and each distributed according to $N(0, \Sigma)$. Let

$$Z_\alpha = \begin{matrix} \begin{pmatrix} Z_\alpha^{(1)} \\ Z_\alpha^{(2)} \end{pmatrix} \begin{matrix} q \\ p-q \end{matrix} \end{matrix}$$

$$A = \begin{bmatrix} \sum_{\alpha=1}^{n} Z_\alpha^{(1)} Z_\alpha^{(1)\prime} & \sum_{\alpha=1}^{n} Z_\alpha^{(1)} Z_\alpha^{(2)\prime} \\ \sum_{\alpha=1}^{n} Z_\alpha^{(2)} Z_\alpha^{(1)\prime} & \sum_{\alpha=1}^{n} Z_\alpha^{(2)} Z_\alpha^{(2)\prime} \end{bmatrix} = \begin{bmatrix} WW' & WY' \\ YW' & YY' \end{bmatrix}$$

where

$$W = \sum_{\alpha=1}^{n} Z_\alpha^{(1)} = \left[Z_1^{(1)} \quad Z_2^{(1)} \quad \cdots \quad Z_\alpha^{(1)}\right]_{q\times n},$$

$$Y = \sum_{\alpha=1}^{n} Z_\alpha^{(2)} = \left[Z_1^{(2)} \quad Z_2^{(2)} \quad \cdots \quad Z_\alpha^{(2)}\right]_{(p-q)\times n}$$

Since $YY' = A_{22}$ is a non-singular matrix, $\exists$ a matrix F such that

$$FYY'F' = I_{(p-q)\times(p-q)}$$

Writing $G_2 = FY$, we have

$$G_2 G_2' = I_{(p-q)\times(p-q)}$$

Thus, $\exists$ a matrix G such that $G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}$ is an orthogonal matrix.

Where G_2 is a matrix of order $(n - p + q) \times n$.

Consider, a transformation

$$U = WG'$$

where, $U' = [U_1 \quad U_2 \quad \cdots \quad U_n]$

Since G is orthogonal matrix, we have

$$GG' = I_n$$

$$\Rightarrow \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} [G_1' \quad G_2'] = \begin{bmatrix} I_{\{n-(p-q)\}} & 0 \\ 0 & I_{(p-q)} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} G_1 G_1' & G_1 G_2' \\ G_2 G_1' & G_2 G_2' \end{bmatrix} = \begin{bmatrix} I_{\{n-(p-q)\}} & 0 \\ 0 & I_{(p-q)} \end{bmatrix}$$

Let $G = \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_n \end{bmatrix}$

Then

$$U_1 = W g_1'$$
$$U_2 = W g_2'$$
$$\vdots$$
$$U_n = W g_n'$$

$$\boxed{U_\alpha = W g_\alpha'}$$

or

$$U_\alpha = \begin{bmatrix} Z_1^{(1)} & Z_2^{(1)} & \cdots & Z_n^{(1)} \end{bmatrix} \begin{bmatrix} g_{\alpha 1} \\ g_{\alpha 2} \\ \vdots \\ g_{\alpha n} \end{bmatrix}$$

$$= \sum_{\beta=1}^{n} g_{\alpha\beta} Z_\beta^{(1)}$$

is a linear combination of n independent normal vector. Therefore, the distribution of U_α is multivariate normal with

$$E[U_\alpha] = 0$$

$$Cov[U_\alpha, U_\beta] = E(U_\alpha U_\beta')$$

$$= E(W g_\alpha' g_\beta W')$$

Since G is an orthogonal matrix $g_\alpha' g_\beta = 0 \; \forall \alpha \neq \beta$. Thus

$$Cov[U_\alpha, U_\beta] = 0 \qquad \text{if} \quad \alpha \neq \beta$$

$U_1, U_2, \ldots, U_n$ are independently distributed.

The conditional distribution of $X^{(1)}$ given $X^{(2)}$ is

$$f\left(X^{(1)}\middle|X^{(2)}\right)\sim N_q\left(\mu_1+\Sigma_{12}\Sigma_{22}^{-1}\left(X^{(2)}-\mu_2\right),\Sigma_{11.2}\right)$$

Since $Z_\alpha\sim N(0,\Sigma)$, the conditional distribution of $Z^{(1)}$ given $Z^{(2)}$ is

$$Z_\alpha^{(1)}|Z_\alpha^{(2)}\sim N_q\left(\Sigma_{12}\Sigma_{22}^{-1}\,Z_\alpha^{(2)},\Sigma_{11.2}\right).$$

Consider

$$E[U|Y]=E[WG'|Y]=E[W|Y]G'=\Sigma_{12}\Sigma_{22}^{-1}YG'$$

Since $G_2=FY\Rightarrow Y=F^{-1}G_2$,

We have

$$E[U|Y]=\Sigma_{12}\Sigma_{22}^{-1}F^{-1}G_2G'$$

$$=\beta F^{-1}\qquad(\beta=\Sigma_{12}\Sigma_{22}^{-1})$$

The conditional distribution of $U_{\{n-(p-q)\}+\alpha},\alpha=1,\dots,(p-q)$ is $N_q(\nu_\alpha,\Sigma_{11.2})$, where, ν_α is the α^{th} column of βF^{-1}.

(i) $U_1,U_2,\dots,U_n$ are independent.

(ii) $U_1,U_2,\dots,U_{\{n-(p-q)\}}\sim N_q(0,\Sigma_{11.2})$

(iii) $U_{\{n-(p-q)\}+1},U_{\{n-(p-q)\}+2},\dots,U_n\sim N_q(\nu_\alpha,\Sigma_{11.2})$

Then

$$\mathrm{A}_{11}=WW'=UGG'U'=UU'$$

Therefore, G is an orthogonal matrix $(G')^{-1}=G\Rightarrow GG'=I$

$$\mathrm{A}_{11}=[U_1\quad U_2\quad\cdots\quad U_n]\begin{bmatrix}U_1\\U_2\\\vdots\\U_n\end{bmatrix}$$

$$= \sum_{\alpha=1}^{n} U_\alpha U_\alpha'$$

and

$$\mathrm{A}_{12}\mathrm{A}_{22}^{-1}\mathrm{A}_{21} = (WY')(YY')^{-1}(YW')$$

$$= (UG)(G'U')$$

$$= U\begin{bmatrix} G_1 \\ G_2 \end{bmatrix}[G_1' \quad G_2']U'$$

$$= U\begin{bmatrix} G_1G_1' & G_1G_2' \\ G_2G_1' & G_2G_2' \end{bmatrix}U'$$

$$= [U_1 \quad U_2 \quad \cdots \quad U_n]\left((a_{ij})\right)\begin{bmatrix} U_1' \\ U_2' \\ \vdots \\ U_n' \end{bmatrix}$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} U_i\, a_{ij} U_i'$$

When $i = j$ $a_{ii} = 0$ and for $i = 1,2,\dots,n-(p-q), a_{ij} = 0$ for $i \neq j$, and $a_{ii} = 1, \forall\, i = \{n-(p-q)\}+1, \{n-(p-q)\}+2, \dots, n.$

Hence

$$\mathrm{A}_{12}\mathrm{A}_{22}^{-1}\mathrm{A}_{21} = \sum_{\alpha=\{n-(p-q)\}+1}^{n} U_\alpha U_\alpha'$$

So that

$$\mathrm{A}_{11.2} = \mathrm{A}_{11} - \mathrm{A}_{12}\mathrm{A}_{22}^{-1}\mathrm{A}_{21}$$

$$= \sum_{\alpha=1}^{n} U_\alpha U_\alpha' - \sum_{\alpha=\{n-(p-q)\}+1}^{n} U_\alpha U_\alpha'$$

$$= \sum_{\alpha=1}^{\{n-(p-q)\}} U_\alpha U_\alpha'$$

where $U_\alpha \sim N_q(0, \Sigma_{11.2})$.

Hence $A_{11.2} \sim W_q(\Sigma_{11.2}, \{n-(p-q)\})$.

Since $A_{12}A_{22}^{-1}A_{21} = \sum_{\alpha=\{n-(p-q)\}+1}^{n} U_\alpha U_\alpha'$ and U_α are independent. This implies that $A_{11.2}$ and $A_{12}A_{22}^{-1}A_{21}$ are independently distributed.

Further $U_\alpha \sim N(\nu_\alpha, \Sigma_{11.2})$, it implies that $A_{12}A_{22}^{-1}A_{21}$ follows a 'non-central Wishart distribution'.

Theorem 4.4.7.: If $A \sim W_p(n, \Sigma)$ and B is any $(q \times p)$ matrix of rank $q(q \leq p)$ then $BAB' \sim W_q(n, B\Sigma B')$.

Proof: $A \sim W_p(n, \Sigma) \Rightarrow A = \sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$, $Z_\alpha \sim N(0, \Sigma)$, $Z_\alpha' s$ are independently distributed.

Then

$$BAB' = B\left(\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'\right)B' = \sum_{\alpha=1}^{n} BZ_\alpha Z_\alpha' B' = \sum_{\alpha=1}^{n} (BZ_\alpha)(BZ_\alpha)' = \sum_{\alpha=1}^{n} Z_\alpha^*(Z_\alpha^*)'$$

where $Z_\alpha^* = BZ_\alpha$. Further,

$Z_\alpha \sim N_p(0, \Sigma)$

Thus

$BZ_\alpha \sim N_p(0, B\Sigma B')$

$Z_\alpha^* \sim N_p(0, B\Sigma B')$

$\Rightarrow BAB' \sim W_q(n, B\Sigma B')$ if $Z_\alpha^{*\prime} s$ are independent.

Since

$Cov(Z_\alpha^*, Z_\beta^*) = E\left[Z_\alpha^*(Z_\beta^*)'\right]$

$\Rightarrow Cov(BZ_\alpha, BZ_\beta) = E\left[BZ_\alpha(BZ_\beta)'\right]$

$= E\left[BZ_\alpha Z_\beta' B'\right]$

$= BE\left[Z_\alpha Z_\beta{}'\right]B'$

Since $Cov(Z_\alpha, Z_\beta') = E\left[Z_\alpha Z_\beta{}'\right] = 0 \quad if\ \alpha \neq \beta$

We have

$Cov(BZ_\alpha, BZ_\beta)$

$= BE\left[Z_\alpha Z_\beta{}'\right]B' = 0$

Therefore Z_α's are i.i.d.$\sim N_p(0, \Sigma)$ implies that Z_α^* $(\alpha = 1,2, \dots, n)$ are i.i.d. $N_q(0, B\Sigma B')$.

4.5 Cochran theorem (From where you have taken this portion. Cochran's theorem is different.)

Cochran's Theorem (1952) is a fundamental result in multivariate analysis that states:

"If a quadratic form $Q = X^T AX$ is distributed as a chi-square with p degrees of freedom, then the matrix A is idempotent ($A^2 = A$) and rank$(A) = p$."

Conversely,

"If a matrix A is idempotent ($A^2 = A$) and rank$(A) = p$, then the quadratic form $Q = X^T AX$ is distributed as chi-squared with p degrees of freedom."

Proof: 1. Suppose $Q = X^T AX \sim \chi_p^2$

a. Show that A is idempotent:

$E(Q) = E(X^T AX)$

$= tr\ (A\ \Sigma)$

where Σ is the covariance matrix of X.

Since $E(Q) = p$, we have

$tr\ (A\ \Sigma) = p$

Using the trace operator's properties, we get:

$tr\ (A^2\ \Sigma) = tr\ (A\ \Sigma) = p$

Thus, $A^2 = A$.

b. Show that rank$(A)\ =\ p$

Since A is idempotent, we have

$A^2 = A$

Taking the determinant of both sides, we get

$|A^2| = |A|$.

Using the property $\det(AB) = \det(A)\det(B)$, we get

$\det(A)\det(A) = \det(A)$

Since $det(A)\ \neq\ 0$ (otherwise, Q would not be chi-squared), we have

$det(A)\ =\ 1$

Thus, rank$(A)\ =\ p$.

1. Conversely, suppose A is idempotent ($A^2 = A$) and rank$(A) = p$.

a. Show that $Q = X^T AX \sim \chi_p^2$.

Since A is idempotent, we have:

$A^2 = A$

Multiplying both sides by X^T and X, we get

$X^T A^2 X = X^T AX$

Using the quadratic form $Q = X^T AX$, we have

$Q = X^T AX = X^T A^2 X$

Thus, Q is a quadratic form in the variables X, and its distribution is chi-squared with p degrees of freedom.

Therefore, Cochran's Theorem is proved.

4.6 Distribution of Characteristic Roots and Vectors of Wishart Matrices

The distribution of characteristic roots and vectors of Wishart matrices is a fundamental concept in statistics and random matrix theory.

In brief, a Wishart matrix is a random matrix formed from the Gram matrix of a multivariate normal distribution. Its characteristic roots (eigenvalues) and vectors (eigenvectors) have the following distributions:

- Roots (eigenvalues): Follow a Wishart distribution, which is a generalization of the chi-squared distribution.
- Vectors (eigenvectors): Are uniformly distributed on the unit sphere, independent of the roots.

These distributions play a crucial role in various areas, such as multivariate analysis, principal component analysis, and signal processing.

A Wishart matrix, denoted by W, is a random $p \times p$ matrix formed from the Gram matrix of a multivariate normal distribution

$W = X^T X$

where X is a $n \times p$ matrix, with each row independently drawn from a p-variate normal distribution with mean vector μ and covariance matrix Σ.

Distribution of Characteristic Roots (Eigenvalues)

The characteristic roots (eigenvalues) of a Wishart matrix, denoted by $\lambda_1, \lambda_2, \ldots, \lambda_p$, follow a Wishart distribution, which is a generalization of the chi-squared distribution.

The Wishart distribution is characterized by two parameters:

- n (degrees of freedom)
- Σ (covariance matrix)

The probability density function (pdf) of the Wishart distribution is:

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

$Q = X^T AX \sim \chi_p^2$

where Γp is the multivariate gamma function.

The characteristic vectors (eigenvectors) of a Wishart matrix, denoted by $v_1, v_2, \ldots, v_p$ are uniformly distributed on the unit sphere, independent of the roots.

4.6 Summary

In this unit, we have covered the concepts of Wishart distribution under following situations:

1. Drive the distribution

2. Discuss properties of Wishart distribution.

3. State and proved Cochran theorem

4. Also derive the distribution of roots and vectors of Wishart Matrix.

4.7 Self-Assessment Exercises

1. State Wishart distribution. Obtain its characteristic function; hence deduce its reproductive properties.
2. Let A follows wishart distribution $W_p(n, \Sigma)$ what is the distribution of $W_1(n, \Sigma)$.
3. If $A \sim W_q(n, \Sigma)$ i.e. Wishart distribution and l' is a non-null $(p \times 1)$ vector. What is the distribution of $l'Al$.
4. If $X \sim W(n, \Sigma)$ then what will be distribution of $Y = CX$ where C is real vector.
5. Let A is distributed according to $W_p(n, \Sigma)$ and A and Σ be partitioned into q and $(p - q)$ rows and column
6. $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$. Then prove that A_{11} is distributed according to $W(n, \Sigma_{11})$.
7. Define the Wishart distribution. Starting from your definition, prove its reproductive property. Give a random sample $X_\alpha, \alpha = 1,2, \dots, N$ from a p-variate normal population with unknown means $\mu_1, \mu_2, \dots, \mu_p$ and unknown dispersion, give the standard procedure for testing the hypothesis $H_0: \mu_1 = \mu_2 = \cdots, = \mu_p = 0$. Obtain the non-null distribution of your statistic.
8. Obtain the characteristic function of the Wishart distribution $W_p(n, \Sigma)$.
9. What is reproductive property of Wishart distribution?
10. If A has wishart $W_p(n, \Sigma)$ then write down the characteristic function of A.
11. Write down the characteristic function of Wishart distribution.
12. Define Wishart distribution. Derive its distribution and also obtain transformation.
13. Define Wishart distribution. Derive its distribution and also obtain its moments.
14. Define Wishart matrix. Write down its probability density function. Obtain its characteristic function. State and prove its reproductive property.
15. Let a $p \times p$ matrix symmetric random matrix $A \sim W_p(n, \Sigma)$ and let be partitioned into q and $(p - q)$ rows and column as

    $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$. Then derive the distribution of $A_{11.2} = A_{11} - A_{12}A_{22}^{-1}A_{21}$

16. Define the characteristic function of the Wishart distribution. Mention any two properties of the Wishart distribution.
17. Suppose that the $A_i (i = 1,2,3)$ are independently distributed according to Wishart distribution $W_p(n_i, \Sigma)$ respectively, then write down the distribution of $A = A_1 + A_2 + A_3$ for $n_1 = 1, n_2 = 5, n_3 = 2$.
19. State and prove the additive property of Wishart distribution.
20. State and prove the Cochran theorem.
21. Let $A \sim W_p(\Sigma, n_i)$ and A and Σ be partitioned into q and $(p-q)$ rows and columns, $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. Then prove that $A_{11.2} \sim W_q(\Sigma_{11.2}, n-(p-q))$.
22. Given $A \sim W(\Sigma, n)$, find the density of inverted Wishart Distribution.

4.8 References

- Johnson, R. A., Wichern, D. W. (2019): Applied Multivariate Statistical Analysis. United Kingdom: Pearson
- Morrison, D.F. (2004): Multivariate Statistical Methods (Fourth Edition). Duxbury Press, New York.
- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Rao, C.R. (2001): Linear Statistical Inference and its Applications (Second Edition), WileyInter Science, New York.
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Giri Narayan, C. (1995): Multivariate Statistical Analysis.
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis : Methods and Applications.
- Kshirsaga,r A. M. (1979): Multivariate Analysis ,Marcel Dekker Inc. New York.
- Mardia, K.V., Kent, J. T and Bibby, J. M. (1979): Multivariate Analysis. Academic Press, New York.
- Kendall, M.G., Stuart, A. and Ord, K.J. (1973): The Advanced Theory of Statistics. (Fourth Edition), Vol. 2, Charles Griffin company Ltd.

4.9 Further Reading

- Kotz, S., Balakrishnan, N. and Johnson, N.L.: Continuous Multivariate Distribution Models and Applications (Second Edition). Volume 1, Wiley - Inter science, New York.
- Khatri, C. G.: Multivariate Analysis.
- Mardia, K. V.: Multivariate Analysis.
- Seber, G.A.F.: *Multivariate Observations.* Wiley, New York.
- Rencher, Alvin C.: Multivariate Statistical Inference and Applications. John Wiley. New York, New York.

UNIT 5: HOTELLING'S T^2 STATISTIC

Structure

5.1 Introduction

Hotelling's T^2, is the multivariate counter part of the t-test. "Multivariate" means that you have data for more than one parameter for each sample. For example, let's say you wanted to compare how well two different sets of students performed in school. You could compare (e.g. mean test scores) with a t-test. Or, you could use Hotelling's T-squared to compare multivariate data, e.g. the multivariate mean of test scores, GPA, and class grades.

5.2 Objectives

After reading this unit, you should be able to:

- formulate the null and alternative hypothesis for mean vector when Σ is known or unknown;
- derive a test statistic for testing the hypothesis of mean vectors;
- Apply the tests to the given data.

5.3 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.

Let $x_1, \dots, x_N$ is a random sample from $N(\mu, \sigma^2)$ and

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad , \quad s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2$$

Then the distribution of

$$t = \frac{\sqrt{N}(\overline{x} - \mu)}{s},$$

is t-distribution with $(N-1)$ degrees of freedom.

Let $x_1, \dots, x_N$ be a random sample from the multivariate normal distribution $N(\mu, \Sigma)$. Then the multivariate analogue of the square of t is

$$T^2 = N(\overline{x} - \mu)' S^{-1}(\overline{x} - \mu)$$

Here

$\overline{x} = \frac{1}{N}\sum_{\alpha=1}^{N} x_\alpha$: sample mean vector

$S = \frac{A}{N-1}$: sample covariance matrix

$$A = \sum_{\alpha=1}^{N} (x_\alpha - \overline{x})(x_\alpha - \overline{x})'$$

5.3.1 Assumptions for 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.

5.3.2 Importance of Hotelling's T^2

Hotelling's $T^2$is an important tool for identifying changes in means between multiple populations. By using linear combinations of variables, it allows us to compare multiple samples at once, instead of having to run separate tests for each sample. This makes it much easier and faster to identify any meaningful changes in means between populations over time or across different groups of people. Additionally, because it uses a chi-squared test statistic, it helps us determine whether or not these changes are statistically significant something that traditional t-tests cannot do on their own.

Example 5.3(1): Random sample with $N = 20$, were collected. The sample mean vector and covariance matrix are given bellow:

$$\bar{x} = \begin{bmatrix} 10 \\ 20 \end{bmatrix}, S = \begin{bmatrix} 40 & -50 \\ -50 & 100 \end{bmatrix}$$

Obtain the value of Hotelling's T^2 statistic.

Solution: Given that, $N = 20$

$$\bar{x} = \begin{bmatrix} 10 \\ 20 \end{bmatrix}, S = \begin{bmatrix} 40 & -50 \\ -50 & 100 \end{bmatrix}$$

We have

$$S^{-1} = \begin{bmatrix} 0.0667 & 0.0333 \\ 0.0333 & 0.0267 \end{bmatrix}$$

Hence

$$T^2 = N(\overline{x} - \mu)' S^{-1} (\overline{x} - \mu)$$

$$= 20 \left[\begin{pmatrix} 10 \\ 20 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right]' \begin{bmatrix} 0.0667 & 0.0333 \\ 0.0333 & 0.0267 \end{bmatrix} \left[\begin{pmatrix} 10 \\ 20 \end{pmatrix} - \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \right]$$

$$= 20[0.0667(10 - \mu_1) + 0.0333(20 - \mu_2) \quad 0.0333(10 - \mu_1) + 0.0267(20 - \mu_2)] \begin{bmatrix} 10 - \mu_1 \\ 20 - \mu_2 \end{bmatrix}$$

$$\Rightarrow T^2 = 1.334(10 - \mu_1)^2 + 1.332(10 - \mu_1)(20 - \mu_2) + 0.534(20 - \mu_2)^2$$

5.4 Distribution of Hotelling's T^2

Theorem 5.4.1: Let $T^2 = y' S^{-1} y$, where $y \sim N(\upsilon, \Sigma)$ and nS is independently distributed as $\sum_{\alpha=1}^{n} Z_\alpha Z_\alpha'$ with Z_α independent, each with distribution $N(0, \Sigma)$. Then $(T^2/n)[(n - p + 1)/p]$ is distributed as a non-central F with p and $(n - p + 1)$ degrees of freedom and non-centrality parameter $\nu' \Sigma^{-1} \nu$. If $\nu = 0$, the distribution is central F-distribution.

Proof: By definition

$$T^2 = N(\overline{x} - \mu)' S^{-1} (\overline{x} - \mu)$$

Let $y = \sqrt{N}(\bar{x} - \mu_0)$, then

$$E(y) = \sqrt{N} E(\bar{x} - \mu_0)$$

$$= \sqrt{N}(\mu - \mu_0) = \nu,$$

and

$$\Sigma_y = E[y - E(y)][y - E(y)]'$$

$= \Sigma$

Therefore $y \sim N(v, \Sigma)$. Then

$T^2 = y' S^{-1} y$.

Let D be a non-singular matrix such that

$D \Sigma D' = I$

$\Rightarrow DD' = \Sigma^{-1}$

Define

$y^* = Dy$,

$S^* = DSD'$.

Then

$E(y^*) = DE(y) = Dv = v^*$

$\Sigma_{y^*} = E[y^* - E(y^*)][y^* - E(y^*)]'$

$= D \Sigma D' = I$

Therefore $y^* \sim N(v^*, \mathrm{I})$. Hence

$T^2 = y^{*\prime} S^{*-1} y^*$

$$n.S^* = \sum_{\alpha=1}^{n} Z_\alpha^* Z_\alpha^{*\prime} = \sum (DZ_\alpha)(DZ_\alpha)'$$

$Z_\alpha^* = DZ_\alpha \sim N(0, I)$

Let us define a $p \times p$ orthogonal matrix Q such that its first row is defined by

$$q_{1i} = \frac{y_i^*}{\sqrt{y^{*\prime} y^*}} \ , \quad i = 1,2, \dots, p$$

This is permissible since $\sum_i q_{1i}^2 = 1$. Remaining $(p-1)$ rows can be defined by some arbitrary rule. Since Q depends on y^* it is a random matrix.

Let $U = Qy^*$ be an orthogonal transformation, also

$$B = \left(\left(b_{ij} \right) \right) n \, QS^*Q'$$

Then the first element of U is given by

$$U_1 = \sum_{i=1}^{p} q_{1i} y_i^* = \sum_{i=1}^{p} \frac{y_i^{*2}}{\sqrt{y_i^{*\prime} y_i^*}} = \sqrt{y_i^{*\prime} y_i^*}$$

and $\forall \, j = 2, \dots, p$

$$U_j = \sum_{i=1}^{p} q_{ji} y_i^* \ , \qquad y_i^* = q_{1i} \sqrt{y_i^{*\prime} y_i^*}$$

$$= \sqrt{y_i^{*\prime} y_i^*} \sum_{i=1}^{p} q_{ji} \, q_{1i} \qquad (\text{since } Q \text{ is an orthogonal matrix})$$

$$= 0 \qquad (\text{by using the property of an orthogonal matrix})$$

Thus

$$T^2 = y^{*\prime} S^{*-1} y^* = (Q^{-1}U)' S^{*-1} (Q^{-1}U) = U'(QS^*Q')^{-1}U = nU'(Q \, NS^*Q')^{-1}U = n \, U'B^{-1}U$$

Then

$$\frac{T^2}{n} = U'B^{-1}U = (U_1 \quad 0 \quad \cdots \quad 0)\begin{pmatrix} b^{11} & b^{12} & \cdots & b^{1p} \\ b^{21} & b^{22} & \cdots & b^{2p} \\ \vdots & \vdots & \cdots & \vdots \\ b^{p1} & b^{p2} & \cdots & b^{pp} \end{pmatrix}\begin{pmatrix} U_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = U_1^2 b^{11}$$

where, $B^{-1} = \left(\left(b^{ij}\right)\right)$.

$$B = \begin{pmatrix} b_{11} & b_{(1)} \\ b'_{(1)} & B_{22} \end{pmatrix}$$

Now

$$B^{-1}B = I$$

$$\begin{pmatrix} b^{11} & b^{(1)} \\ b^{(1)\prime} & B^{22} \end{pmatrix}\begin{pmatrix} b_{11} & b_{(1)} \\ b'_{(1)} & B_{22} \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & I \end{bmatrix}$$

$$\begin{pmatrix} b^{11}b_{11} + b^{(1)}b'_{(1)} & b^{11}b_{(1)} + b^{(1)}B_{22} \\ \cdots & \cdots \end{pmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & I \end{bmatrix}$$

Hence,

$$b^{11}b_{11} + b^{(1)}b'_{(1)} = 1$$

$$b^{11}b_{(1)} + b^{(1)}B_{22} = 0$$

$$b^{(1)} = -b^{11}b_{(1)}{B_{22}}^{-1}$$

$$b_{11}b^{11} - b^{11}b_{(1)}{B_{22}}^{-1}b'_{(1)} = 1$$

or

$$b_{11} = \frac{1}{b^{11} - b_{(1)}{B_{22}}^{-1}b'_{(1)}} = \frac{1}{b_{11.2,\ldots,p}}$$

Therefore

$$\frac{T^2}{n} = \frac{U_1^2}{b_{11.2,\ldots,p}} = \frac{y_i^{*\prime} y_i^*}{b_{11.2,\ldots,p}}$$

Now conditional distribution of B is given Q is that of $B = Q\, n\, S^*\, Q' = Q \sum_{\alpha=1}^{n} Z_\alpha^*\, Z_\alpha^{*\prime} Q' = \sum_{\alpha=1}^{n} \nu_\alpha \nu_\alpha'$, where conditionally $\nu_\alpha = QZ_\alpha^*$ are independent each following distribution $N(0, I)$. Hence, $b_{11.2,\ldots,p}$ is conditionally distributed as $\sum_{\alpha=1}^{n-(p-1)} W_\alpha^2$, where conditionally W_α are independent, i.e. $W_\alpha \sim N(0,1)$.

Hence $b_{11.2,\ldots,p}$ is a conditionally distributed as χ^2 with $n-(p-1)$ degree of freedom. Since the conditional distribution of $b_{11.2,\ldots,p}$ does not depend on Q, therefore the unconditional distribution of $b_{11.2,\ldots,p}$ is $\chi^2_{n-(p-1)}$.

Since $y^* \sim N(\nu^*, I)$, $y^{*\prime} y^* \sim \text{non central} - \chi^2$ with p degrees of freedom and non-centrality parameter $\nu^{*\prime}\nu^* = \nu' \Sigma^{-1} \nu = \lambda^2$. Hence $\frac{T^2}{n}$ is a ratio of a $\text{non central} - \chi^2$ with p degree of freedom to an independent χ^2 with $(n-p+1)$ degree of freedom. Therefore

$$\frac{n-p+1}{p}\left(\frac{T^2}{n}\right) = \frac{\frac{\chi_p^2(\lambda^2)}{p}}{\frac{\chi^2_{n-(p-1)}}{n-p+1}} \sim non\ central\ F\ with\ p\ \&\ n-p+1\ d.f.$$

If $\mu = \mu_0$, then F-distribution is central.

We shall call this distribution as $T^2 - distribution$ with n d.f.

Corollary 5.4.2: Let $x_1, \ldots, x_N$ be a sample from $N(\mu, \Sigma)$ and

$T^2 = N(\overline{x} - \mu_0)' S^{-1} (\overline{x} - \mu_0)$, then distribution of $\frac{T^2}{N-1}\frac{N-p}{p}$ is

noncentral $F_{p,N-p}[N(\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0)]$. If $\mu = \mu_0$, the distribution is central.

5.4.1 T^2-Statistic as a Function of Likelihood Ratio Criterion

Let $x_1, \ldots, x_N$ be a random sample from $N_p(\mu, \Sigma)$ $(N > p)$. The likelihood function is

$$L(\mu,\Sigma)=\frac{|\Sigma^{-1}|^{\frac{1}{2}N}}{(2\pi)^{\frac{1}{2}pN}}\exp\left[-\frac{1}{2}\sum_{\alpha=1}^{N}(x_\alpha-\mu)'\Sigma^{-1}(x_\alpha-\mu)\right]$$

For testing $H_0:\mu=\mu_0$ against $H_1:\mu=\mu_1$, the likelihood ratio criterion is

$$\lambda=\frac{\max\limits_{\Sigma^{-1}} L(\mu_0,\Sigma^{-1})}{\max\limits_{\mu,\Sigma^{-1}} L(\mu,\Sigma^{-1})}$$

$\max\limits_{\Sigma^{-1}} L(\mu_0,\Sigma^{-1})$: maximum of the likelihood function for μ,Σ^{-1} in the parameter space restricted by the null hypothesis.

$\max\limits_{\mu,\Sigma^{-1}} L(\mu,\Sigma^{-1})$: maximum of L over the entire parameter space (μ,Σ^{-1}

Maximum of $L(\mu,\Sigma^{-1})$ over the entire parameter space are defined by the maximum likelihood estimates of μ and Σ

$$\hat{\mu}_\Omega=\overline{x}\,,\hat{\Sigma}_\Omega=\frac{1}{N}\sum_{\alpha}(x_\alpha-\overline{x})(x_\alpha-\overline{x})'=\frac{A}{N}$$

When $\mu=\mu_0$, the likelihood function is maximized at

$$\hat{\Sigma}_\omega=\frac{1}{N}\sum_{\alpha=1}^{N}(x_\alpha-\mu_0)(x_\alpha-\mu_0)'$$

Therefore,

$$\max_{\mu,\Sigma^{-1}} L(\mu,\Sigma^{-1})=\frac{1}{(2\pi)^{\frac{1}{2}pN}\left|\hat{\Sigma}_\Omega\right|^{\frac{1}{2}N}}e^{\left[-\frac{1}{2}\Sigma\alpha(x_\alpha-\overline{x})(\hat{\Sigma}_\Omega)^{-1}(x_\alpha-\overline{x})'\right]}=\frac{1}{(2\Pi)^{\frac{1}{2}pN}\left|\hat{\Sigma}_\Omega\right|^{\frac{1}{2}N}}e^{-\frac{1}{2}pN}$$

$\max\limits_{\Sigma^{-1}} L(\mu_0,\Sigma^{-1})$

$$= \frac{1}{(2\pi)^{\frac{1}{2}pN}\left|\hat{\Sigma}_{\omega}\right|^{\frac{1}{2}N}} \exp\left[-\frac{1}{2} tr\, N\left\{\sum_{\alpha=1}^{N}(x_\alpha-\mu_0)(x_\alpha-\mu_0)'\right\}^{-1}\left\{\sum_{\alpha=1}^{N}(x_\alpha-\mu_0)(x_\alpha-\mu_0)'\right\}\right]$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}pN}\left|\hat{\Sigma}_{\omega}\right|^{\frac{1}{2}N}} \exp\left[-\frac{pN}{2}\right]$$

Consider

$$\sum_{\alpha=1}^{N}(x_\alpha-\mu_0)(x_\alpha-\mu_0)' = \sum_{\alpha=1}^{N}\{(x_\alpha-\bar{x})+(\bar{x}-\mu_0)\}\{(x_\alpha-\bar{x})+(\bar{x}-\mu_0)\}'$$

$$= A + N(\bar{x}-\mu_0)(\bar{x}-\mu_0)'$$

Thus

$$\lambda = \frac{\left|\hat{\Sigma}_{\Omega}\right|^{\frac{1}{2}N}}{\left|\hat{\Sigma}_{\omega}\right|^{\frac{1}{2}N}} = \frac{|\sum_\alpha(x_\alpha-\overline{x})(x_\alpha-\overline{x})'|^{\frac{1}{2}N}}{|\sum_\alpha(x_\alpha-\mu_0)(x_\alpha-\mu_0)'|^{\frac{1}{2}N}} = \frac{|A|^{\frac{1}{2}N}}{|A+N(\overline{x}-\mu_0)(\overline{x}-\mu_0)'|^{\frac{1}{2}N}} \quad ;$$

Hence

$$|A+N(\overline{x}-\mu_0)(\overline{x}-\mu_0)'| = \left|A+\left\{\sqrt{N}(\overline{x}-\mu_0)\right\}\left\{\sqrt{N}(\overline{x}-\mu_0)'\right\}\right|$$

$$= \begin{vmatrix} A & \sqrt{N}(\overline{x}-\mu_0) \\ -\sqrt{N}(\overline{x}-\mu_0)' & 1 \end{vmatrix}_{1}^{p} = |A||1+N(\overline{x}-\mu_0)'A^{-1}(\overline{x}-\mu_0)|$$

Notice that $|\Sigma| = |\Sigma_{11}||\Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|$.

Therefore

$$\lambda^{2/N} = \frac{1}{1+N(\overline{x}-\mu_0)A^{-1}(\overline{x}-\mu_0)'} = \frac{1}{1+\frac{N}{N-1}(\overline{x}-\mu_0)S^{-1}(\overline{x}-\mu_0)'} = \frac{1}{1+T^2/(N-1)}$$

where $T^2 = N(\overline{x}_0 - \mu_0)' S^{-1}(\overline{x}_0 - \mu_0)$, $S = \frac{A}{N-1}$

The likelihood ratio test is defined by the critical region $\lambda \leq \lambda_0$, where λ_0 is chosen, so that

$P(\lambda \leq \lambda_0 | H_0) = \alpha$: level of significance.

Thus

$$\lambda^{\frac{2}{N}} \leq \lambda_0^{\frac{2}{N}}$$

$$\Rightarrow \frac{1}{1 + T^2/(N-1)} \leq \lambda_0^{\frac{2}{N}}$$

$$\Rightarrow \lambda_0^{-\frac{2}{N}} \leq 1 + T^2/(N-1)$$

$$\Rightarrow T^2 \geq (N-1)\left\{\lambda_0^{-\left(\frac{2}{N}\right)} - 1\right\} = T_0^2$$

Hence the critical region is

$P[T^2 \geq T_0^2 | H_0] = \alpha$

where $T_0^2 = (N-1)(\lambda_0^{-2/N} - 1)$

This test is the likelihood ratio test for testing the hypothesis $H_0: \mu = \mu_0$.

5.4.2 Invariance Property of T^2

Let $X \sim N(\mu, \Sigma)$, then $T_x^2 = N(\overline{x} - \mu)' S_x^{-1}(\overline{x} - \mu)$

Where

$$S_x = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(x_i - \overline{x})'$$

$$= \frac{1}{(N-1)}(x_1x_1' + \cdots + x_Nx_N' - N\bar{x}\bar{x}')$$

Make a non-singular transformation

$Y = DX$

$\Rightarrow y_i = D\, x_i$

Now

$$S_y = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})(y_i - \overline{y})' = \frac{1}{(N-1)}(y_1y_1' + \cdots + y_Ny_N' - N\bar{y}\bar{y}')$$

$$= \frac{1}{(N-1)}(Dx_1x_1'D' + \cdots + Dx_Nx_N'D' - N.D\bar{x}\bar{x}'D')$$

$$= D\left[\frac{1}{(N-1)}(x_1x_1' + \cdots + x_Nx_N' - N\bar{x}\bar{x}')\right]D'$$

$\Rightarrow S_y = DS_xD'$

By definition

$T_y^2 = N(\overline{y} - \mu)'S_y^{-1}(\overline{y} - \mu) = N(D\overline{x} - D\mu)'(DS_xD')^{-1}(D\overline{x} - D\mu)$

$$= N(\overline{x} - \mu)'D'D'^{-1}(S_x)^{-1}(D^{-1}D)(\overline{x} - \mu)$$

$= N(\overline{x} - \mu)'(S_x)^{-1}(\overline{x} - \mu)$

$= T_x^2$

$\Rightarrow T_y^2 = T_x^2$

Hence T_x^2 is invariant under nonsingular transformation $Y = DX$.

5.5. Application of T^2-Statistic

(i) One Sample Problem

Let $x_1, \dots, x_N$ be a random sample from $N(\mu, \Sigma)$. Suppose Σ is unknown and we test $H_0: \mu = \mu_0$. Let $y = \sqrt{N}(\overline{x} - \mu_0)$, then, under null hypothesis, $E(y) = 0$, and $\Sigma_y = \Sigma$. Thus $y \sim N(0, \Sigma)$

Further

$$S = \frac{1}{N-1}\sum_{\alpha=1}^{N}(x_\alpha - \overline{x})(x_\alpha - \overline{x})'$$

$$\Rightarrow A = \sum_{\alpha=1}^{N}(x_\alpha - \overline{x})(x_\alpha - \overline{x})' = \sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha{}'$$

where $Z_\alpha \sim N(0, \Sigma)$. Therefore, by the definition $T^2 = y^{*\prime} S^{*-1} y^*$ and

$$\frac{N-p+1}{p}\left(\frac{T^2}{N}\right) \sim F \text{ with } p \;\&\; N-p+1 \; d.f.$$

Thus, adopting a level of significance α, we reject the hypothesis if $T^2 \geq T_0^2$, where

$$T_0^2 = \frac{(N-1)p}{N-p} F_{p,N-p}(\alpha)$$

$$T^2 = N(\overline{x} - \mu_0)' S^{-1} (\overline{x} - \mu_0)$$

$$= N(N-1)(\overline{x} - \mu_0)' A^{-1} (\overline{x} - \mu_0)$$

$$A^{-1}(\overline{x} - \mu_0) = b$$

$$Ab = (\overline{x} - \mu_0) \qquad (5.1)$$

Thus, the computation of A^{-1} or S^{-1} is not required. The vector b can be directly obtained by solving (5.1), and then

$$\frac{T^2}{N-1} = N(\overline{x} - \mu_0)'b$$

Example 5.5.1: The length of Centrum (x_1) and width of Centrum (x_2) of a sample of 12 fishes of Serrandae family were observed as given in Table

i	x_1	x_2
1	7.5	6.7
2	6.8	6.2
3	8.5	7.1
4	5.8	6
5	5.2	5.8
6	7	6.2
7	8.2	7.5
8	6.9	7.3
9	7.4	6.8
10	8.4	7.3
11	7.6	7
12	9.2	7.8

Test the researcher's claim about mean length and width of Centrum for Serrandae fishes to be 8.94 and 6.76 respectively at 5% level of significance and no information is available about the population covariance matrix.

Solution: We have to test the null hypothesis $H_0: \mu = \begin{pmatrix} 8.94 \\ 6.76 \end{pmatrix} = \mu_0$ against the alternative hypothesis

$H_1: \bar{\mu} \neq \begin{pmatrix} 8.94 \\ 6.76 \end{pmatrix}$.

The sample covariance matrix S is calculated as

$$S = \begin{bmatrix} 1.191 & 0.5877 \\ 0.5877 & 0.3741 \end{bmatrix}$$

Where

$$s_{11} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})' = \frac{1}{12} \times 14.3025 = 1.191$$

$$s_{12} = s_{21} = \frac{1}{N}\sum_{i \neq j=1}^{N}(x_i - \bar{x})\left(x_j - \bar{x}\right)' = \frac{1}{12} \times 7.0525 = 0.5877$$

$$s_{22} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})' = \frac{1}{12} \times 4.4892 = 0.3741$$

The test statistic under H_0 can be calculated as

$$T^2 = (N-1)(\bar{x} - \bar{\mu})'\mathrm{S}^{-1}(\bar{x} - \bar{\mu})$$

The sample mean is

$$\bar{x} = \begin{pmatrix} 7.375 \\ 6.808 \end{pmatrix}$$

Where

$$\bar{x_1} = \frac{1}{12}\sum_{i=1}^{12} x_i = \frac{88.5}{12} = 7.375$$

$$\bar{x_2} = \frac{1}{12}\sum_{i=1}^{12} x_i = \frac{81.7}{12} = 6.808$$

$$\bar{\mu} = \begin{pmatrix} 8.94 \\ 6.76 \end{pmatrix}$$

$$\mathrm{S}^{-1} = \frac{|\mathrm{S}|}{adj\ \mathrm{S}} = \begin{bmatrix} 3.73496 & -5.86751 \\ -5.86750 & 11.89075 \end{bmatrix}$$

Thus

$$T^2 = 11(1.565 \quad -0.048)\begin{bmatrix} 3.735 & -5.868 \\ -5.868 & 11.891 \end{bmatrix}\begin{pmatrix} 7.375 - 8.94 = 1.565 \\ 6.808 - 6.76 = -0.048 \end{pmatrix}$$

$$= 11(6.127 \quad -9.754)\begin{pmatrix} 1.565 \\ -0.048 \end{pmatrix} = 11 \times 10.057 = 110.63$$

And

$$F_{(p,N-p+1)} = \frac{N-p+1}{p}\left(\frac{T^2}{N}\right)$$

$$= \frac{12-2+1}{2} \times \frac{110.63}{12}$$

$$= \frac{9 \times 110.63}{24} = \frac{995.643}{24}$$

$$= 41.485$$

$$\Rightarrow F_{(p,N-p+1)} = 41.485$$

Now $F_{(2,11)}$ tabulated value is 3.98. Therefore $F_{cal} > F_{tab}$ i.e. we reject hypothesis at 5% level of significance and conclude that the researcher claim is not true in the light of observed data.

(ii) Two Sample Problem

Let $x_\alpha^{(i)}$ $(\alpha = 1,2,\cdots,N_i\,; i = 1,2)$ be a random sample from $N(\mu^{(i)}, \Sigma)$ respectively.

The hypothesis is

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

Let

$$\bar{x}^{(i)} = \frac{1}{N_i}\sum_{\alpha=1}^{N_i} x_\alpha^{(i)}$$

be the sample mean vector, and $\bar{x}^{(i)} \sim N(\mu^{(i)},\ \Sigma/N_i)$, then $\{\bar{x}^{(1)} - \bar{x}^{(2)}\} \sim N\left\{0,\ \left(\frac{1}{N_1} + \frac{1}{N_2}\right)\Sigma\right\}$, under H_0

Let

$$y = \sqrt{\frac{N_1 N_2}{N_1 + N_2}}\{\bar{x}^{(1)} - \bar{x}^{(2)}\}$$

Then, under null hypothesis, $E(y) = 0$, and

$$\Sigma_y = \frac{N_1 N_2}{N_1 + N_2} E\left[\{\bar{x}^{(1)} - \bar{x}^{(2)}\}\{\bar{x}^{(1)} - \bar{x}^{(2)}\}'\right]$$

$$= \Sigma$$

Thus $y \sim N(0, \Sigma)$

Let

$$S^{(i)} = \frac{1}{N_i - 1}\sum_{\alpha=1}^{N_i}\left(x_\alpha^{(i)} - \overline{x}^{(i)}\right)\left(x_\alpha^{(i)} - \overline{x}^{(i)}\right)'$$

and

$$S = \frac{1}{N_1 + N_2 - 2}\sum_{i=1}^{2}\sum_{\alpha=1}^{N_i}\left(x_\alpha^{(i)} - \overline{x}^{(i)}\right)\left(x_\alpha^{(i)} - \overline{x}^{(i)}\right)'$$

$$\Rightarrow S = \frac{(N_1 - 1)S^{(1)} + (N_2 - 1)S^{(2)}}{N_1 + N_2 - 2}$$

S is the pooled sample variance covariance matrix. Hence

$$(N_1 + N_2 - 2)S = (N_1 - 1)S^{(1)} + (N_2 - 1)S^{(2)}$$

$$= A^{(1)} + A^{(2)}$$

$$= \sum_{\alpha=1}^{(N_1+N_2-2)} Z_\alpha Z_\alpha{}'$$

where $Z_\alpha \sim N(0, \Sigma)$. Therefore, $(N_1 + N_2 - 2)S$ is distributed as $\sum_{\alpha=1}^{(N_1+N_2-2)} Z_\alpha {Z_\alpha}'$. By the definition

$$T^2 = y' S^{-1} y$$

$$= \left(\frac{N_1 N_2}{N_1 + N_2}\right) \{\bar{x}^{(1)} - \bar{x}^{(2)}\}' S^{-1} \{\bar{x}^{(1)} - \bar{x}^{(2)}\}$$

is distributed as T^2 with $(N_1 + N_2 - 2)$ degree of freedom, i.e.

$$\frac{N_1 + N_2 - 2 - p + 1}{p} \left\{\frac{T^2}{(N_1 + N_2 - 2)}\right\} \sim F \text{ with } p \text{ and } (N_1 + N_2 - p)\ d.f.$$

Thus, adopting a level of significance α, we reject the hypothesis if $T^2 \geq T_0^2$, where

$$T_0^2 = \frac{\{N_1 + N_2 - 1\}p}{N_1 + N_2 - p} F_{p, N_1 + N_2 - p}(\alpha)$$

(iii) *k*-Sample Problem

Let $x_\alpha^{(i)}$ $(\alpha = 1,2, \cdots, N_i\ ; i = 1,2, \cdots, k)$ be a random sample from $N\left(\mu^{(i)}, \Sigma\right)$ respectively.

Suppose we are required to test

$$H_0 : \sum_{i=1}^{k} \beta_i\ \mu^{(i)} = \mu$$

Where $\beta_1, \beta_2, \cdots, \beta_k$ are scalars and μ is mean vector. Let

$$\bar{x}^{(i)} = \frac{1}{N_i} \sum_{\alpha=1}^{N_i} x_\alpha^{(i)}$$

Be the sample mean vector, and $\bar{x}^{(i)} \sim N\left(\mu^{(i)},\ \Sigma/N_i\right)$. Then $\sum_{i=1}^{k} \beta_i\, \bar{x}^{(i)} \sim N\left(\mu^{(i)}, \Sigma/D\right)$ under H_0, where $E\left(\sum_{i=1}^{k} \beta_i\, \bar{x}^{(i)}\right) = \mu$ and the variance covariance matrix is

$$Cov\left(\sum_{i=1}^{k}\beta_i\,\bar{x}^{(i)}\right)=\sum_{i=1}^{k}\beta_i^2\;\;Cov\left(\bar{x}^{(i)}\right)=\sum_{i=1}^{k}\frac{\beta_i^2}{N_i}\Sigma=\frac{\Sigma}{D}$$

Let $y=\sqrt{D}\left(\sum_{i=1}^{k}\beta_i\,\bar{x}^{(i)}-\mu\right)$, then $E(y)=0$, under null hypothesis and $\Sigma_y=\Sigma$. Thus $y\sim N(0,\Sigma)$, and

$$S^{(i)}=\frac{1}{N_i-1}\sum_{\alpha=1}^{N_i}\left(x_\alpha^{(i)}-\overline{x}^{(i)}\right)\left(x_\alpha^{(i)}-\overline{x}^{(i)}\right)'$$

$$S=\frac{1}{\sum_{i=1}^{k}N_i-k}\sum_{i=1}^{k}\sum_{\alpha=1}^{N_i}\left(x_\alpha^{(i)}-\overline{x}^{(i)}\right)\left(x_\alpha^{(i)}-\overline{x}^{(i)}\right)'$$

$$\Rightarrow\left(\sum_{i=1}^{k}N_i-k\right)S=\sum_{i=1}^{k}\sum_{\alpha=1}^{N_i}\left(x_\alpha^{(i)}-\overline{x}^{(i)}\right)\left(x_\alpha^{(i)}-\overline{x}^{(i)}\right)'=\sum_{\alpha=1}^{\left(\sum_{i=1}^{k}N_i-k\right)}Z_\alpha Z_\alpha{}'$$

where $Z_\alpha\sim N(0,\Sigma)$

Therefore by the definition, $T^2=y'S^{*-1}y=D\left(\sum_{i=1}^{k}\beta_i\,\bar{x}^{(i)}-\mu\right)'S^{-1}\left(\sum_{i=1}^{k}\beta_i\,\bar{x}^{(i)}-\mu\right)$ is distributed as T^2 with $\left(\sum_{i=1}^{k}N_i-k\right)$ degree of freedom, i.e.

$$\frac{\left(\sum_{i=1}^{k}N_i-k\right)-p+1}{p}\left\{\frac{T^2}{\left(\sum_{i=1}^{k}N_i-k\right)}\right\}\sim F\text{ with }p\;\&\;\left(\sum_{i=1}^{k}N_i-k\right)-p+1\;d.f.$$

Thus, adopting a level of significance α, then, we reject the hypothesis if

$$T^2\geq T_0^2$$

where

$$T_0^2=\frac{\left\{\left(\sum_{i=1}^{k}N_i-k\right)-1\right\}p}{\left(\sum_{i=1}^{k}N_i-k\right)-p}F_{p,\left(\sum_{i=1}^{k}N_i-k\right)-p}(\alpha)$$

5.6 Summary

In this unit, we have covered the concepts of Hotelling's T^2 under following situations:

1. Derived its distribution.

2. Discuss likelihood criterion.

3. Test for mean vector for one sample case when population covariance is known.

4. Test for mean vector for one sample case when population covariance is unknown.

5. Test for mean vector for k sample case when population covariance is known.

5.7 Self-Assessment Exercises

1. Define Hotelling's T^2 statistic. How is it related with F-distribution? For a multivariate normal distribution$N_p(\mu, \Sigma)$, derive a likelihood ratio test for testing $H_0: \mu = \mu_0$ against the alternative $H_1: \mu \neq \mu_0$.
2. Derive the null distribution of Hotelling's T^2 statistic.
3. Discuss various applications of Hotelling's T^2 statistic in testing for the mean vector of one and more multivariate normal population.
4. A sample of 10 industrial corporations was considered for the pairs of observations of their sales (x_1) and profits (x_2). The observations are given in the following Table:

Corporation No.	x_1 (Rs. In Lakhs)	x_2 (Rs. In Lakhs)
1	40	8
2	42	10
3	34	6
4	16	6
5	50	10
6	24	4
7	37	6
8	42	8
9	25	7
10	20	5

The expected mean vector and variance-covariance matrix is

$\bar{\mu} = \begin{bmatrix} 40 \\ 10 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 13 & 10 \\ 10 & 6 \end{bmatrix}$

Test whether the sample confirms its truth ness of mean vector at 5% level of significance.

5. Two samples of size 50 bars and 60 bars were taken from the lots produced by method 1 and method 2. Two characteristics X_I = lather and X_z = mildness were measured. The summary statistics for bars produced by methods 1 and 2 is given by

$\bar{x}^{(1)} = \begin{bmatrix} 8 \\ 4 \end{bmatrix}, \bar{x}^{(2)} = \begin{bmatrix} 10 \\ 4 \end{bmatrix}$

$S_1 = \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}, S_2 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}$

Test at 5% level of significance whether $\mu^{(1)} = \mu^{(2)}$ or not.

5.8 References

- Johnson, R. A., Wichern, D. W. (2019): Applied Multivariate Statistical Analysis. United Kingdom: Pearson
- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Giri Narayan, C. (1995): Multivariate Statistical Analysis.
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis : Methods and Applications.
- Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York.

5.9 Further Reading

- Khatri, C. G.: Multivariate Analysis.
- Mardia, K. V.: Multivariate Analysis.
- Seber, G.A.F.: *Multivariate Observations*. Wiley, New York.
- Rencher, Alvin C.: Multivariate Statistical Inference and Applications. John Wiley, New York.

UNIT: 6 MAHALANOBIS D^2

Structure

6.1 Introduction

Let X be a $(p \times 1)$ vector random variable having $N(\mu, \Sigma)$ distribution. Let X $(p \times n)$ be a data matrix observed from a random sample of size n. The population distribution involves as parameters p components of mean μ and $\frac{1}{2}p(p+1)$ components of variance-covariance Σ. For

these parameters, minimum $\{2^{p(p+3)} - 1\}$ null hypothesis can be formulated. These null hypotheses can specify the values of a subset of parameters. In this unit, we shall consider the problem of testing hypothesis about the mean μ under both the situations when variance-covariance matrix Σ is known and when variance-covariance matrix Σ is unknown.

6.2 Objectives

After reading this unit, you should be able to:

- Formulate the null and alternative hypothesis for mean vector when Σ is known or unknown
- Derive a test statistic for testing the hypothesis of mean vectors
- Apply the tests to the given data.
- Derive the Mahalanobis D^2.

6.3 Equality of the Component of a Mean Vector in a Multivariate Normal Population

Suppose $x_1^{(1)}, \ldots, x_{N_1}^{(1)}$ *and* $x_2^{(2)}, \ldots, x_{N_2}^{(2)}$ are samples from $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ respectively First, we will consider the two hypotheses about the equality of mean vectors sample problem under the various assumptions about the covariance matrices.

6.3.1 Test for Equality of Two Mean Vectors when Covariance Matrices are known

Let $x_1^{(1)}, \ldots, x_{N_1}^{(1)}$ *and* $x_2^{(2)}, \ldots, x_{N_2}^{(2)}$ be the random sample of sizes N_1 and N_2 drawn from $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$ respectively. If $\overline{x}$ is the sample mean, then it is unbiased estimate of corresponding mean vector μ and covariance matrix $(N_i^{-1}\Sigma_i) i = 1,2$. Let us define a new variable $\overline{x_d} = \overline{x_1} - \overline{x_2}$. Then $\bar{x}_d$ will be follow a multivariate normal distribution with mean vector $\mu = \mu_1 - \mu_2$ and covariance matrix $\Sigma = (N_1^{-1}\Sigma_1 + N_2^{-1}\Sigma_2)$. Therefore,

$(\bar{x}_d - \mu)'\Sigma^{-1}(\bar{x}_d - \mu)$ be a chi-square distribution with p degrees of freedom. Under the null hypothesis $H_0: \mu = \mu_1 - \mu_2 = 0$, it reduces to $(\bar{x}_d)'\Sigma^{-1}(\bar{x}_d)$, and shall follow a chi-square distribution. Hence, we reject null hypothesis H_0, if calculated value of$(\bar{x}_d)'\Sigma^{-1}(\bar{x}_d)$ is greater

than the tabulated value of chi-square with p degrees of freedom and at specified level of significance.

Example 6.3.1: The length of Centrum (x_1) and width of Centrum (x_2) of a sample of 12 fishes of Serrandae family were observed as given in Table

i	x_1	x_2
1	7.5	6.7
2	6.8	6.2
3	8.5	7.1
4	5.8	6
5	5.2	5.8
6	7	6.2
7	8.2	7.5
8	6.9	7.3
9	7.4	6.8
10	8.4	7.3
11	7.6	7
12	9.2	7.8

A researcher claims that the mean length and width of Centrum of fishes belonging to Serrandae family is 8.94 and 6.76, respectively with a variance-covariance matrix of x_1, x_2 as

$$\Sigma = \begin{bmatrix} 13.2496 & 9.0418 \\ 9.0418 & 7.6176 \end{bmatrix}$$

Test this claim at 5% level of significance.

Solution: The null hypothesis is $H_0: \bar{\mu} = \begin{pmatrix} 8.94 \\ 6.76 \end{pmatrix} = \bar{\mu}_0$ against the alternative hypothesis

$H_1: \bar{\mu} \neq \begin{pmatrix} 8.94 \\ 6.76 \end{pmatrix}$.

The population covariance matrix Σ is known as claimed by researcher. The value of test statistics under H_0 is

$$\chi^2 = N(\bar{x} - \bar{\mu})' \Sigma^{-1} (\bar{x} - \bar{\mu})$$

The sample mean is

$$\bar{x} = \begin{pmatrix} 7.375 \\ 6.808 \end{pmatrix}$$

Where

$$\bar{x}_1 = \frac{1}{12}\sum_{i=1}^{12} x_i = \frac{88.5}{12} = 7.375$$

$$\bar{x}_2 = \frac{1}{12}\sum_{i=1}^{12} x_i = \frac{81.7}{12} = 6.808$$

$$\bar{\mu} = \begin{pmatrix} 8.94 \\ 6.76 \end{pmatrix}$$

$$\Sigma^{-1} = \frac{|\Sigma|}{adj\ \Sigma} = \begin{bmatrix} 0.3972 & -0.4715 \\ -0.4715 & 0.6909 \end{bmatrix}$$

Thus

$$\chi^2 = 12(1.565 \quad -0.048)\begin{bmatrix} 0.3972 & -0.4715 \\ -0.4715 & 0.6909 \end{bmatrix}\begin{pmatrix} 7.375 - 8.94 = 1.565 \\ 6.808 - 6.76 = -0.048 \end{pmatrix}$$

$$= 12(0.64425 \quad -0.77106)\begin{pmatrix} 1.565 \\ -0.048 \end{pmatrix} = 12 \times 1.045 = 12.54$$

$$\Rightarrow \chi^2 = 12.54$$

Since χ^2 tabulated value is 10.60.

Therefore $\chi^2_{cal} > \chi^2_{tab}$ i.e. we reject hypothesis at 5% level of significance and conclude that the data contradicts the claim of the researcher.

6.3.2 Test for Equality of Two Mean Vectors when Covariance Matrices are Equal and Unknown

Let $x_1^{(1)}, \ldots, x_{N_1}^{(1)}$ *and* $x_2^{(2)}, \ldots, x_{N_2}^{(2)}$ be the random sample of sizes N_1 and N_2 drawn from $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ respectively. Assume that $\Sigma_1 = \Sigma_2 = \Sigma$ (unknown). Then Mahalanobis distance is defined as

$$\Delta^2 = \left(\mu^{(1)} - \mu^{(2)}\right)' \Sigma^{-1} \left(\mu^{(1)} - \mu^{(2)}\right)$$

The value of Δ^2 cannot be calculated, as the parameters μ_1, μ_2 and Σ are not known. However, Mahalanobis Distance for sample observations is given by

$$D^2 = (\overline{x_1} - \overline{x_2})' S^{-1} (\overline{x_1} - \overline{x_2})$$

where,

$\overline{x}_i$ $(i = 1,2)$ is the mean of i^{th} sample and

$$S = \frac{N_1 S_1 + N_2 S_2}{N_1 + N_2 - 2}$$

Here S_i is the sample covariance matrix of i^{th} sample $(i = 1,2)$. It may be noted here that for the given situation, $\overline{x_1}$ and $\overline{x_2}$, are the unbiased estimators of μ_1 and μ_2 respectively. Similarly, $N_1 S_1$ and $N_2 S_2$ are maximum likelihood estimators of Σ and the pooled unbiased estimator of Σ is S. Further $\overline{x}_1, \overline{x}_2, S_1$ and S_2, are independently distributed.

It may be recalled that if we denote $\overline{x}_d = \overline{x}_1 - \overline{x}_2$, then it follows multivariate normal distribution with mean vector $(\mu_1 - \mu_2)$ and covariance matrix $(N_1^{-1} + N_2^{-1})\Sigma$, under the null hypothesis $H_0: \mu_1 = \mu_2$, $\overline{x}_d \sim \{0, (N_1^{-1} + N_2^{-1})\Sigma\}$. Since the samples are independent, $N_1 S_1$ and $N_2 S_2$ are independently distributed, following Wishart distribution, i.e.

$$N_i S_i \sim W(\Sigma, N_i - 1) \ (i = 1,2)$$

$$\therefore \quad (N_1^{-1} + N_2^{-1})(N_1 S_1 + N_2 S_2) \sim W[(N_1^{-1} + N_2^{-1})\Sigma, (N_1 + N_2 - 2)]$$

and is independent of $\overline{x}_d$.

From the definition of Hotelling's T^2-statistics, we get that

$$(N_1 + N_2 - 2)(\overline{x}_d)'[(N_1^{-1} + N_2^{-1})(N_1 S_1 + N_2 S_2)]^{-1} \overline{x}_d$$

is distributed as $T^2(p, N_1 + N_2 - 2)$.

We may note that $D^2 = (\overline{x}_d)' S^{-1} \overline{x}_d$ and

$$(N_1 + N_2 - 2)(\overline{x}_d)'[(N_1^{-1} + N_2^{-1})(N_1 S_1 + N_2 S_2)]^{-1} \overline{x}_d = (N_1^{-1} + N_2^{-1})^{-1} (\overline{x}_d)' S^{-1} \overline{x}_d$$

Hence, D^2 can be transformed to T^2, by the relation

$$T^2(p, N_1 + N_2 - 2) = \frac{N_1 N_2}{N_1 + N_2} D^2$$

Therefore, the significance of the hypothesis $H_0: \mu_1 = \mu_2$ is tested by the statistic

$$F = \frac{N_1 + N_2 - p - 1}{(N_1 + N_2 - 2)p} T^2\{p, (N_1 + N_2 - 2)\}$$

$$= \frac{N_1 + N_2 - p - 1}{(N_1 + N_2 - 2)p} \left(\frac{N_1 N_2}{N_1 + N_2}\right) [(\overline{x_1} - \overline{x_2})' S^{-1} (\overline{x_1} - \overline{x_2})]$$

which follows F-distribution with $[p, (N_1 + N_2 - p - 1)]$ degrees of fieedom. Therefore, the test procedure would be to reject the null hypothesis H_0 if calculated value of the above statistics is greater than tabulated value of F-statistics at specified level of significance and above-mentioned degree of freedom.

Example 6.3.2.: The following data show the number of ever born children and dead children to a number of couples belonging to low and medium socio-economic status:

Low Socio Economic		Medium Socio Economic	
Number of Children		Number of Children	
10	3	2	0
3	0	5	0
5	0	4	1
2	0	6	1
12	2	3	0
1	0	4	0
8	1	5	1
7	2	10	2
4	0	8	3
2	0	6	1
1	0	7	0

5	1	6	0
4	0	5	0
6	2	4	0
7	1	8	1
4	1	7	2
8	2	3	0
3	1	2	0
6	1	3	0
5	0	1	0
		2	0
		4	1
		5	1
		4	0
		6	0
		9	0
		6	0
		5	0

$H_0: \mu^{(1)} = \mu^{(2)}$ against $H_1: \mu^{(1)} \neq \mu^{(2)}$ assuming that $\Sigma = \Sigma_1 = \Sigma_2$ (unknown)

Here $N_1 = 20$, $N_2 = 28$ and $p = 2$.

$\bar{x}^{(1)} = \begin{bmatrix} 5.15 \\ 0.85 \end{bmatrix}$, $\bar{x}^{(2)} = \begin{bmatrix} 5.00 \\ 0.50 \end{bmatrix}$

Where

$$\bar{x}^{(1)} = \frac{1}{N_1} \sum_{\alpha=1}^{N_1} x_\alpha^{(1)}$$

$$\bar{x}^{(2)} = \frac{1}{N_2} \sum_{\alpha=1}^{N_2} x_\alpha^{(2)}$$

$$S_1 = \begin{bmatrix} 8.555 & 2.181 \\ 2.181 & 0.871 \end{bmatrix}, S_2 = \begin{bmatrix} 4.888 & 0.962 \\ 0.962 & 0.629 \end{bmatrix}$$

where

$$S_1 = \frac{1}{N_1 - 1} \sum_{\alpha=1}^{N_1} \left(x_\alpha^{(1)} - x^{(1)}\right)\left(x_\alpha^{(1)} - x^{(1)}\right)'$$

$$S_2 = \frac{1}{N_2 - 1} \sum_{\alpha=1}^{N_2} \left(x_\alpha^{(2)} - x^{(2)}\right)\left(x_\alpha^{(2)} - x^{(2)}\right)'$$

The pooled dispersion matrix is

$$S = \frac{1}{N_1 + N_2 - 2}\left[\sum_{\alpha=1}^{N_1} \left(x_\alpha^{(1)} - x^{(1)}\right)\left(x_\alpha^{(1)} - x^{(1)}\right)' + \sum_{\alpha=1}^{N_2} \left(x_\alpha^{(2)} - x^{(2)}\right)\left(x_\alpha^{(2)} - x^{(2)}\right)'\right]$$

$$S = \begin{bmatrix} 6.403 & 0.068 \\ 0.068 & 0.729 \end{bmatrix}$$

And

$$S^{-1} = \begin{bmatrix} 0.156 & -0.014 \\ -0.014 & 1.373 \end{bmatrix}$$

$$D^2 = \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right)' S^{-1} \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right) = [0.15 \quad 0.35] \begin{bmatrix} 0.156 & -0.014 \\ -0.014 & 1.373 \end{bmatrix} \begin{bmatrix} 0.15 \\ 0.35 \end{bmatrix} = 0.17025$$

Hence

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} D^2 = \frac{20 \times 28}{20 + 28} \times 0.17025 = 1.986$$

$$F = \frac{N_1 + N_2 - p - 1}{p(N_1 + N_2 - 2)} T^2 = \frac{20 + 28 - 2 - 1}{2(20 + 28 - 2)} \times 1.986 = 0.971$$

The tabulated value $F(2, 45)$ at 5% level of significance is 3.27. Since calculated value is less than tabulated value at 5% level of significance, we conclude that data provide no evidence for rejection of hypothesis.

6.4 Fisher-Behrens Problem

The Fisher-Behrens Problem is the problem of testing for the equality of means from two multivariate normal distributions when the covariance matrices are unknown and possibly not equal. Since this is a generalization of the univariate Fisher-Behrens Problem, it inherits all the difficulties that arise in the univariate problem. The main challenge lies in the fact that unequal covariance matrices (in the multivariate case) introduce additional complexity, making traditional methods like the Student's t-test inapplicable.

6.4.1 Two samples Problem

Let $x_1^{(1)}, \dots, x_{N_1}^{(1)}$ and $x_2^{(2)}, \dots, x_{N_2}^{(2)}$ be the random sample of sizes N_1 and N_2 drawn from $N(\mu^{(1)}, \Sigma_1)$ and $N(\mu^{(2)}, \Sigma_2$. Under the null hypothesis $H_0: \mu^{(1)} = \mu^{(2)}$. The mean $(x)^{-}\hat{}((1))$ of the first sample is normally distributed with expected value

$$E\left(\bar{x}^{(1)}\right) = \mu^{(1)}$$

and covariance matrix is

$$E\left(\bar{x}^{(1)} - \mu^{(1)}\right)\left(\bar{x}^{(1)} - \mu^{(1)}\right)' = \frac{1}{N_1}\Sigma_1.$$

Thus

$$\bar{x}^{(1)} \sim N\left(\mu^{(1)}, \frac{\Sigma_1}{N_1}\right)$$

The mean $\bar{x}^{(2)}$ of the second sample is normally distributed with expected value

$$E\left(\bar{x}^{(2)}\right) = \mu^{(2)}$$

and covariance matrix is

$$E\left(\bar{x}^{(2)} - \mu^{(2)}\right)\left(\bar{x}^{(2)} - \mu^{(2)}\right)' = \frac{1}{N_2}\Sigma_2$$

Therefore

$$\bar{x}^{(2)} \sim N\left(\mu^{(2)}, \frac{\Sigma_2}{N_2}\right)$$

Thus,

$$\left(\bar{x}^{(1)} - \bar{x}^{(2)}\right) \sim N\left[\mu^{(1)} - \mu^{(2)}, \left(\frac{\Sigma_1}{N_1} + \frac{\Sigma_2}{N_2}\right)\right]$$

If $N_1 = N_2 = N$

Let

$$y_\alpha = x_\alpha^{(1)} - x_\alpha^{(2)}$$

Then

$$y_\alpha \sim N(0, \Sigma_1 + \Sigma_2)$$

$$\Rightarrow \bar{y} = \frac{1}{N}\sum_{\alpha=1}^{N} y_\alpha$$

$$= \left(\bar{x}^{(1)} - \bar{x}^{(2)}\right) \sim N\left(0, \frac{\Sigma_1 + \Sigma_2}{N}\right)$$

$$\Rightarrow \sqrt{N}\bar{y} \sim N(0, \Sigma_1 + \Sigma_2)$$

Let

$$S_y = \frac{1}{N-1}\sum_{\alpha=1}^{N}\left(y_\alpha - \underline{\bar{y}}\right)\left(y_\alpha - \underline{\bar{y}}\right)'$$

$$\Rightarrow (N-1)S = \sum_{\alpha=1}^{N-1} Z_\alpha Z_\alpha{}'$$

By definition,

$$T^2 = N\bar{y}' S_y^{-1} \bar{y} \sim T^2_{(N-1)}$$

The critical region is

$$T^2 \geq \frac{(N-1)p}{N-p} F_{p,(N-p)}\,(\alpha)$$

If $N_1 \neq N_2$ and $N_1 < N_2$.

Define,

$$y_\alpha = x_\alpha^{(1)} - \sqrt{\frac{N_1}{N_2}}\, x_\alpha^{(2)} + \frac{1}{\sqrt{N_1\, N_2}} \sum_{\beta=1}^{N_1} x_\beta^{(2)} - \frac{1}{N_2} \sum_{\gamma=1}^{N_2} x_\gamma^{(2)}$$

Then

$$E(y_\alpha) = \mu^{(1)} - \sqrt{\frac{N_1}{N_2}}\, \mu^{(2)} + \frac{1}{\sqrt{N_1\, N_2}} \sum_{\beta=1}^{N_1} \mu^{(2)} - \frac{1}{N_2} \sum_{\gamma=1}^{N_2} \mu^{(2)}$$

$$= \mu^{(1)} - \sqrt{\frac{N_1}{N_2}}\, \mu^{(2)} + \sqrt{\frac{N_1}{N_2}}\, \mu^{(2)} - \mu^{(2)}$$

$$= \mu^{(1)} - \mu^{(2)}$$

The covariance matrix of y_α and y_β is

$$E[y_\alpha - E(y_\alpha)][y_\beta - E(y_\beta)]'$$

$$= E\left[\left(x_\alpha^{(1)} - \mu^{(1)}\right) - \sqrt{\frac{N_1}{N_2}}\left(x_\alpha^{(2)} - \mu^{(2)}\right) + \frac{1}{\sqrt{N_1\, N_2}} \sum_{\beta=1}^{N_1}\left(x_\beta^{(2)} - \mu^{(2)}\right) - \frac{1}{N_2} \sum_{\gamma=1}^{N_2}\left(x_\gamma^{(2)} - \mu^{(2)}\right)\right]$$

$$\left[\left(x_\alpha^{(1)} - \mu^{(1)}\right)' - \sqrt{\frac{N_1}{N_2}}\left(x_\alpha^{(2)} - \mu^{(2)}\right)' + \frac{1}{\sqrt{N_1\, N_2}} \sum_{\beta=1}^{N_1}\left(x_\beta^{(2)} - \mu^{(2)}\right)' - \frac{1}{N_2} \sum_{\gamma=1}^{N_2}\left(x_\gamma^{(2)} - \mu^{(2)}\right)'\right]$$

$$= E\left(x_\alpha^{(1)} - \mu^{(1)}\right)\left(x_\alpha^{(1)} - \mu^{(1)}\right)' + \frac{N_1}{N_2} E\left(x_\alpha^{(2)} - \mu^{(2)}\right)\left(x_\alpha^{(2)} - \mu^{(2)}\right)'$$

$$+ \frac{1}{N_1 N_2} \sum_{\beta=1}^{N_1}\left(x_\beta^{(2)} - \mu^{(2)}\right)\left(x_\beta^{(2)} - \mu^{(2)}\right)' + \frac{1}{N_2^2} \sum_{\gamma=1}^{N_2}\left(x_\gamma^{(2)} - \mu^{(2)}\right)\left(x_\gamma^{(2)} - \mu^{(2)}\right)'$$

$$-2\sqrt{\frac{N_1}{N_2}}\frac{1}{\sqrt{N_1\,N_2}}E\left(x_\alpha^{(2)}-\mu^{(2)}\right)\left(x_\alpha^{(2)}-\mu^{(2)}\right)'+2\sqrt{\frac{N_1}{N_2}}\frac{1}{N_2}E\left(x_\alpha^{(2)}-\mu^{(2)}\right)\left(x_\alpha^{(2)}-\mu^{(2)}\right)'$$

$$-2\frac{1}{\sqrt{N_1\,N_2}}\frac{1}{N_2}\sum_{\beta=1}^{N_1}E\left(x_\beta^{(2)}-\mu^{(2)}\right)\left(x_\beta^{(2)}-\mu^{(2)}\right)'$$

+terms having expectation zero

$$=\Sigma_1+\frac{N_1}{N_2}\Sigma_2+\frac{1}{N_1N_2}N_1\Sigma_2+\frac{1}{N_2^2}N_2\,\Sigma_2-2\sqrt{\frac{N_1}{N_2}}\,\frac{1}{\sqrt{N_1\,N_2}}\,\Sigma_2+2\sqrt{\frac{N_1}{N_2}}\left(\frac{1}{N_2}\right)\Sigma_2$$

$$-2\frac{1}{\sqrt{N_1\,N_2}}\left(\frac{1}{N_2}\right)N_1\Sigma_2$$

$$=\Sigma_1+\left[\frac{N_1}{N_2}+\frac{1}{N_2}+\frac{1}{N_2}\,-\frac{2}{N_2}+2\sqrt{\frac{N_1}{N_2}}\left(\frac{1}{N_2}\right)-2\sqrt{\frac{N_1}{N_2}}\left(\frac{1}{N_2}\right)\right]\Sigma_2$$

$$=\Sigma_1+\frac{N_1}{N_2}\Sigma_2$$

Hence, under H_0

$$y_\alpha\sim N\left(0,\Sigma_1+\frac{N_1}{N_2}\Sigma_2\right)$$

$$\Rightarrow\bar{y}=\frac{1}{N_1}\sum_{\alpha=1}^{N_1}y_\alpha\;\sim N\left[0,\frac{1}{N_1}\left(\Sigma_1+\frac{N_1}{N_2}\Sigma_2\right)\right]$$

$$\Rightarrow\sqrt{N_1}\bar{y}\sim N\left(0,\Sigma_1+\frac{N_1}{N_2}\Sigma_2\right)$$

Let

$$S = \frac{1}{N_1 - 1}\sum_{\alpha=1}^{N_1}(y_\alpha - \bar{y})(y_\alpha - \bar{y})'$$

$$\Rightarrow (N_1 - 1)S = \sum_{\alpha=1}^{N_1-1} Z_\alpha {Z_\alpha}'$$

By definition,

$$T^2 = N_1\, \underline{\bar{y}}' S_{\bar{y}}^{-1} \underline{\bar{y}} \sim T^2_{(N_1-1)}$$

The critical region is

$$T^2 \geq \frac{(N_1 - 1)p}{N_1 - p} F_{p,(N_1-p)}\,(\alpha)$$

6.4.2 *k*-Sample Problem

Let $x_\alpha^{(i)}$ $(\alpha = 1,2,\dots,N_i\,; i = 1,2,\dots,k)$ be the random sample from $N(\mu^{(i)}, \Sigma_i$ respectively. Under the null hypothesis

$$H_0: \sum_{i=1}^{k} \beta_i \mu^{(i)} = \mu$$

where $\beta_1, \beta_2, \dots., \beta_k$ are given scalars and μ is a given vector. If N_i are unequal take N_1 to be the smallest.

Define,

$$y_\alpha = \beta_1 x_\alpha^{(1)} + \sum_{i=2}^{k} \beta_i \sqrt{\frac{N_1}{N_2}} \left[x_\alpha^{(i)} - \frac{1}{N_1}\sum_{\beta=1}^{N_1} x_\beta^{(i)} + \frac{1}{\sqrt{N_1\, N_i}} \sum_{\gamma=1}^{N_i} x_\gamma^{(i)} \right] \quad , \alpha = 1,2,\dots., N_1$$

Then

$$E(y_\alpha) = E\left[\beta_1 x_\alpha^{(1)} + \sum_{i=2}^{k} \beta_i \sqrt{\frac{N_1}{N_2}} \left\{ x_\alpha^{(i)} - \frac{1}{N_1} \sum_{\beta=1}^{N_1} x_\beta^{(i)} + \frac{1}{\sqrt{N_1\, N_i}} \sum_{\gamma=1}^{N_i} x_\gamma^{(i)} \right\}\right]$$

$$= \beta_1 \mu^{(1)} + \sum_{i=2}^{k} \beta_i \sqrt{\frac{N_1}{N_2}} \left[\mu^{(i)} - \frac{1}{N_1} N_1 \mu^{(i)} + \frac{1}{\sqrt{N_1\, N_i}} N_i \mu^{(i)}\right]$$

$$= \beta_1 \mu^{(1)} + \sum_{i=2}^{k} \beta_i\, \mu^{(i)}$$

$$= \sum_{i=1}^{k} \beta_i\, \mu^{(i)}$$

The covariance matrix of y_α and y_β is

$$E[y_\alpha - E(y_\alpha)]\left[y_\beta - E\left(y_\beta\right)\right]' = \sum_{i=1}^{k} \frac{N_1 \beta_i^2}{N_i} \Sigma_i$$

Hence, under H_0

$$y_\alpha \sim N\left(\mu, \sum_{i=1}^{k} \frac{N_1 \beta_i^2}{N_i} \Sigma_i\right)$$

And

$$\bar{y} = \frac{1}{N_1} \sum_{\alpha=1}^{N_1} y_\alpha \sim N\left(\mu, \sum_{i=1}^{k} \frac{N_1 \beta_i^2}{N_i} \Sigma_i\right)$$

$$\Rightarrow \sqrt{N_1}(\bar{y} - \mu) \sim N\left(0, \sum_{i=1}^{k} \frac{N_1 \beta_i^2}{N_i} \Sigma_i\right)$$

Let

$$S = \frac{1}{N_1 - 1}\sum_{\alpha=1}^{N_1}(y_\alpha - \bar{y})(y_\alpha - \bar{y})'$$

$$\Rightarrow (N_1 - 1)S = \sum_{\alpha=1}^{N_1}(y_\alpha - \bar{y})(y_\alpha - \bar{y})'$$

By definition,

$$T^2 = N_1\,(\bar{y} - \mu)'S^{-1}(\bar{y} - \mu) \sim T^2_{(N_1-1)}$$

The critical region is

$$T^2 \geq \frac{(N_1 - 1)p}{N_1 - p} F_{p,(N_1-p)}\,(\alpha)$$

6.5 Mahalanobis D^2

The **Mahalanobis** D^2 is a distance measure used in statistics to calculate the distance between a point and a distribution, or between two distributions, in a multivariate space. It accounts for correlations between variables, making it more general and robust than the standard Euclidean distance in high-dimensional spaces.

Given the mean vectors μ_1 and μ_2 of two multivariate normal populations and their common covariance matrix Σ, the quantity $\Delta^2 = \left(\mu^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$ was proposed by Mahalanobis as a measure of the divergence or distance between the two population $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$. The corresponding sample quantity, obtained by replacing $(\mu_1 - \mu_2)$ and Σ by their sample estimates is denoted by D^2, which is given by

$D^2 = (\bar{x} - \bar{y})'S^{-1}(\bar{x} - \bar{y})$ and is known as Mahalanobis Studentized D^2.

Here

$$S = \frac{(N_1 - 1)S^{(1)} + (N_2 - 1)S^{(2)}}{N_1 + N_2 - 2}$$

$$S^{(1)} = \frac{1}{N_1 - 1} \sum_{\alpha=1}^{N_1} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'$$

$$S^{(2)} = \frac{1}{N_2 - 1} \sum_{\alpha=1}^{N_2} (y_\alpha - \bar{y})(y - \bar{y})'$$

Obviously

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} D^2$$

So that T^2 and D^2 are almost same, except for the constant $\frac{N_1 N_2}{N_1 + N_2}$.

Let

$$U = \frac{N_1 N_2}{N_1 + N_2} (\bar{x} - \bar{y})$$

Then the expected value of U is

$$E(U)$$

$$= \frac{N_1 N_2}{N_1 + N_2} \left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= \delta \text{ (say)}$$

The variance covariance matrix of$U$ is

$$\Sigma_U$$

$$= \left(\frac{N_1 N_2}{N_1 + N_2}\right)^2 E\left[(\bar{x} - \bar{y}) - \left(\mu^{(1)} - \mu^{(2)}\right)\right]\left[(\bar{x} - \bar{y}) - \left(\mu^{(1)} - \mu^{(2)}\right)\right]'$$

$$= \left(\frac{N_1 N_2}{N_1 + N_2}\right)^2 \left(\frac{1}{N_1} + \frac{1}{N_2}\right)^2 \Sigma$$

$$= \left(\frac{N_1 N_2}{N_1 + N_2}\right)^2 \left(\frac{N_1 + N_2}{N_1 N_2}\right)^2 \Sigma$$

$$= \Sigma$$

Therefore, $U \sim N(\delta, \Sigma)$. Further

$$\left(\frac{N_1 N_2}{N_1 + N_2}\right)^2 D^2 = U' S^{-1} U$$

Since Σ is positive definite matrix there exist a non-singular matrix C such that

$$C\Sigma C' = I$$

$$\Longrightarrow CC' = \Sigma^{-1}$$

Define

$$U^* = CU,$$

$$S^* = CSC'$$

and

$$\delta^* = C\delta.$$

Then,

$$\left(\frac{N_1 N_2}{N_1 + N_2}\right)^2 D^2 = U^{*'} S^{*^{-1}} U^*,$$

and the expected value is

$$E(U^*) = CE(U)$$

$$= C\delta$$

$$= \delta^*$$

The variance covariance matrix is

$$\Sigma_{U^*} = CE[U - E(U)][U - E(U)]'C'$$

$$= C\Sigma C'$$

$$= I$$

Thus,

$U^* \sim N(\delta^*, I)$,

$\Rightarrow U^{*'} U^* \sim \chi_p^2(\delta^{*'} \delta^*)$

where

$$\delta^{*'} \delta^* = \delta' C' C \delta$$

$$= \delta' \Sigma^{-1} \delta$$

$$= \lambda^2$$

Let

$$(N_1 + N_2 - 2)S = \sum_{\alpha=1}^{N_1+N_2-2} (Z_\alpha)(Z_\alpha)'$$

Where $Z_\alpha \sim N(0, \Sigma)$. Then

$$(N_1 + N_2 - 2)S^* = \sum_{\alpha=1}^{N_1+N_2-2} (C\, Z_\alpha)(C\, Z_\alpha)'$$

$(C\, Z_\alpha) \sim N(0, I)$.

Therefore, under H_0

$$\left(\frac{N_1 N_2}{N_1 + N_2}\right)^2 D^2 = Z^{*'} S^{*^{-1}} Z^*$$

$$= (N_1 + N_2 - 2) \frac{\chi_p^2(\lambda^2)}{\chi^2_{N_1+N_2-p-1}}$$

$$\frac{N_1 + N_2 - p - 1}{p} \left(\frac{N_1 N_2}{N_1 + N_2}\right) \frac{D^2}{N_1 + N_2 - 2}$$

$$= \frac{\chi_p^2(\lambda^2)/p}{\chi^2_{N_1+N_2-p-1}/N_1 + N_2 - p - 1} \sim F_{p, N_1+N_2-p-1}$$

6.6 Applications

1. Outlier Detection: It helps identify outliers in multivariate data, which are points that are far away from the center of the distribution.

2. Classification: It is used in classification algorithms, such as discriminant analysis, to determine the class membership of a new observation.

3. Clustering: It is used in clustering algorithms, such as k-means and hierarchical clustering, to determine the similarity between observations.

4. Dimensionality Reduction: It is used in dimensionality reduction techniques, such as principal component analysis, to select the most important variables.

5. Anomaly Detection: It is used in anomaly detection to identify data points that are far away from the normal data points.

6. Quality Control: It is used in quality control to detect deviations in multivariate data.

7. Image Processing: It is used in image processing to detect outliers and anomalies in images.

8. Finance: It is used in finance to detect fraud and anomalies in financial data

6.7 Summary

In this unit, we have covered the concepts of testing of hypothesis regarding population mean vector under following situations:

1. Test for mean vector for two independent sample cases when population covariances are known.
2. Test for mean vector for two independent sample cases when population covariances are equal but unknown.
3. Discuss about Fisher-Behrens Problems.
4. Also discuss Mahalanobis D^2 and its application.

6.8 Self-Assessment Exercises

1. Explain the equality component of mean vector when covariance matrix is known.
2. Explain the equality component of mean vector when covariance matrix is unknown.
3. Two samples of size 50 bars and 60 bars were taken from the lots produced by method 1 and method 2. Two characteristics X_1 = lather and X_2 = mildness were measures. The summary statistics for bars produced by methods 1 and 2 is given by

   $\bar{x}^{(1)} = \begin{bmatrix} 8 \\ 4 \end{bmatrix}, \bar{x}^{(2)} = \begin{bmatrix} 10 \\ 4 \end{bmatrix}$

   $S_1 = \begin{bmatrix} 2 & 1 \\ 1 & 5 \end{bmatrix}, S_2 = \begin{bmatrix} 2 & 1 \\ 1 & 6 \end{bmatrix}$

   Test at 5% level of significance whether $\mu^{(1)} = \mu^{(2)}$ or not.

4. Derive the relation between T^2 and D^2.
5. Define Mahalanobis D^2 and its application.

6.9 References

- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Johnson, R. A., Wichern, D. W. (2019): Applied Multivariate Statistical Analysis. United Kingdom: Pearson
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.
- Giri Narayan C. (1995): Multivariate Statistical Analysis.
- Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York.

6.10 Further Reading

- Khatri C G.: Multivariate Analysis.
- Mardia K V.: Multivariate Analysis.
- Seber G.A.F.: *Multivariate Observations*, Wiley, New York.

UNIT - 7: DISCRIMINANT ANALYSIS

Structure

7.1 Introduction

Discriminant Analysis is a statistical technique used for classifying a set of observations into predefined classes or groups based on predictor variables. It is primarily used when the dependent variable is categorical (i.e., it represents groups or classes), and the independent variables are continuous or interval in nature. The objective of discriminant analysis is to build a predictive model that best separates the groups based on the independent variables.

The key components of discriminant analysis are

(i) Dependent Variable (Grouping Variable): This is categorical and defines the groups or classes into which the data will be classified. For example, the type of customer (high, medium, low).

(ii) Independent Variables (Predictor Variables): These are continuous or interval variables used to differentiate between the groups. For example, Age, income, expenditure, etc. on the basis of which one may categorize the type of customer.

We may broadly divide the Discriminant Analysis into two types, (i) the Linear Discriminant Analysis (LDA), which is used when there are two or more groups, and it assumes that the independent variables are normally distributed and that the variance-covariance matrices of each group are equal. Here the goal is to find a linear combination of the independent variables that best separates the groups., and (ii) Multiple Discriminant Analysis (MDA), which is an extension of LDA that deals with more than two groups. Here, multiple discriminant functions are created to best separate the observations into multiple classes and each discriminant function maximizes the separation between the groups.

7.2 Objectives

Upon completion of this unit, you should be able to:

- Determine whether linear or quadratic discriminant analysis should be applied to a given data set
- Be able to apply the linear discriminant function to classify a subject by its measurements;
- Understand how to assess the efficacy of discriminant analysis.

7.3 Discriminant Analysis

The problem of discriminant analysis deals with assigning an individual to one of several categories based on measurements on a p component vector of variable x on that individual. For example, we take certain measurements on the skull of an animal and want to know whether it was male or female, a patient is to be classified as diabetic or not, a person has to be classified as successful or unsuccessful on different psychological tests.

7.4 Classification and Discrimination Procedures for Discrimination between Two Multivariate Normal Populations

Classification $\Rightarrow$ Several measurements on an individual are available and objective is to classify the individual into one of several categories based on these measurements.

Examples:

- An anthropologist tries to identify a jaw bone excavated from a burial ground as having belonged to a male or a female.
- A doctor decides based on some diagnosis tests whether a patient suffering from Jaundice requires surgery or not.
- A biologist wants to identify an observed specimen as a member of one out of k possible known spices.
- Based on gene expression data, one wants to classify the stage of a cancer patient.

In above examples, the decision must be taken among several alternative hypotheses.

We assume that there are finite numbers of categories or populations for which an individual may have come and each population is characterized by a probability distribution of the measurements. An individual is considered as a random observation from this population. Given an individual with certain measurements, the decision must be taken regarding the population from which the individual arise.

If probability distributions of the measurements are completely known then the categories are specified beforehand.

If of each distribution may be known, but the parameters of the distribution are unknown, we use estimators of parameters from a sample from that population.

7.4.1 Standards of Good Classification

Minimize the probability of misclassification or, equivalently, it is desired to minimize the bad effects of misclassification.

Let an individual be an observation from either population Π_1 or population Π_2. $x' = (x_1, \dots, x_p)$: Vector of observations on that individual.

We set up a rule that if an individual is characterized by certain sets of values of $x_1, \dots, x_p$ he will be classified as from Π_1; if he has other values he is classified as from Π_2.

We consider the observation vector x as a point in a p-dimensional space. We divide the space into two regions, say, R_1 and R_2. If $x \in R_1$, we classify it as coming from population Π_1, and if $x \in R_2$, we classify it as coming from Π_2.

7.4.2 The Two Kinds of Error

I. The individual is from Π_1 but classified as coming from the population Π_2.

II. The individual is from Π_2 but classified as coming from the population Π_1.

$C(2|1)(> 0)$: Cost of first kind of misclassification

$C(1|2)(> 0)$: Cost of second kind of misclassification

Population	Statistician's Decision		
		Π_1	Π_2
	Π_1	0	$C(2\|1)$
	Π_2	$C(1\|2)$	0

7.4.3 Two Cases of Two Populations

Case I: Assume that prior probabilities of two populations are given:

Let

q_i: probability that an individual comes from Π_i $(i = 1,2)$

$p_1(x)$: density of Π_1

$p_2(x)$: density of Π_2

R_1: region of classification as from Π_1

R_2: region of classification as from Π_2

Then, probability of correctly classifying an observation that is drawn from population Π_1, is

$$P(1|1,R) = \int_{R_1} p_1(x)dx$$

Probability of misclassification of an observation from Π_1 is

$$P(2|1,R) = \int_{R_2} p_1(x)dx$$

Probability of correctly classifying an observation from Π_2 is

$$P(2|2,R) = \int_{R_2} p_2(x)dx$$

Probability of misclassifying an observation from Π_2 is

$$P(1|2,R) = \int_{R_1} p_2(x)dx$$

The probability of drawing an observation from Π_i is $\mathrm{q}_i; i = 1,2$.

$\mathrm{q}_i = P(i|i,R)$= Probability of drawing an observation from Π_i and correctly classifying it ($i = 1,2$).

$\mathrm{q}_2 P(1|2,R)$= Probability of drawing an observation from Π_2 and misclassifying it.

$\mathrm{q}_1 P(2|1,R)$= Probability of drawing an observation from Π_1 and misclassifying it.

Expected cost of misclassification=$C(2|1)P(2|1,R)\mathrm{q}_1 + C(1|2)P(1|2,R)\mathrm{q}_2$.

We wish to divide the space R into regions $\mathrm{R}_1$and R_2 such that expected loss is as small as possible. A procedure that minimizes expected cost is called a Bayes procedure.

Case II: No known a priori probabilities:

The expected loss if the observation is from Π_1 is

$C(2|1)P(2|1,R) = r(1,R)$

and the expected loss if the observation is from Π_2 is

$C(1|2)P(1|2,R) = r(2,R)$

We do not know whether the observation is from Π_1 or from Π_2, and we do not know probabilities of these two instances.

For two procedure R and R*, R is said to be as good as R*if

$r(1,R) \leq r(1,\mathrm{R}^*)$ and $r(2,R) \leq r(2,\mathrm{R}^*)$.

If at least one of these inequalities is strict, then R is said to be better than R*. Usually, there is no procedure which is better than or as good as all other procedures.

7.4.4 Some Definitions

(1) For two procedure R and R*, R is said to be as good as R*if

$$r(1,R) \leq r(1,\mathrm{R}^*) \text{ and } r(2,R) \leq r(2,\mathrm{R}^*)$$

If at least one of these inequalities is strict, then R is said to be better than R*. Usually, there is no procedure that is better than or as good as all other procedures.

(2) A procedure R is said to be admissible if there is no procedure better than R. Under certain conditions the class of all admissible procedures is same as the class of Bayes procedures.

(3) A class of procedure is complete if for any procedure outside this class there is one in the class which is better.

(4) A minimal complete class (if it exists) is a complete class such that no proper subset of it is a complete class. A similar definition holds for a minimal essentially complete class. Under certain conditions we shall show that the admissible class is minimal complete.

(5) A procedure is minimax if the maximum expected loss, $r(i,R)$, is a minimum.

(6) A class is called of procedures is essentially complete if for any procedure outside the class, there is at least one in the class which is at least as good.

7.4.5 Procedure of Classification into One of Two Populations with Known Probability Distribution

Case I: When A Priori Probabilities are Known

We now turn to the problem of choosing regions $R_1 and R_2$ so as to minimize the expected loss.

Theorem 7.4.1: If q_1 and q_2 are a priori probabilities of drawing an observation from population Π_1 with density $p_1(x)$ and Π_2 with density $p_2(x)$ respectively, and if the cost of misclassifying an observation from Π_1 as from Π_2 is $C(2|1)$ and an observation Π_2 as from Π_1 is $C(1|2)$. Consider the regions of classification $R_1 and R_2$, defined by

$R_1: [C(2|1)q_1]p_1(x) \geq [C(1|2)q_2]p_2(x)$

$R_2: [C(2|1)q_1]p_1(x) < [C(1|2)q_2]p_2(x)$

Then these regions minimize the expected loss.

If

$$P\left\{\frac{p_1(x)}{p_2(x)} = \frac{q_2 C(1|2)}{q_1 C(2|1)}\middle| \Pi_i\right\} = 0; i = 1,2$$

then the procedure is unique except for sets of probability zero.

Proof: The posterior probability that an x observation came from Π_1 is

$$\frac{q_1 p_1(x)}{q_1 p_1(x) + q_2 p_2(x)}$$

Similarly, the posterior probability that an observation x came from Π_2 is

$$\frac{q_2 p_2(x)}{q_1 p_1(x) + q_2 p_2(x)}$$

For a moment, we take $C(1|2) = C(2|1) = 1$. Then the expected loss is

$$q_1 \int p_1(x)dx + q_2 \int p_2(x)dx$$

This is also the probability of misclassification and we have to minimize it.

For a given observed point x we minimize the probability of misclassification by assigning it to the population that has the higher posterior probability. Thus, if

$$\frac{q_1 p_1(x)}{q_1 p_1(x) + q_2 p_2(x)} \geq \frac{q_2 p_2(x)}{q_1 p_1(x) + q_2 p_2(x)}$$

we choose population Π_1 otherwise Π_2. Since we minimize the probability of misclassification at each point, we minimize it over the whole space. Thus, the rule is

$R_1: q_1 p_1(x) \geq q_2 p_2(x)$

$R_2: q_1 p_1(x) < q_2 p_2(x)$

If $q_1 p_1(x) = q_2 p_2(x)$, the point could be classified as either from Π_1 or Π_2. We have arbitrarily put it into R_1.

If $q_1 p_1(x) + q_2 p_2(x) = 0$ for a given x, that point also may go into either region.

To show that this procedure is best, we consider any other procedure, which partitions R into (R_1^*, R_2^*).

The probability of misclassification is

$$q_1 \int_{R_2^*} p_1(x)dx + q_2 \int_{R_1^*} p_2(x)dx = \int_{R_2^*} [q_1 p_1(x) - q_2 p_2(x)]\, dx + q_2 \int_R p(x)dx \qquad (7.1)$$

On the RHS of second term of (7.1) is a fixed number. The first term is minimized if R_2^* includes the points x such that $q_1 p_1(x) - q_2 p_2(x) < 0$ and excludes the points for which $q_1 p_1(x) - q_2 p_2(x) > 0$. If we assume that

$$P\left\{\frac{p_1(x)}{p_2(x)} = \frac{q_2}{q_1}\middle| \Pi_i\right\} = 0; i = 1,2$$

then the Bayes procedure is unique except for sets of points with probability zero.

Now, we consider the general case. The problem is to minimize

$[C(2|1)\mathrm{q}_1]\int_{R_2} p_1(x)dx + [C(1|2)\mathrm{q}_2]\int_{R_1} p_2(x)dx,$

We choose $\mathrm{R}_1 \text{and} \mathrm{R}_2$ according to

$\mathrm{R}_1: [C(2|1)\mathrm{q}_1]p_1(x) \geq [C(1|2)\mathrm{q}_2]p_2(x)$

$\mathrm{R}_2: [C(2|1)\mathrm{q}_1]p_1(x) < [C(1|2)\mathrm{q}_2]p_2(x)$

or,

$$\mathrm{R}_1: \frac{p_1(x)}{p_2(x)} \geq \frac{\mathrm{q}_2 C(1|2)}{\mathrm{q}_1 C(2|1)}$$

$$\mathrm{R}_2: \frac{p_1(x)}{p_2(x)} < \frac{\mathrm{q}_2 C(1|2)}{\mathrm{q}_1 C(2|1)}$$

Case II: When No a Priori Probabilities are Known

In this case when a priori probabilities are not given, we shall look for the class of admissible procedures.

Theorem 7.4.2.: If $P\{p_1(x) = 0|\Pi_1\} = 0 = P\{p_2(x) = 0|\Pi_2\}$ then every Bayes procedure is admissible.

Proof: Let $R = (\mathrm{R}_1, \mathrm{R}_2)$ is a Bayes procedure for a given a priori probability $\mathrm{q}_1, \mathrm{q}_2$. Since R is a Bayes procedure, for any other procedure $\mathrm{R}^* = (\mathrm{R}_1^*, \mathrm{R}_2^*)$, we have

$$\mathrm{q}_1 P(2|1, R) + \mathrm{q}_2 P(1|2, R) \leq \mathrm{q}_1 P(2|1, R^*) + \mathrm{q}_2 P(1|2, R^*) \quad (7.2)$$

$$\text{or, } \mathrm{q}_1[P(2|1, R) - P(2|1, R^*)] \leq \mathrm{q}_2[P(1|2, R^*) - P(1|2, R)] \quad (7.3)$$

Now, we have to prove that there exists no procedure R^* such that

$P(1|2, R^*) \leq P(1|2, R) \text{and} P(2|1, R^*) \leq P(2|1, R)$

If $P(1|2, R^*) \leq P(1|2, R)$ then by (7.2), if $\mathrm{q}_2 > 0, P(2|1, R) \leq P(2|1, R^*)$.

Similarly, for $\mathrm{q}_1 > 0, P(2|1, R^*) \leq P(2|1, R)$ implies $P(1|2, R) \leq P(1|2, R^*)$.

Thus $R^*$is not better than R and R is admissible. If $\mathrm{q}_1 = 0$, then (7.3) implies

$P(1|2, R^*) - P(1|2, R) \geq 0$

For a Bayes procedure, R_1 includes only points for which $p_2(x) = 0$. Therefore,

$P(1|2, R) = 0$ and if R^* is to be better $P(1|2, R^*) = 0$. If $P\{p_2(x) = 0|\Pi_1\} = 0$, then $P(2|1, R) = P\{p_2(x) > 0|\Pi_1\} = 1$.

If $P(1|2, R^*) = 0$, then R_1^* contains only those points for which $p_2(x) = 0$. Then $P(2|1, R^*) = P\{R_2^*|\Pi_1\} = P\{p_2(x) > 0|\Pi_1\} = 1$, and R^* is not better than R.

Now let us prove the converse.

Theorem 7.4.3.: If

$$P\left\{\frac{p_1(x)}{p_2(x)} = k|\Pi_i\right\} = 0; i = 1,2; 0 \leq k \leq \infty \tag{7.4}$$

then every admissible procedure is a Bayes procedure.

Proof: Under (7.4), for any q_1 Bayes procedure is unique. Moreover, the cdf of $\frac{p_1(x)}{p_2(x)}$ for Π_1 and Π_2 is continuous. Let R be an admissible procedure. Then there exists a k such that

$$P(2|1, R) = P\left\{\frac{p_1(x)}{p_2(x)} \leq k|\Pi_1\right\} = P(2|1, R^*)$$

where R^* is the Bayes procedure corresponding to $\frac{q_2}{q_1} = k \quad [i.e., q_1 = \frac{1}{1+k}, q_2 = \frac{k}{1+k}]$.

Since R is admissible,

$P(1|2, R) \leq P(1|2, R^*)$

Since, by previous theorem, R^* is admissible, $P(1|2, R) \geq P(1|2, R^*)$. By uniqueness of Bayes procedure R is the same as R^*.

Theorem 7.4.4.: If (7.4) holds, the class of Bayes procedure is minimal complete.

Proof: Let R be any procedure outside the class of Bayes Procedures. We can construct a Bayes procedure R^* so that $P(2|1, R) = P(2|1, R^*)$. Then, since R^* is admissible $P(1|2, R) \geq P(1|2, R^*)$. Furthermore, the class of Bayes procedures is minimal complete since it is identical with the class of admissible procedures.

Let us now consider the minimax procedure. Let

$P(i|j, \mathrm{q}_1) = P(i|j, R)$

where R is the Bayes procedure corresponding to q_1. $P(i|j, \mathrm{q}_1)$ is a continuous function of q_1.

As q_1 varies from 0 to 1, $P(2|1, \mathrm{q}_1)$ varies from 1 to 0 and $P(1|2, \mathrm{q}_1)$ varies from 0 to 1. Thus, there is a value of q_1, say. Q_1^*, such that $P(2|1, \mathrm{q}_1^*) = P(1|2, \mathrm{q}_1^*)$. This is the minimax solution, for if there is another procedure R^* such that

$\mathrm{Max}\{P(2|1, \mathrm{q}_1^*), P(1|2, \mathrm{q}_1^*)\} \leq P(2|1, \mathrm{q}_1^*) = P(1|2, \mathrm{q}_1^*)$

this would contradict the fact that every Bayes solution is admissible.

7.5 Sample Discriminant Function

Suppose that we have a sample $x_1^{(1)}, \ldots, x_{N_1}^{(1)}$ *and* $x_2^{(2)}, \ldots, x_{N_2}^{(2)}$ be the random sample of sizes N_1 and N_2 drawn from $N\left(\mu^{(1)}, \Sigma\right)$ and $N\left(\mu^{(2)}, \Sigma\right)$ respectively and the unbiased estimate of $\mu^{(1)}$ is

$$\overline{x}^{(1)} = \frac{1}{N_1} \sum_{\alpha=1}^{N_1} x_\alpha^{(1)}$$

and $\mu^{(2)}$ is

$$\overline{x}^{(1)} = \frac{1}{N_1} \sum_{\alpha=1}^{N_1} x_\alpha^{(1)}$$

and Σ is S defined by

$$S = \frac{1}{N_1 + N_2 - 2} \left[\sum_{\alpha=1}^{N_1} \left\{ x_\alpha^{(1)} - \overline{x}^{(1)} \right\} \left\{ x_\alpha^{(1)} - \overline{x}^{(1)} \right\}' + \sum_{\alpha=1}^{N_2} \left\{ x_\alpha^{(2)} - \overline{x}^{(2)} \right\} \left\{ x_\alpha^{(2)} - \overline{x}^{(2)} \right\}' \right]$$

Substitute these estimates for the parameters in the function $x'\delta$, Fisher's discriminant function becomes

$x' S^{-1} \left(\overline{x}^{(1)} - \overline{x}^{(2)} \right)$

This is known as sample discriminant function. The classification procedure now becomes i) Compute

$$x'S^{-1}(\overline{x}^{(1)} - \overline{x}^{(2)}) = x'\delta$$

ii) Compute

$$\frac{1}{2}(\overline{x}^{(1)} + \overline{x}^{(2)})'S^{-1}(\overline{x}^{(1)} - \overline{x}^{(2)}) = \frac{1}{2}(\overline{x}^{(1)} + \overline{x}^{(2)})'\delta$$

iii) Assign the individual with measurements x to population first or population second, according as $x'\delta - \frac{1}{2}(\overline{x}^{(1)} + \overline{x}^{(2)})'\delta$ is ≥ 0 or < 0.

7.6 Classification into One of Two Known Multivariate Normal Populations (Wald's Procedure,1944)

Suppose we have two normal populations $N(\mu^{(1)}, \Sigma) and N(\mu^{(2)}, \Sigma)$ with common variance-covariance matrix Σ. $\mu^{(1)}, \mu^{(2)}$ and Σ are known. Then

$$p_i(x) = \frac{1}{2\pi^{p/2}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\ (x - \mu^{(i)}\ )'\Sigma^{-1}(x - \mu^{(i)})\right\} \quad ; i = 1,2 \tag{7.5}$$

Theorem 7.6.1.: If Π_i has the density (7.5) (i=1,2), the best regions of classification are given by

$$\begin{cases} \mathrm{R}_1 : x'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) \geq \log k \\ \mathrm{R}_2 : x'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) < \log k \end{cases} \tag{7.6}$$

ratio of densities is

$$\frac{p_1(x)}{p_2(x)} = \exp\left\{-\frac{1}{2}\left[\left(x - \mu^{(1)}\right)'\Sigma^{-1}\left(x - \mu^{(1)}\right) - \left(x - \mu^{(2)}\right)'\Sigma^{-1}\left(x - \mu^{(2)}\right)\right]\right\} \tag{7.7}$$

The region of classification into Π_1, R_1, is the set of x's for which

$$\frac{p_1(x)}{p_2(x)} \geq k (k \text{ suitably choosen}) \tag{7.8}$$

Since logarithmic function is monotonic increasing, the inequality (7.6) can be written in terms of the logarithm of (7.8) as

$$-\frac{1}{2}\left[\left(x-\mu^{(1)}\right)'\Sigma^{-1}\left(x-\mu^{(1)}\right)-\left(x-\mu^{(2)}\right)'\Sigma^{-1}\left(x-\mu^{(2)}\right)\right]\geq \log k$$

or,

$$x'\Sigma^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)-\frac{1}{2}\left(\mu^{(1)}+\mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)\geq \log k$$

The first term is a linear function of components of x and is called the discriminant function.

If a priori probabilities q_1 and q_2 are know, then k is given by

$$k=\frac{\mathrm{q}_2 C(1|2)}{\mathrm{q}_1 C(2|1)}$$

If two population are equally likely ($\mathrm{q}_1=\mathrm{q}_2$) and costs being equal, k=1 and log k=0

Then the region of classification into Π_1 is -

$\mathrm{R}_1: x'\Sigma^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)\geq\frac{1}{2}\left(\mu^{(1)}+\mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)$

And the region of classification into Π_2 is -

$R_2: x'\Sigma^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)<\frac{1}{2}\left(\mu^{(1)}-\mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)$

If priori probabilities $\mathrm{q}_1 and \mathrm{q}_2$ are not given, we may select log k= c (say) based on making expected losses due to misclassification equal.

Theorem7.6.2.: If the Π_i have densities (7.5) (i=1,2), the minimax regions of classification are given by (7.6) where C = log k is chosen by the condition

$$C(1|2)\int_{\frac{\left(c+\frac{1}{2}\alpha\right)}{\sqrt{\alpha}}}^{\infty}\frac{1}{\sqrt{2\Pi}}e^{-\frac{1}{2}y^2}dy=C(2|1)\int_{-\infty}^{\frac{\left(c-\frac{1}{2}\alpha\right)}{\sqrt{\alpha}}}\frac{1}{\sqrt{2\Pi}}e^{-\frac{1}{2}y^2}dy$$

Proof: Let

$$U=x'\Sigma^{-1}\left(\mu^{(1)}-\mu^{(2)}\right)-\tfrac{1}{2}\left(\mu^{(1)}+\mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)}-\mu^{(2)}\right),$$

When x is distribution according to $N\left(\mu^{(1)},\Sigma\right)$, the distribution of U is normal with mean

$$E_{\pi_1}(U) = \mu^{(1)'}\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= \left(\mu^{(1)} - \frac{1}{2}\mu^{(1)} - \frac{1}{2}\mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= \frac{1}{2}\left(\mu^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= \frac{\alpha}{2}.$$

Here, the distance between $N\left(\mu^{(1)}, \Sigma\right)$ and $N\left(\mu^{(2)}, \Sigma\right)$ is

$$\alpha = \left(\mu^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

The variance covariance matrix is

$$var_{\pi_1}(U) = E_{\pi_1}[UU']$$

$$= E_{\pi_1}[U - E(U)]^2$$

$$= E_{\pi_1}[\{U - E(U)\}\{U - E(U)\}']$$

$$= E_{\pi_1}\left[\left(\mu^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\left(x - \mu^{(1)}\right)\left(x - \mu^{(1)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)\right]$$

$$= \left(\mu^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}E_{\pi_1}\left(x - \mu^{(1)}\right)\left(x - \mu^{(1)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= \left(\mu^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\Sigma\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= \left(\mu^{(1)} - \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= \alpha \text{ (distance between the two populations)}$$

Thus, $U \sim N(\frac{1}{2}\alpha, \alpha)$. if $x \sim N\left(\mu^{(1)}, \Sigma\right)$

If $x \sim N\left(\mu^{(2)}, \Sigma\right)$, then

$$E_{\pi_2}(U)$$

$$= \mu^{(2)'}\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right) - \frac{1}{2}\left(\mu^{(1)} + \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

$$= \left(\mu^{(2)} - \frac{1}{2}\mu^{(1)} - \frac{1}{2}\,\mu^{(2)}\right)' \Sigma^{-1}\left(\mu^{(1)} - \,\mu^{(2)}\right)$$

$$= \frac{1}{2}\left(\mu^{(1)} - \,\mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \,\mu^{(2)}\right)$$

$$= -\frac{\alpha}{2}$$

and variance covariance matrix is

$$var_{\pi_2}(U) = E_{\pi_2}[UU'] = E_{\pi_2}[U - E(U)]^2 = E_{\pi_2}[\{U - E(U)\}\{U - E(U)\}']$$

$$= E_{\pi_2}\left[\left(\mu^{(1)} - \,\mu^{(2)}\right)'\Sigma^{-1}\left(x - \mu^{(2)}\right)\left(x - \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \,\mu^{(2)}\right)\right]$$

$$= \left(\mu^{(1)} - \,\mu^{(2)}\right)'\Sigma^{-1}E_{\pi_2}\left(x - \mu^{(2)}\right)\left(x - \mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \,\mu^{(2)}\right)$$

$$= \left(\mu^{(1)} - \,\mu^{(2)}\right)'\Sigma^{-1}\Sigma\Sigma^{-1}\left(\mu^{(1)} - \,\mu^{(2)}\right)$$

$$= \left(\mu^{(1)} - \,\mu^{(2)}\right)'\Sigma^{-1}\left(\mu^{(1)} - \,\mu^{(2)}\right) = \alpha$$

If $x \sim N\left(\mu^{(2)}, \Sigma\right)$, then $U \sim N\left(-\frac{1}{2}\alpha, \alpha\right)$.

If the observation is from Π_1, the probability of misclassification is

$$P(2|1)$$

$$= \int_{-\infty}^{c} \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{1}{2}\left(u - \frac{1}{2}\alpha\right)^2} du$$

$$= \int_{-\infty}^{\frac{(c - \frac{1}{2}\alpha)}{\sqrt{\alpha}}} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \qquad\qquad y = \frac{u - \frac{1}{2}\alpha}{\sqrt{\alpha}} \quad \text{and } dy = \frac{du}{\sqrt{\alpha}}$$

If observation is from Π_2, the probability of misclassification is

$$P(1|2)$$

$$= \int_{c}^{\infty} \frac{1}{\sqrt{2\pi\alpha}} e^{-\frac{1}{2}\left(u + \frac{1}{2}\alpha\right)^2} du$$

$$= \int_{\frac{\left(c+\frac{1}{2}\alpha\right)}{\sqrt{\alpha}}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy \qquad y = \frac{u+\frac{1}{2}\alpha}{\sqrt{\alpha}} \quad \text{and } dy = \frac{du}{\sqrt{\alpha}}$$

For the minimax solution we choose c, so that

$$C(1|2)P(1|2) = C(2|1)P(2|1)$$

$$C(1|2) \int_{\frac{\left(c+\frac{1}{2}\alpha\right)}{\sqrt{\alpha}}}^{\infty} \frac{1}{\sqrt{2\Pi}} e^{-\frac{1}{2}y^2} dy = C(2|1) \int_{-\infty}^{\frac{\left(c-\frac{1}{2}\alpha\right)}{\sqrt{\alpha}}} \frac{1}{\sqrt{2\Pi}} e^{-\frac{1}{2}y^2} dy$$

If $C(1|2) = C(2|1)$, then $c = 0$ and the probability of misclassification is given by

$$P(1|2)$$

$$= \int_{\frac{\sqrt{\alpha}}{2}}^{\infty} \frac{1}{\sqrt{2\Pi}} e^{-\frac{1}{2}y^2} dy$$

$$= \int_{-\infty}^{\frac{\sqrt{\alpha}}{2}} \frac{1}{\sqrt{2\Pi}} e^{-\frac{1}{2}y^2} dy$$

$$= P(2|1)$$

In case the costs of misclassification are unequal, c is determined by trial and error method using normal tables.

The term

$$\delta = \Sigma^{-1}\left(\mu^{(1)} - \mu^{(2)}\right)$$

may be obtained as a solution of

$$\Sigma\delta = \left(\mu^{(1)} - \mu^{(2)}\right)$$

If we have a sample of size N either from $\Pi_1 \text{or} \Pi_2$, we use the mean of the sample and classify it as from

$N\left(\mu^{(1)}, \frac{1}{N}\Sigma\right)$ or $N\left(\mu^{(2)}, \frac{1}{N}\Sigma\right)$.

7.6.1 Classification into One of Two Multivariate Normal Populations When the Parameters are Estimated (Fisher Procedure)

Generally, $\mu^{(1)}, \mu^{(2)}$ and Σ are unknown and we infer them from the samples, one from each population and we want to use that information in classifying another observation as coming from one of the two populations.

Let a sample $x_1^{(1)}, \ldots, x_{N_1}^{(1)}$ be taken from $N(\mu^{(1)}, \Sigma)$ and another sample $x_1^{(2)}, \ldots, x_{N_2}^{(2)}$ from $N(\mu^{(2)}, \Sigma)$. Based on this information, we wish to classify the observation x as coming from Π_1 or Π_2. We estimate $\mu^{(1)}$ by

$$\overline{x}^{(1)} = \frac{1}{N_1}\sum_{\alpha=1}^{N_1} x_\alpha^{(1)},$$

and $\mu^{(2)}$ by

$$\overline{x}^{(2)} = \frac{1}{N_2}\sum_{\alpha=1}^{N_2} x_\alpha^{(2)}$$

Further, estimator of Σ, say S, is given by

$$S = \frac{1}{(N_1 + N_2 - 2)}\left[\sum_{\alpha=1}^{N_1}\left(x_\alpha^{(1)} - \overline{x}^{(1)}\right)\left(x_\alpha^{(1)} - \overline{x}^{(1)}\right)' + \sum_{\alpha=1}^{N_2}\left(x_\alpha^{(2)} - \overline{x}^{(2)}\right)\left(x_\alpha^{(2)} - \overline{x}^{(2)}\right)'\right]$$

Hence, we classify the observation into Π_1 if

$$\mathrm{R}_1: x'\mathrm{S}^{-1}\left(x^{(1)} - x^{(2)}\right) - \frac{1}{2}\left(x^{(1)} + x^{(2)}\right)'\mathrm{S}^{-1}\left(x^{(1)} - x^{(2)}\right) \geq c$$

and in Π_2 if

$$\mathrm{R}_2: x'\mathrm{S}^{-1}\left(x^{(1)} - x^{(2)}\right) - \frac{1}{2}\left(x^{(1)} + x^{(2)}\right)'\mathrm{S}^{-1}\left(x^{(1)} - x^{(2)}\right) < c$$

c is chosen suitably.

The first term is the Fisher's discriminant function based on two samples

Suppose we have a sample $x_1, \dots, x_N$ from either $\Pi_1 \text{or} \Pi_2$ and we wish to classify the sample as a whole. Then we define S by

$$S = \frac{1}{(N_1 + N_2 + N - 3)} \left[\sum_{\alpha=1}^{N_1} \left(x_\alpha^{(1)} - \overline{x}^{(1)} \right) \left(x_\alpha^{(1)} - \overline{x}^{(1)} \right)' + \sum_{\alpha=1}^{N_2} \left(x_\alpha^{(2)} - \overline{x}^{(2)} \right) \left(x_\alpha^{(2)} - \overline{x}^{(2)} \right)' + \sum_{\alpha=1}^{N} (x_\alpha - \overline{x}) (x_\alpha - \overline{x})' \right]$$

7.7 Rao's U-Statistic

It is used in the context of discriminant analysis to determine whether adding a set of additional variables (q variables) to an existing set (p variables) enhances the discrimination between two populations.

A measure used to test whether the addition of a new set of variables provides significant additional discriminative power.

7.7.1 Procedure

1. **Initial Setup**:
    - Let X_p be the initial set of p variables.
    - Let X_q be the additional set of q variables.
2. **Hypotheses**:
    - Null Hypothesis (H_0): The additional set of q variables do not provide significant additional discrimination.

- Alternative Hypothesis (H_1): The additional set of q variables provide significant additional discrimination.

3. **Calculate Discriminant Functions**:
    - Construct discriminant functions using X_p and $X_p + X_q$.
4. **Compute Rao's U-Statistic**:
    - Rao's U-Statistic is calculated based on the eigenvalues of the covariance matrices of the discriminant functions.
5. **Test Statistic**:
    - The U-Statistic follows a chi-square distribution under the null hypothesis.
    - Compare the calculated U-Statistic to the critical value from the chi-square distribution with appropriate degrees of freedom.
6. **Decision Rule**:
    - If the U-Statistic is greater than the critical value, reject H_0, suggesting that the additional variables provide significant additional discrimination.
    - If the U-Statistic is less than or equal to the critical value, do not reject H_0, suggesting that the additional variables do not provide significant additional discrimination.

7.7.2 Benefits

- **Reduction in Computational Load**: By identifying unnecessary variables, the computational complexity of the analysis can be reduced.
- **Efficiency in Discrimination**: Focuses on variables that contribute most to the discrimination, enhancing the effectiveness of the model.

Rao's methodology is particularly useful in high-dimensional data where the number of variables can be large, and it is important to determine the most informative subset for effective discrimination.

7.7.3 Mathematical Formulation

The mathematical formulation of Rao's U-Statistic in the context of discriminant analysis involves the use of eigenvalues derived from the covariance matrices of the data. Here is a step-by-step outline of the formulation:

1. Initial Setup

- Let X be the $N \times (p + q)$ data matrix, where N is the number of observations, p is the number of initial variables, and q is the number of additional variables.
- Partition X into X_p (the initial p variables) and X_q (the additional q variables).

2. Compute Sample Covariance Matrices

- Compute the sample covariance matrix for the initial set of variables S_p and the augmented set S_{p+q}.

3. Discriminant Functions

- Construct the within-group and between-group scatter matrices for both sets of variables:
    - W_p and B_p for the initial set.
    - W_{p+q} and B_{p+q} for the augmented set.

4. Eigenvalue Problem

Solve the eigenvalue problem for the ratio of between-group to within-group scatter matrices:

- For the initial set: $B_p w = \lambda W_p w$

- For the augmented set: $B_{p+q}w = \lambda W_{p+q}w$

5. Rao's U-Statistic

- Let $\lambda_1, \lambda_2, \ldots, \lambda_p$ be the eigenvalues from the initial set.
- Let $\lambda'_1, \lambda'_2, \ldots, \lambda'_{p+q}$ be the eigenvalues from the augmented set.
- Compute Rao's U-Statistic using the following formula:

$$U = N \left(\sum_{i=1}^{p+q} log(1 + \lambda'_i) - \sum_{i=1}^{p} log(1 + \lambda_i) \right)$$

6. Test Statistic

- Under the null hypothesis H0H_0H0, the U-Statistic asymptotically follows a chi-square distribution with $q(p + 1)$ degrees of freedom.

7. Decision Rule

Compare the calculated U-Statistic to the critical value from the chi-square distribution with $q(p + 1)$ degrees of freedom:

If U is greater than the critical value, reject H_0, indicating that the additional variables provide significant additional discrimination.

If U is less than or equal to the critical value, do not reject H_0, indicating that the additional variables do not provide significant additional discrimination.

7.7.4 Summary of Notations

- N: Number of observations
- p: Number of initial variables

- q: Number of additional variables
- X_p Data matrix for the initial p variables
- X_q: Data matrix for the additional q variables
- S_p: Sample covariance matrix for the initial p variables
- S_{p+q}: Sample covariance matrix for the augmented set of $p + q$ variables
- W_p, B_p: Within-group and between-group scatter matrices for the initial set
- W_{p+q}, B_{p+q}: Within-group and between-group scatter matrices for the augmented set
- λ_i: Eigenvalues from the initial set
- λ_i': Eigenvalues from the augmented set

By using this formulation, Rao's U-Statistic helps determine the additional discriminative power provided by a new set of variables in discriminant analysis.

7.8 Summary

In this unit, we have covered the concepts of Discriminant Analysis under following situations:

1. Define classification and discrimination procedure.
2. Discuss sample discriminant analysis.
3. Derive the classification into One of Two Multivariate Normal Population.
4. Derive the classification into One of Two Multivariate Normal Populations when the Parameters are Estimated.
5. Discuss Rao U statistic and its procedure.

7.9 Self-Assessment Exercises

1. What is the probability of misclassification?
2. Define Rao U-statistic and explain its mathematical formulation.

3. If q_1 and q_2 are a priori probabilities of drawing an observation from population Π_1 with density $p_1(x)$ and Π_2 with density $p_2(x)$ respectively, and if the cost of misclassifying an observation from Π_1 as from Π_2 is $C(2|1)$ and an observation from Π_2 as from Π_1 is $C(1|2)$. Consider the regions of classification:

$$R_1: [C(2|1)q_1]p_1(x) \geq [C(1|2)q_2]p_2(x)$$
$$R_2: [C(2|1)q_1]p_1(x) < [C(1|2)q_2]p_2(x)$$

Show that these regions minimize the expected loss.

4. A researcher has enough data available to estimate the density functions $p_1(x)$ and $p_2(x)$ associated with the populations Π_1 & Π_2 , respectively. Suppose $C(2|1) = 5$ units and $C(1|2) = 10$ units. In addition, it is known that about 20% of all objects (for which the measurements X can be recorded) belongs to Π_2. Obtain the classification rule.
5. Consider the two data sets

$$X_1 = \begin{bmatrix} 3 & 2 & 4 \\ 7 & 4 & 7 \end{bmatrix}, \qquad X_2 = \begin{bmatrix} 6 & 5 & 4 \\ 9 & 7 & 8 \end{bmatrix}$$

from two bivariate normal populations with same covariance matrix. Calculate the linear discriminant function.

7.10 References

- Johnson, R. A., Wichern, D. W. (2019): Applied Multivariate Statistical Analysis. United Kingdom: Pearson
- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Giri Narayan, C. (1995): Multivariate Statistical Analysis.
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.
- Williams, E. J. (1955): Significance tests for discriminant functions and linear functional relationships. Biometrika 42, 360—381.

7.11 Further Readings

- Khatri, C. G.: Multivariate Analysis.
- Mardia, K. V.: Multivariate Analysis.
- Seber, G.A.F.: *Multivariate Observations*. Wiley, New York.
- Rencher, Alvin C.: Multivariate Statistical Inference and Applications. John Wiley. New York, New York.

UNIT: 8 ADVANCE ANALYSIS

Structure

8.1 Introduction

The **Stein estimator** is a key result in the field of statistical estimation, particularly in the context of estimating the mean of a multivariate normal distribution. It is a fundamental example of how traditional estimators can be improved by considering both bias and variance, leading to more efficient estimators in high-dimensional settings. It was introduced by Charles Stein in 1956, and later expanded by James and Stein in 1961, demonstrating that the commonly used maximum likelihood estimator (MLE) for the mean of a multivariate normal distribution can be improved upon when considering the mean squared error (MSE) loss function. These estimators are obtained by multiplying the maximum likelihood estimator by properly chosen shrinkage factors and known as the Stein estimators. A refinement of the James-Stein estimator is the **positive part estimator**, where the shrinkage factor is truncated to be non-negative and it provides further improvement in terms of mean squared error.

8.2 Objectives

After going through this unit, you will be able to:

- Inadmissibility of maximum likelihood estimator of mean vector of multivariate normal distribution when dimension is greater than three
- James-Stein estimator of the mean vector
- Improved estimation of dispersion matrix of an MND.

8.3 Inadmissibility of Maximum Likelihood Estimator of Mean Vector of Multivariate Normal Distribution when Dimension is Greater than Three

In classical statistics, the maximum likelihood estimator (MLE) is often considered a natural and efficient estimator for parameters of a distribution. However, in the case of estimating the mean vector of a multivariate normal distribution, Charles Stein proved that the MLE is inadmissible when the dimension of the mean vector is greater than 2. This result, later developed in collaboration with James, forms the basis of the James-Stein estimator, which dominates the MLE in terms of mean squared error (MSE).

8.4 James-Stein Estimation

The James-Stein estimator is a shrinkage estimator used to estimate the mean of a multivariate normal distribution. It's particularly notable in high-dimensional statistics for outperforming the traditional sample mean estimator under certain conditions.

8.4.1 Background

Let's consider a multivariate normal distribution:

$X \sim N_p(\mu, \sigma^2 I_p)$

Where $X = (X_1, X_2, \ldots X_p)$ is a p-dimensional random vector. Obviously, on the basis of a single observation vector X, the MLE of μ is X.

Note: In case we have a set of n observation vectors from $N_p(\mu, \sigma^2 I_p)$, we simply replace X by the sample mean vector and proceed.

8.4.2 James-Stein Estimator

The James-Stein estimator is given by:

$$\hat{\mu}_{JS} = \left[1 - \frac{(p-2)\sigma^2}{||X||^2}\right] X$$

where $\left|\left|X\right|\right|^2$ is the squared Euclidean norm of the sample mean vector and defined as

$$\left|\left|X\right|\right|^2 = \sum_{i=1}^{p} X_i^2$$
$$= X'X$$

The James-Stein estimator "shrinks" the sample mean vector X towards the origin (or some other fixed point if modified). The amount of shrinkage depends on the ratio $\frac{(p-2)\sigma^2}{\left|\left|X\right|\right|^2}$. The larger the $\frac{(p-2)\sigma^2}{\left|\left|X\right|\right|^2}$., the less shrinkage is applied, and conversely, the smaller the $\frac{(p-2)\sigma^2}{\left|\left|X\right|\right|^2}$, the more shrinkage is applied.

The James-Stein estimator works due to a phenomenon known as **Stein's Paradox**, where the combined estimation of several parameters can lead to more accurate results than estimating each parameter separately. For $p \geq 3$, the shrinkage reduces the variance more than it increases the bias, resulting in a lower overall mean squared error compared to the sample mean estimator.

Stein (1956) and James and Stein (1961) showed that, if $p \geq 3$, X is inadmissible and, under the mean squared error $E(\hat{\mu} - \mu)'(\hat{\mu} - \mu)$, it has lower risk than X. For proving this result, we use the following lemma.

8.4.3 Stein's Multivariate Lemma

Let $Z_{(k\times 1)} \sim N(0, I_k)$ and $g: R^k \to R$ be an absolutely continuous and differentiable with derivative

$$\frac{\partial g(z)}{\partial z} = \nabla g(z).$$

Then

$$E[Zg(Z)] = E[\nabla g(z)]. \tag{8.1}$$

Proof: Let Z_i be the i^{th} component of Z. Then, using integration by parts, we obtain

$$E_{Z_i}[Z_i g(Z)] = \int_{-\infty}^{\infty} z_i g(z) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} dz_i$$

$$= g(z)\int z_i \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} dz_i \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{\partial}{\partial z_i}\big(g(z)\big) \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} dz_i$$

$$= -g\,(z) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{\partial}{\partial z_i}\big(g(z)\big) \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z_i^2}$$

$$= 0 + E_{Z_i}\left[\frac{\partial}{\partial Z_i}\big(g(Z)\big)\right].$$

Hence

$E[Zg(Z)] = E[\nabla g(z)]$

Which leads to the required result .

We state the following result, which can be easily proved using the previous Stein's lemma.

Theorem 8.4.1.: When $X \sim N(\mu, I_p)$, we have

$$E[(X-\mu)g(X)] = E[\nabla g(X)]. \tag{8.2}$$

Result 8.4.1.: Under the under the mean squared error $E(\hat{\mu}-\mu)'(\hat{\mu}-\mu)$, the Stein-rule estimator $\hat{\mu}_{JS}$ has lower MSE than the MLE X as long as $k \geq 3$.

Proof: The MSE for $\hat{\mu}$ is given by

$$\mathrm{M}\big(\hat{\mu}_{JS}, \mu\big)$$

$$= E\big(\hat{\mu}_{JS} - \mu\big)'\big(\hat{\mu}_{JS} - \mu\big)$$

$$= E\left[X - \mu - \frac{p-2}{X'X}X\right]'\left[X - \mu - \frac{p-2}{X'X}X\right]$$

$$= \left\{E(X-\mu)'(X-\mu) - 2(p-2)E\left[(X-\mu)'X\frac{1}{X'X}\right] + (p-2)^2 E\left[\frac{1}{(X'X)^2}X'X\right]\right\}.$$

Hence

$$\Delta = E(X-\mu)'(X-\mu) - \mathrm{E}\big(\hat{\mu}_{JS} - \mu\big)'\big(\hat{\mu}_{JS} - \mu\big)$$

$$= \left\{2(p-2)E\left[(X-\mu)'X\frac{1}{X'X}\right] - (p-2)^2 E\left[\frac{1}{(X'X)^2}X'X\right]\right\}$$

$$= 2(p-2)E\left[(X-\mu)'X\frac{1}{X'X}\right] - (p-2)^2 E\left[\frac{1}{X'X}\right]$$

$$= 2(p-2)E\left[\frac{\partial}{\partial X'}\frac{1}{(X'X)^2}X\right] - (p-2)^2 E\left[\frac{1}{X'X}\right]$$

$$= (p-2)E\left[2p\frac{1}{X'X} - 4\frac{X'X}{(X'X)^2} - (p-2)\frac{1}{X'X}\right]$$

$$= (p-2)^2 \mathrm{E}\left(\frac{1}{X'X}\right) \geq 0.$$

Therefore $\Delta > 0$, whenever $p \geq 3$. This leads to condition.

Since X is the minimax estimator for μ, (8.2) is the condition for $\hat{\mu}_{JS}$ to be a minimax estimator.

The above result also shows that the maximum likelihood estimator X is inadmissible under the mean squared error criterion and uniformly dominated by the JS estimator.

8.4.4 Key Properties

1. Dominance: The James-Stein estimator dominates the sample mean estimator in terms of lower risk (mean squared error) for any μ when $p \geq 3$.

2. Shrinkage: It shrinks estimates towards the origin, effectively reducing the variance of the estimator.

3. Bias-Variance Trade-off: The estimator introduces some bias, but the reduction in variance typically outweighs this, leading to a lower total mean squared error.

8.4.5 Practical Considerations

The original James-Stein estimator assumes σ^2 is known. In practice, σ^2 might need to be estimated from the data, which leads to a modified version of the estimator.

The choice of shrinkage target (origin in the classical case) can be generalized to other values, depending on prior knowledge or other considerations.

The James-Stein estimator illustrates a counterintuitive result in statistics: sometimes, by pooling information across different components and allowing for some bias, one can achieve better overall estimation accuracy.

8.4.6 Applications

The James-Stein (JS) estimator has a range of applications across different fields, particularly where high-dimensional data and estimation problems are common. The main idea behind using the JS estimator is to improve the estimation accuracy by shrinking the estimates toward a central point, thus reducing variance and improving mean squared error (MSE). Here are some key applications of the JS estimator:

1. Genomics and Bioinformatics

Gene Expression Analysis: In genomics, researchers often estimate the expression levels of thousands of genes simultaneously. Since the number of genes (parameters) is large compared to the number of samples (observations), the sample mean can be a poor estimator. The JS estimator can be used to improve the accuracy of these estimates by shrinking the estimated expression levels toward a central value, thereby reducing variance.

DNA Microarray Data: Similar to gene expression analysis, microarray data analysis involves estimating a large number of parameters (gene expressions). The JS estimator helps in obtaining more accurate estimates of gene activity levels by borrowing strength across different genes.

2. Econometrics and Finance

Portfolio Optimization: In finance, estimating the expected returns of various assets is crucial for portfolio optimization. When the number of assets (stocks, bonds, etc.) is large relative to the available data points, the sample mean of returns can be unreliable. The JS estimator can provide more stable estimates by shrinking the returns toward the mean return of all assets, which can lead to more robust portfolio choices.

Risk Management: The JS estimator can be applied to estimate risk metrics, such as Value at Risk (VaR) and expected shortfall, more accurately when dealing with high-dimensional risk factor models.

3. Machine Learning and Data Mining

Regularization in High-Dimensional Models: The concept of shrinkage in the JS estimator is related to regularization techniques (like Lasso and Ridge regression) used in machine learning. In situations where models have a large number of features compared to observations, shrinkage estimators can help prevent overfitting and improve predictive performance.

Ensemble Learning: In ensemble methods, the JS estimator can be used to combine different models by shrinking their predictions towards a common central model, thus reducing variance and improving overall predictive accuracy.

4. Sports and Epidemiology

Player Performance Metrics: In sports analytics, the performance of players (like batting averages in baseball or shooting percentages in basketball) often needs to be estimated based on a limited number of observations. The JS estimator can improve these estimates by shrinking individual performance metrics towards the league average, leading to more reliable performance assessments.

Disease Rate Estimation: In epidemiology, estimating disease rates in small populations (e.g., rare diseases in small towns) can suffer from high variance. The JS estimator can provide more reliable estimates by shrinking the rates toward the overall mean rate, which helps in understanding disease prevalence more accurately.

5. Psychometrics and Social Sciences

Test Scores and Ability Estimation: In psychometrics, when estimating abilities or traits based on test scores, especially in small sample settings, the JS estimator can provide more

accurate estimates by shrinking individual scores toward the group mean. This can help in obtaining more stable assessments of abilities or personality traits.

Survey Data Analysis: In the analysis of survey data, where multiple related outcomes or questions are analyzed simultaneously, the JS estimator can be used to improve the estimation of mean responses by shrinking towards a common overall mean, thus reducing estimation error.

6. Image Processing and Computer Vision

Denoising: In image processing, estimating pixel intensities in noisy images can benefit from shrinkage techniques like the JS estimator. It can reduce noise by shrinking pixel intensity estimates towards a global mean or a structured model, improving the clarity of the image.

Reconstruction: In computer vision tasks like 3D reconstruction, where estimates of spatial coordinates are made from multiple noisy observations, the JS estimator can reduce the variance of the estimated coordinates, leading to more accurate reconstructions.

7. Astronomy and Astrophysics

Star Luminosity Estimation: When estimating the brightness of stars or other celestial bodies from telescope data, especially when dealing with high-dimensional data sets (e.g., spectra), the JS estimator can provide more reliable estimates by shrinking the luminosity estimates towards a central value.

8.4.7 Drawbacks

The James-Stein estimator is widely applicable in fields dealing with high-dimensional data and estimation problems where the number of parameters is large relative to the number of observations. It improves estimation accuracy by introducing a bias (shrinking toward a central point) that is more than compensated by a reduction in variance, leading to lower overall mean

squared error. This makes it valuable in many scientific, engineering, and social science applications where robust estimation is crucial.

While the James-Stein (JS) estimator offers significant advantages in terms of reducing the mean squared error (MSE) compared to the standard sample mean estimator, especially in high-dimensional settings, it also has some drawbacks and limitations. Here are some key drawbacks of the JS estimator:

1. **Assumption of Known Variance:** The standard form of the JS estimator assumes that the variance $\sigma^2$of the underlying normal distribution is known. In practice, σ^2 is often unknown and must be estimated from the data. Estimating σ^2 introduces additional variability, which can affect the performance of the JS estimator. The estimator needs to be adjusted when σ^2 is unknown, potentially reducing its effectiveness.

2. **Shrinkage Toward the Origin:** The JS estimator shrinks the estimates towards a common point, typically the origin (zero vector) or the overall mean. This is based on the assumption that the true means are closer to the origin or that there is some prior belief about the central tendency of the parameters. However, in cases where the true mean vector is far from the origin, this shrinkage can introduce significant bias, leading to poor estimates.

3. **Lack of Interpretability:** The amount of shrinkage applied by the JS estimator is determined by the ratio of the number of parameters and the squared Euclidean norm of the data vector. This can make the estimator less interpretable because the shrinkage factor is data-dependent and may change significantly with small changes in the data. It can be difficult to explain why the estimator behaves a certain way for particular datasets.

4. **Only Applicable for $p \geq 3$:** The JS estimator is only advantageous for $(p \geq 3)$, meaning it's only useful when estimating at least three parameters. For $(p = 1)$or $(p = 2)$, the sample mean is actually the optimal estimator in terms of MSE. Thus, the JS estimator is not applicable in low-dimensional settings.

5. **Bias Introduction:** Although the JS estimator reduces the variance of the estimates, it introduces bias. The trade-off between bias and variance may not always result in a lower mean squared error for every dataset or application. In cases where unbiasedness is a

critical requirement (e.g., in certain inferential statistical tasks), the bias introduced by the JS estimator can be seen as a drawback.

6. **Sensitivity to Outliers:** The JS estimator's shrinkage mechanism can make it sensitive to outliers. If the data vector X contains extreme values, the squared norm $||X||^2$ can become disproportionately large, reducing the shrinkage effect. Consequently, the estimator may perform poorly in the presence of outliers, which can skew the results and reduce the benefits of shrinkage.

7. **Less Effective in Small Samples:** The JS estimator assumes that shrinkage will improve estimation accuracy by reducing variance. However, in small sample sizes, the effectiveness of shrinkage may be limited because the estimator relies on having enough data to accurately determine how much to shrink. In cases with very few observations, the JS estimator may not provide substantial gains over the sample mean, or its performance could even deteriorate.

8. **No Clear Choice of Shrinkage Target:** The standard JS estimator shrinks towards the origin, but this choice is somewhat arbitrary and may not always align with the underlying structure of the data. Other shrinkage targets (e.g., the grand mean or some other central point) could be more appropriate depending on the context. However, selecting an appropriate target is not straightforward and often requires domain knowledge or subjective judgment.

9. **Computational Considerations:** While the JS estimator itself is not computationally expensive, in practice, its performance needs to be evaluated against other estimators or methods. This involves additional computation, especially when dealing with high-dimensional data or when selecting the optimal shrinkage factor. As a result, the implementation and validation of the JS estimator may require more computational resources and expertise compared to simpler estimators.

10. **Limited Applicability in Non-Gaussian Settings:** The JS estimator is specifically designed for multivariate normal distributions. Its effectiveness relies on the assumption of normally distributed data with a common variance structure. If the data do not follow a

multivariate normal distribution (e.g., heavy-tailed distributions, skewed distributions), the performance of the JS estimator may degrade, and other estimators or robust techniques might be more appropriate.

8.4.8 Some Notable Alternatives to the JS Estimator

1. Empirical Bayes Estimators: Empirical Bayes (EB) methods estimate the shrinkage parameter (or hyperparameters) from the data itself rather than using a fixed formula. This makes EB methods flexible and well-suited to situations where prior information or a reasonable prior distribution can be estimated.

(i) Stein's Estimator with Empirical Bayes: This method adapts the shrinkage factor by estimating it from the data, which can improve performance, especially when the underlying variance (σ^2) is unknown or needs to be estimated.

(ii) Generalized Empirical Bayes Estimators: These extend the EB framework to different types of data and priors, providing flexibility and robustness in various settings.

2. Ridge Regression: Ridge regression is a regularization method used primarily in linear regression but also for estimating parameters in multivariate settings. It penalizes large coefficients by adding a shrinkage penalty to the least square's estimation.

(i) Application: Ridge regression is useful when multicollinearity exists among predictor variables or when the number of predictors is large relative to the number of observations.

3. Lasso (Least Absolute Shrinkage and Selection Operator): Lasso is another shrinkage and selection method that, unlike ridge regression, can produce sparse models by setting some coefficients exactly to zero. This is particularly useful when variable selection is desired along with parameter estimation.

(i) Application: Lasso is particularly useful in high-dimensional settings where it helps in both reducing variance and performing variable selection, thus providing a more interpretable model.

4. Bayesian Shrinkage Estimators: Bayesian shrinkage estimators provide a principled framework for incorporating prior information into the estimation process. The posterior distribution is used to derive estimates, often leading to shrinkage towards the prior mean.

(i) Bayesian Linear Models: For a normal prior on the regression coefficients, the posterior mean can provide shrinkage similar to ridge regression but with more flexibility to include other types of priors and hierarchical models.

(ii) Hierarchical Bayesian Models: These models allow for varying degrees of shrinkage depending on the grouping of data, which can be particularly useful in cases with nested or hierarchical structures.

5. Tikhonov Regularization: Tikhonov regularization (a generalization of ridge regression) applies to ill-posed problems, particularly in inverse problems and other contexts where direct estimation is not feasible.

(i) Application: Useful in image reconstruction, signal processing, and other fields where the model is often under-determined, and regularization is needed to stabilize the solution.

6. Principal Component Analysis (PCA) and Factor Analysis: While not shrinkage estimators per se, PCA and factor analysis reduce dimensionality by identifying the main directions (principal components) that capture the most variance in the data. By focusing on the principal components, one effectively performs shrinkage in the direction of these components, reducing the impact of less informative directions.

(i) Application: Commonly used in data preprocessing for high-dimensional data, particularly when noise reduction is needed.

7. Shrinkage Towards a General Target: Shrinkage towards a general target allows for the shrinkage of estimates towards a target other than the origin or the grand mean, which might be more appropriate in specific applications.

(i) Application: Used when there is prior knowledge about the true mean being close to some known vector, not necessarily the origin.

8. Adaptive Shrinkage Estimators: Adaptive shrinkage estimators adjust the amount of shrinkage based on the data, often using methods like cross-validation to find the optimal shrinkage parameter.

(i) Application: Useful in non-parametric settings and when a data-driven approach is needed to determine the appropriate amount of shrinkage dynamically.

9. Stein-type Shrinkage Estimators: Stein-type shrinkage estimators are a broad class of estimators inspired by the original James-Stein estimator, but they adapt the shrinkage based on different loss functions or constraints.

(i) Application: These are useful when the assumptions of the JS estimator are not fully met or when the loss function differs from the quadratic loss.

10. Nonparametric and Semi-parametric Methods: In some cases, nonparametric methods (e.g., kernel density estimation) and semi-parametric methods (e.g., partially linear models) may provide better estimates than parametric shrinkage estimators when the underlying distribution does not meet the normality assumption or is highly skewed.

(i) Application: Used in settings where the form of the underlying distribution is unknown or cannot be easily specified.

Estimating the variance-covariance matrix of a multivariate distribution is a fundamental task in statistics and data analysis, especially when dealing with multivariate normal data.

Stein estimation for the variance-covariance matrix is a powerful method that extends the principles of the James-Stein estimator to the covariance matrix, offering more stable and accurate estimates, especially in high-dimensional settings. It is designed to improve the estimation accuracy by shrinking the sample covariance matrix towards a structured target. By shrinking the sample covariance matrix towards a structured target, these estimators reduce variance, ensure positive definiteness, and improve overall estimation accuracy.

8.5 Improved Estimation of Dispersion Matrix of an MND

8.5.1 Background and Motivation

The sample covariance matrix $\hat{\Sigma}$ is the most common estimator for the covariance matrix Σ of a multivariate normal distribution. However, in high-dimensional settings (where the number of variables p is large relative to the number of observations n), $\hat{\Sigma}$ becomes unstable, can be singular, and typically has high variance. Stein-type estimators aim to address these issues by

shrinking $\hat{\Sigma}$ towards a more stable target, similar to how the James-Stein estimator shrinks mean vectors.

8.5.2 Challenges with the Sample Covariance Matrix

While the sample covariance matrix $\hat{\Sigma}$ is an unbiased estimator of the true covariance matrix Σ, it has several drawbacks, especially when the number of variables p is large relative to the number of observations n:

- **High Variance**: The sample covariance matrix can have high variance, leading to noisy estimates.
- **Instability**: When p approaches or exceeds n, $\hat{\Sigma}$ becomes ill-conditioned or singular, making it difficult to invert or use in further analyses.
- **Overfitting**: In high dimensions, $\hat{\Sigma}$ can overfit the sample data, capturing noise rather than the true underlying structure.

8.5.3 Stein's Loss Function

One of the key components in Stein estimation is the use of Stein's loss function, which measures the difference between the true covariance matrix Σ and an estimator $\hat{\Sigma}$. The most common form of the Stein's Loss (Risk) Function is

$$L(\hat{\Sigma}, \Sigma) = tr(\hat{\Sigma}^{-1}\Sigma) - logdet(\hat{\Sigma}^{-1}\Sigma) - p$$

This loss function has some desirable properties:

- It is invariant under linear transformations of the data.
- It penalizes both underestimation and overestimation of the eigenvalues of Σ.
- It encourages shrinkage of the estimator towards a more structured form, reducing the risk associated with the high variability of the sample covariance matrix.

8.5.4 James-Stein-Type Shrinkage Estimators for Covariance Matrices

The James-Stein-type shrinkage estimator for the covariance matrix takes a form similar to the James-Stein estimator for the mean vector, shrinking the sample covariance matrix towards a target matrix T. The general form is:

$\hat{\Sigma}_{Stein} = (1 - \lambda)\hat{\Sigma} + \lambda T$

where:

$\hat{\Sigma}$ is the sample covariance matrix.

T is the target covariance matrix (often chosen as the identity matrix, I, or a diagonal matrix).

λ is the shrinkage parameter, typically estimated from the data.

8.5.5 Choosing the Shrinkage Target T

The target matrix T is chosen based on prior knowledge or assumptions about the structure of the covariance matrix:

- **Identity Matrix (I)**: Assumes that variables are uncorrelated and have unit variance. This is a common choice in the absence of strong prior information.
- **Diagonal Matrix**: Assumes that variables are uncorrelated, but allows for different variances for each variable.
- **Scaled Identity Matrix**: Shrinks towards an identity matrix scaled by the average variance.
- **Average of the Sample Covariance Matrices**: In a hierarchical or multi-group setting, this could be the average of the sample covariance matrices across groups.

8.5.6 Estimating the Shrinkage Parameter λ

The optimal shrinkage parameter λ is typically estimated to minimize the mean squared error or Stein's loss. There are various methods for estimating λ including:

- **Ledoit-Wolf Estimator**: A popular method that provides an analytic solution for the optimal shrinkage parameter λ by minimizing the Frobenius norm or a similar loss function.
- **Cross-Validation**: A data-driven approach where λ is chosen to minimize an empirical risk measure (e.g., cross-validated Stein’s loss).

8.5.7 Properties and Advantages

- **Improved Estimation in High Dimensions**: The shrinkage reduces the variance of the estimator, making it more stable in high-dimensional settings.
- **Guaranteed Positive-Definiteness**: Shrinkage towards a positive definite target ensures that the resulting covariance matrix is positive definite, which is crucial for many applications (e.g., portfolio optimization, multivariate analysis).
- **Reduced Sensitivity to Sampling Variability**: By shrinking the covariance matrix, the estimator becomes less sensitive to sampling noise and outliers.

8.5.8 Applications

- **Finance**: Covariance matrix estimation is crucial in portfolio optimization, where accurate estimates of the covariance matrix of asset returns are necessary to construct efficient portfolios. Stein shrinkage estimators are widely used to improve the stability of these estimates.
- **Genomics**: In gene expression studies, covariance matrices are used to understand the relationships between genes. High-dimensional data (many genes, few samples) make shrinkage techniques particularly valuable.
- **Multivariate Statistics**: In various multivariate techniques like Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA), accurate estimation of the covariance matrix is crucial, and shrinkage estimators provide more reliable estimates.

8.6 Summary

In this unit, we have covered the concepts of Linear Regression Model under following situations:

- We have discussed Inadmissibility of maximum likelihood estimator of mean vector of multivariate normal distribution when dimension is greater than three.
- We have derived James-Stein estimator of the mean vector.
- We have explained Improved estimation of dispersion matrix of an MND.

8.7 Self-Assessment Exercises

1. Show that $E[Zg(Z)] = E[\nabla g(z)]$ where $Z_{(k\times 1)} \sim N(0, I_k)$ and $g: R^k \to R$ be an absolutely continuous and differentiable with derivative $\frac{\partial g(z)}{\partial z} = \nabla g(z)$.

2. Under the under the mean squared error $E(\hat{\mu} - \mu)'(\hat{\mu} - \mu)$, prove that the Stein-rule estimator $\hat{\mu}_{JS}$ has lower MSE than the MLE X as long as $k \geq 3$.

3. Write down the application and Drawbacks of James-Stein (JS) estimator.
4. Explain some notable alternatives to the JS estimator.
5. Discuss the estimation of the Shrinkage Parameter λ in Shrinkage Estimators for Covariance Matrix.

8.8 References

- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.
- Everitt B.S. & Dunn G. (1991): Applied Multivariate Data Analysis. Edward Arnold. London. pp. 219-220.
- Giri Narayan C. (1995): Multivariate Statistical Analysis
- Härdle WK, Simar L (2015): Applied Multivariate Statistical Analysis. Springer-Verlag, Berlin.4^{th} Berkeley Symp. Math. Statist. Prob.
- James, W. and Stein, C.M. (1961): Estimation with quadratic loss function, Proceedings of the fourth Berkeley Symposium on Math. Statist. and Probability, vol. 1 (University of California Press, Berkeley)
- Johnson, R. A., Wichern, D. W. (2019): Applied Multivariate Statistical Analysis. United Kingdom: Pearson.
- Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York.

- Mardia, K. V., Bibby, J. M., Kent, J. T. (1979): Multivariate Analysis. United Kingdom: Academic Press.
- Muirhead, R. J. (2009): Aspects of Multivariate Statistical Theory. Germany: Wiley.
- Stein, C. M. (1956): Inadmissibility of the Usual Estimator for the Mean of a Multivariate Normal Distribution, Proceedings of the Third Berkeley Symposium on Math. Statist. and Probability, vol. 1 (University of California Press, Berkeley).
- Stein, C. M. (1966): An approach to the recovery of inter-block information in balanced incomplete block designs, In Festschrift for J. Neyman: Research papers in Statistics (F. N. David, ed.), Wiley, New York.

8.9 Further Readings

- Brenner, D., Bilodeau, M.: Theory of Multivariate Statistics. Germany: Springer.
- Kotz, S., Balakrishnan, N. and Johnson, N.L.: Continuous Multivariate Distribution Models and Applications (Second Edition). Volume 1, Wiley - Inter science, New York.
- Khatri, C. G.: Multivariate Analysis.
- Mardia, K. V.: Multivariate Analysis.
- Seber, G.A.F.: *Multivariate Observations*. Wiley, New York.
- Rencher, Alvin C.: Multivariate Statistical Inference and Applications. John Wiley. New

UNIT: 9 PRINCIPAL COMPONENT ANALYSIS

Structure

9.1 Introduction

Analysis of various fields of sciences involve high-dimensional data sets. We adopt possible projection methods to project it to a lower-dimensional subspace without losing important information regarding some characteristic of variables. One way is by creating a reduced set of linear or nonlinear transformations of the input variables. Linear methods such as principal component analysis (PCA) (Hotelling, 1933) and canonical variate and correlation analysis (CVA or CCA) (Hotelling, 1936), are two of the most popular dimensionality-reducing techniques. Both PCA and CVA are eigenvalue-eigenvector problems.

Principal Component Analysis is a technique for deriving a reduced set of orthogonal linear projections of a single collection of correlated variables where projections are ordered by decreasing variance. The variance is an important measure of amount of information in that variable.

Canonical correlation is a technique to identify and quantify the association between two sets of variables. Each set can contain several variables. Simple and multiple correlations are special cases of canonical correlation in which one or both sets contain a single variable. This technique was given by H. Hotelling in 1935-36, for relating the arithmetic speed and arithmetic power to reading speed and reading power based on a sample data received from 140 seventh grade students. Other examples where canonical correlations can be helpful are: relating governmental policy variables with economic goal variables; relating college performance variables (grades in courses in different subjects) with pre-college achievement variables (percentage of marks in high school, number of extracurricular activities in height school, etc.); relating yield attributing parameters (test weight, plant height, number of grains per panicle, etc.) and quality parameters (protein content, carbohydrate content, etc.) in case of a certain crop; relating job satisfaction variables (supervisor satisfaction, workload satisfaction, general satisfaction, etc.) and job characteristic variables (feedback, task identity, task variety, etc.); relating physiological variables (weight in kg, waist in inches, pulse rate, etc.) with exercise variables (number of sit ups, jumps, etc.) and many others such pairs. Canonical correlation analysis focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in the second set. The idea is first to determine the pair of linear combinations having the largest correlation. Next, we determine the pair of linear combinations having the largest correlation among all pairs

uncorrelated with the initially selected pairs. This process continues until the number of pairs of canonical variables equals the number of variables in the smaller group. The pairs of linear combinations are called the canonical variables and their correlations are called canonical correlations. The canonical correlations measure the strength of association between the two sets of variables. The maximization aspect of the technique represents an attempt to concentrate a high-dimensional relationship between two sets of variables into a few pair of canonical variables. The purpose of canonical correlation is to explain the relation of the two sets of variables, not to model the individual variables.

9.2 Objectives

After studying this unit, one will be able to

• define principal components

• derive principal components from covariance matrix

• derive principal components from a correlation matrix

• understand the meaning and concept of canonical correlation

9.3 Principal Component Analysis

Principal component are linear combinations of random and statistical variables which have special properties in terms of variances. For examples, the first principal component is the normalized linear combination (the sum of square of the coefficient being one) with maximum variance. The principal components turn out to be the characteristic vectors of the covariance matrix. Thus, the study of principal component can be considered as the usual developments of characteristic roots and vectors (for positive semi-definite matrices). In statistical practice the method of principal component is used to find the linear combination with large variance. In many exploratory studies where the numbers of variables are too large, a way of reducing the number of variables to be treated is to discard the linear combinations which have small variances and study only those with large variances.

9.3.1 Uses of Principle Component Analysis

PCA has also been referred to as a method for decorrelating the observation matrix**.** PCA can be used for lossy data compression (compress data to be transmitted without losing much information), pattern recognition, and image analysis.

The first few principal component scores reveal whether most of the data live on a linear subspace of R^r and can be used to identify outliers, distributional peculiarities, and clusters of points. The last few principal component scores can be used to detect collinearity.

9.4 Derivation

Theorem 9.4.1: Let X be a $p \times 1$ random vector with mean vector $E(X) = 0$ and covariance matrix $E(XX') = \Sigma$.. Then there exists a linear transformation $U = \beta' X$, such that the covariance matrix of U is $E(UU') = \Lambda = diag(\lambda_1, \dots, \lambda_p)$, with $\lambda_1 \geq \cdots \geq \lambda_p$, where $\lambda_1, \dots, \lambda_p$ are roots of $|\Sigma - \lambda I| = 0$. The r^{th} column of B, say $\beta^{(r)}$ satisfies $(\Sigma - \lambda_r I)\beta^{(r)} = 0$. The r^{th} component of U is $u_r = \beta^{(r)'} X$.

Proof: Let β be a $p \times 1$ normalized vector ($\beta'\beta = 1$), then variance of $\beta' X$ is

$V(\beta' X) = E(\beta' X)^2$

$= E(\beta' XX' \beta)$

$= \beta' E(XX')\beta = \beta' \Sigma \beta$

To determine normalized linear combination with maximum variance, we find β, such that $\beta'\beta = 1$, which maximizes $\beta' \Sigma \beta$. Let

$\varphi_1 = \beta' \Sigma \beta - \lambda(\beta'\beta - 1)$

$$= \sum_{i=1}^{p} \sum_{j=1}^{p} \beta_i \sigma_{ij} \beta_j - \lambda \left(\sum_{i=1}^{p} \beta_i^2 - 1 \right)$$

λ is the Lagrange's multiplier.

Then, for maximizing $\beta'\Sigma\beta$ subject to $\beta'\beta = 1$, the vector of partial derivatives is

$$\frac{\partial}{\partial\beta}\varphi_1 = 0$$

$$\Rightarrow 2\Sigma\beta - 2\lambda\beta = 0 \tag{9.1}$$

$$\Rightarrow \Sigma\beta - \lambda\beta = 0$$

$$\Rightarrow (\Sigma - \lambda I)\beta = 0 \tag{9.2}$$

To have solution of (9.2) *for* $\beta'\beta = 1$, λ *must satisfy*

$$|\Sigma - \lambda I| = 0 \tag{9.3}$$

$|\Sigma - \lambda I|$ is a polynomial of degree p in λ and thus has p roots. Let these roots be $\lambda_1 \geq \cdots \geq \lambda_p$. Since Σ is positive definite, all these roots are real and positive. If we pre multiply (9.4.2) with β', we get

$$\beta'\Sigma\beta - \lambda\beta'\beta = 0$$

$$\Rightarrow \beta'\Sigma\beta = \lambda \tag{9.4}$$

Thus if β satisfies (9.2), then variance of $\beta'X$ is λ. Thus, for maximum variance, we should use the largest eigenvalue λ_1. Let $\beta^{(1)}$ be the corresponding normalized solution (normalized eigen vector) of $(\Sigma - \lambda_1 I)\beta = 0$. Then $u_1 = \beta^{(1)\prime}X$ is a normalized linear combination (first principal component) with maximum variance. i.e.

$$V(u_1) = \beta^{(1)'}\Sigma\beta^{(1)} = \lambda$$

(If $\Sigma - \lambda I$ is of rank $(p-1)$, then there is only one solution to $(\Sigma - \lambda I)\beta = 0$ and $\beta'\beta = 1$).

Let us obtain linear combination $\beta'X$ which has maximum variance among all linear combinations uncorrelated with u_1. Then

$$0 = Cov(\beta'X, u_1) = E\{\beta'X - E(\beta'X)\}\left\{\beta^{(1)'}X - E\left(\beta^{(1)\prime}X\right)\right\}' = E\left(\beta'XX'\beta^{(1)}\right) = \beta'\Sigma\beta^{(1)}$$

Since $\Sigma\,\beta^{(1)} = \lambda\,\beta^{(1)}$, we have

$$\beta'\Sigma\beta^{(1)} = \lambda_1\beta'\beta^{(1)} = 0 \tag{9.5}$$

Which implies that β is orthogonal to $\beta^{(1)}$, i.e., $\beta'\beta^{(1)} = 0$ since $\lambda_1 \neq 0$.

Consider

$$\varphi_2 = \beta'\Sigma\beta - \lambda(\beta'\beta - 1) - 2\gamma_1\beta'\Sigma\beta^{(1)}$$

where λ and γ_1 are Lagrange multipliers. The vector of partial derivatives is

$$\frac{\partial}{\partial\beta}\varphi_2 = 0$$

$$\Rightarrow \quad 2\Sigma\beta - 2\lambda\beta - 2\gamma_1\Sigma\beta^{(1)} = 0$$

Pre multiplying by $\beta^{(1)'}$, we obtain

$$2\beta^{(1)'}\Sigma\beta - 2\lambda\beta^{(1)'}\beta - 2\gamma_1\beta^{(1)'}\Sigma\beta^{(1)} = 0$$

Since $\beta^{(1)'}\Sigma\beta = \lambda_1\beta^{(1)'}\beta = 0$, so that $-\gamma_1\beta^{(1)'}\Sigma\beta^{(1)} = 0$ or $\gamma_1\lambda_1 = 0$, because $\beta^{(1)'}\Sigma\beta^{(1)} = \lambda_1$. This implies that $\gamma_1 = 0$, because $\lambda_1 \neq 0$.

This means that the condition of uncorrelated ness is itself satisfied when β satisfies (9.2) and λ satisfies (9.3).

Let $\lambda_{(2)}$ be the maximum of $\lambda_1, \ldots, \lambda_p$ such that β is a vector satisfying

$$(\Sigma - \lambda_{(2)}I)\beta = 0, \beta'\beta = 1,$$

and (9.5).

Let this vector be $\beta^{(2)}$ and corresponding linear combination is $u_2 = \beta^{(2)'}X$. This will be shown later that $\lambda_{(2)} = \lambda_2$. Continuing this way, at the $(r+1)^{th}$ step, we find a vector β such that $\beta'X$ has maximum variance of all normalized linear combinations which are uncorrelated with $u_1, \ldots, u_r$, i.e.,

$0 = Cov\,(\beta' X, u_i)$

$= E(\beta' X u_i)$

$= E\left(\beta' X X' \beta^{(i)}\right)$

$= \beta' \Sigma \beta^{(i)} = \lambda_{(i)} \beta' \beta^{(i)}, i = 1,2, \ldots, r$ (9.6)

We want to maximize

$$\varphi_{r+1} = \beta \Sigma \beta - \lambda(\beta' \beta - 1) - 2 \sum_{i=1}^{r} \gamma_i \beta' \Sigma \beta^{(i)} \quad (9.7)$$

Then

$$\frac{\partial}{\partial \beta} \varphi_{r+1} = 2\Sigma\beta - 2\lambda\beta - 2 \sum_{i=1}^{r} \gamma_i \Sigma \beta^{(i)} = 0 \quad (9.8)$$

Pre multiplying (9.8) by $\beta^{(j)'}$ $(j = 1, \ldots, r)$, we obtain

$2\beta^{(j)'} \Sigma\beta - 2\lambda\beta^{(j)'}\beta - 2\gamma_j \beta^{(j)'} \Sigma \beta^{(j)} = 0$

Which leads to $\gamma_j = 0$ $(j = 1, \ldots, r)$. $\Sigma\beta^{(j)} = \lambda_{(j)}\beta^{(j)} = 0$ and the j^{th} term in the sum is (9.8) vanishes. Thus β must satisfy (9.2) and (9.3). Let $\lambda_{(r+1)}$ be the maximum of $\lambda_1, \ldots, \lambda_p$ such that β is a vector satisfying $\left(\Sigma - \lambda_{(r+1)} I\right)\beta = 0, \beta'\beta = 1,$ and (9.6), call it $\beta^{(r+1)}$ and the corresponding linear combination as $u_{(r+1)} = \beta^{(r+1)'} X$. If $\lambda_{(r+1)} = 0$ and $\lambda_{(j)} = 0$, then $\beta^{(j)'} \Sigma \beta^{(r+1)} = 0$ does not imply $\beta^{(j)'} \beta^{(r+1)} = 0$. However, $\beta^{(r+1)}$ can be replaced by a linear combination of $\beta^{(r+1)}$ and $\beta^{(j)}$'s with λ_j's being 0 so that new $\beta^{(r+1)}$ is orthogonal to all $\beta^{(j)}; j = 1,2, \ldots r$. The procedure is continued until at the $(m+1)^{th}$ stage one cannot find a vector β satisfying $\beta'\beta = 1$, (9.2) and (9.6). $\beta^{(1)}, \ldots, \beta^{(m)}$ must be linearly independent. Then either $m < p$ or $m = p$. Since $\beta^{(1)}, \ldots, \beta^{(m)}$ must be linearly independent. Later we will show that $m = p$.

Let $B = \left(\beta^{(1)}, \ldots \beta^{(p)}\right)$ and $\Lambda = diag\left(\lambda_{(1)}, \ldots \lambda_{(p)}\right)$.

In matrix notations, the equation $\Sigma\beta^{(r)} = \lambda_{(r)}\beta^{(r)}$ can be written as

$\Sigma B = B\Lambda$

Further $B'B = I$. Hence $B'\Sigma B = B'B\Lambda = \Lambda$. Further

$$|\Sigma - \lambda I| = |B'||\Sigma - \lambda I||B| = |B'\Sigma B - \lambda B'B| = |\Lambda - \lambda I| = \prod_{i=1}^{p} (\lambda_{(i)} - \lambda) \tag{9.9}$$

Hence roots of (9.9) are diagonal elements of Λ. Thus $\lambda_{(i)} = \lambda_i$ for all i.

The vector U is defined as the vector of principal components of X.

Now we show that $m = p$. If $m < p$, there exist $p - m$ vectors e_j $(j = m + 1, \dots, p)$ such that $\beta^{(i)'} e_j = 0 \; \forall j$, $e_i' e_j = 1$ if $i = j$ and 0 if $i \neq j$. Let $E = (e_{m+1}, \dots e_p)$. Then there exist a $(p - m)$ component vector c such that $Ec = \sum_{i=1}^{(p-m)} c_i e_i$ is a solution to $(\Sigma - \lambda I)\beta = 0$ with $\lambda = \theta$. Consider a root of $|E'\Sigma E - \theta I| = 0$ and a corresponding vector c satisfying $E'\Sigma E = \theta c$. The vector ΣEc is orthogonal to $\beta^{(1)}, \dots, \beta^{(m)}$, as $\beta^{(i)'} \Sigma Ec = \lambda_{(i)} \sum_j^{p-m} c_j \beta^{(i)'} e_j = 0$. Hence $\beta^{(i)}$ is a vector spanned by $e_{m+1}, \dots e_p$ and can be written as Eg, where g is a $(p - m)$ component vector. Multiplying $\Sigma Ec = E'Eg$ by E' we obtain $E'\Sigma Ec = E'Eg = g$. Thus $g = \theta c$ and we have $\Sigma Ec = \theta Ec$. Thus $(Ec)'X$ is uncorrelated with $\beta^{(j)'} X; j = 1, \dots, m$ and this leads to a new $\beta^{(m+1)}$. This contradicts the assumption that $m < p$ and hence we must have $m = p$.

Notes: 1. Contribution of first principal component is

$$\frac{\lambda_1}{\sum_{i=1}^{p} \lambda_i}$$

Contribution of first and second principal component is

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{p} \lambda_i}$$

Corollary: Suppose $\lambda_{(r+1)} = \lambda_{(r+2)} = \cdots = \lambda_{(r+m)} = \upsilon$ (i.e. υ is a root of multiplicity m); then $|\Sigma - \upsilon I|$ is of rank $(p - m)$. Further $\mathcal{B}^* = (\beta^{(r+1)}, \dots \beta^{(r+m)})$ is uniquely determined except for multiplication on the right by an orthogonal matrix.

Proof: From the derivation of theorem, we have $|\Sigma - \upsilon I|\beta^{(i)} = 0, i = r + 1, r + 2, \dots, r + m$, $i.e.$ $\beta^{(r+1)}, \beta^{(r+2)}, \dots, \beta^{(r+m)}$, are linearly independent solutions of $|\Sigma - \lambda I|\beta = 0$. To show there cannot be another linearly independent solution, take $\sum_{i=1}^{p} x_i \beta^{(i)}$, where x_i are scalars. If it is a solution, we have

$\upsilon \sum_{i=1}^{p} x_i \beta^{(i)} = \Sigma \left(\sum_{i=1}^{p} x_i \beta^{(i)} \right) = \sum_{i=1}^{p} x_i \Sigma \, \beta^{(i)} = \sum_{i=1}^{p} x_i \lambda_i \beta^{(i)}$

Since $\upsilon x_i = \lambda_i x_i$, we must have $x_i = 0$ unless $i = r+1, r+2, \ldots r+m$. Thus, the rank is $(p-m)$. If B^* is one set of solution to $|\Sigma - \upsilon I|\beta = 0$ then anyother set of solution are linear combinations of the others, i.e. are B^*A for A is non-singular. However, orthogonality conditions. $B^*B = I$, applied to the linear combinations give $I = (B^*A)'(B^*A) = A'B^{*\prime}B^*A = A'A$ and thus A must be orthogonal.

9.5 Sample Principal Components

$\{X_i, i = 1, 2, \ldots, n\}$: n independent observations on X.

$$\hat{\mu}_X = \bar{X} = n^{-1} \sum_{i=1}^{n} X_i$$

Let $X_{ci} = X_i - \bar{X}, \; i = 1, 2, \ldots, n,$

$X_c = (X_{c1}, \cdots, X_{cn})$ be a $(p \times n)$-matrix

$\hat{\Sigma}_{XX} = n^{-1} S = n^{-1} X_c X_c'$

The ordered eigenvalues of $\hat{\Sigma}_{XX}$ are denoted by

$\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_p \geq 0$

and $\hat{v}_j$ is the sample eigenvector associated with the j^{th} largest sample eigenvalue $\hat{\lambda}_j \; \forall \, j \in \{1,2,..,p\}$. Hence,

$\hat{A}^{(k)} = (\hat{v}_1, \cdots, \hat{v}_k) = \hat{B}^{(k)\prime}$

where $\hat{v}_j$ is the j^{th} sample eigenvector of $\hat{\Sigma}_{\mathrm{XX}}, j = 1, 2, \ldots, k \; (k \leq p)$.

The best rank-k reconstruction of X is given by

$\hat{X}^{(k)} = \bar{X} + \hat{C}^{(k)}(X - \bar{X}),$

Where $\hat{C}^{(k)} = A^{(k)} B^{(k)} = \sum_{j=1}^{k} \hat{v}_j \hat{v}_j$

The j^{th} sample PC score of X is given by

$\hat{\xi}_j = \hat{v}_j' X_c$

A sample estimate of the measure of how well the first k principal components represent the p original variables is given by the statistic

$$V_p^{(k)} = \frac{\hat{\lambda}_{k+1} + \cdots + \hat{\lambda}_p}{\hat{\lambda}_1 + \cdots + \hat{\lambda}_p}.$$

$V_p^{(k)}$ is the proportion of the total sample variation that is explained by the last $p - k$ sample principal components.

9.5.1 How many PC to Retain?

Scree Plot: The plot of ordered eigen values against ordered numbers shows the amount of variance explained by each eigen value.

Principal component analysis can be applied to covariance matrix or correlation matrix (rescaled data). Rescaling is required when different variables measure different characteristics.

9.5.2 PC Rank Trace

k: number of PC retain

p: total number of PC

$$\Delta\hat{C}(k) = \left(1 - \frac{k}{p}\right)^{1/2}$$

$$\Delta\Sigma^{(k)} = \left(\frac{\lambda_{k+1}^2 + \cdots + \lambda_p^2}{\lambda_1^2 + \cdots + \lambda_p^2}\right)^{1/2}$$

Plot of $\Delta\Sigma^{(k)}$ against $\Delta\hat{C}(k)$ is called PC Rank Trace plot.

k is chosen as smallest positive integer between 1 and p at which an elbow can be detected in PC rank trace plot.

Example: Let $R = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$. Solve $|R - \lambda I| = 0$ or $\begin{vmatrix} (1-\lambda) & r_{12} \\ r_{12} & (1-\lambda) \end{vmatrix} = 0$

or $(1-\lambda)^2 - r_{12}^2 = 0$

or $(1 - 2\lambda + \lambda^2 - r_{12}^2) = 0$

$$\lambda = \frac{2 \pm \sqrt{4 - 4(1 - r^2)}}{2} = 1 \pm r \qquad \text{as } r_{12}^2 = r^2$$

If $r > 0$ $\quad \lambda_1 = 1 + r, \lambda_2 = 1 - r$

If $r < 0$ $\quad \lambda_1 = 1 - r, \lambda_2 = 1 + r$

If $r = 0$ $\quad \lambda_1 = 1 = \lambda_2$

Thus, in case of perfect correlation, we need one principal component which explains fully but in case of zero correlation, no principal component.

9.6 Canonical Correlation Analysis

In Canonical Correlation analysis, we consider the correlation between a linear combination of the variable in one set and a linear combination of the variables in another set.

9.6.1 Assumptions

Most of the multivariate technique assumptions apply to Canonical Correlation.

- Assumes the linear relationship between the dependent and independent variables
- Independent variables should not be highly correlated
- Uniform variability
- Additionally, multivariate normality is necessary to perform a statistical test.

9.6.2 Canonical Correlations and Canonical Variables

Suppose the random vector x has covariance matrix Σ (which is assumed to be positive definite) without loss of generality. Let $E(x) = 0$. Since we are interested only in variances and covariances. Suppose

$$x = \begin{bmatrix} x^{(1)} \\ x^{(2)} \end{bmatrix} \tag{9.10}$$

For covariance, let us assume $p_1 \leq p_2$. Then

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{9.11}$$

Consider an arbitrary linear combination $U = \alpha' x^{(1)}$ of the components of $x^{(1)}$ and an arbitrary function $V = \gamma' x^{(2)}$ of the components of $x^{(2)}$. We first look for the linear functions that have maximum correlation. Since the correlation of a multiple of U and multiple of V is the same as that of U and V, we can make an arbitrary normalization of α and γ. Thus α and γ are such that U and V have unit variance. i.e.

$$1 = E(U^2) = E\left(\alpha' x^{(1)}\, x^{(1)'}\alpha\right) = \alpha' E\left(x^{(1)}\, x^{(1)'}\right)\alpha$$

$$= \alpha' \Sigma_{11} \alpha \tag{9.12}$$

$$1 = E(V^2) = E\left(\gamma' x^{(2)}\, x^{(2)'}\gamma\right) = \gamma' E\left(x^{(2)}\, x^{(2)'}\right)\gamma$$

$$= \gamma' \Sigma_{22} \gamma \tag{9.13}$$

Note that $E(U) = E\left(\alpha' x^{(1)}\right) = \alpha' E\left(x^{(1)}\right) = 0$ and similarly

$$E(V) = E\left(\gamma' x^{(2)}\right) = \gamma' E\left(x^{(2)}\right) = 0$$

The correlation coefficient between U and V is

$$r(u, v) = \frac{cov(U, V)}{\sqrt{Var(U) Var(V)}}$$

$$= Cov(U, V)$$

$$= E(UV')$$

$$= E\left(\alpha' x^{(1)}\, x^{(2)'}\gamma\right)$$

$$= \alpha' E\left(x^{(1)}\, x^{(2)'}\right)\gamma$$

$$= \alpha' \Sigma_{12} \gamma \tag{9.14}$$

Thus, the algebraic problem is to find $\alpha \; and \; \gamma$ to maximise (9.14) subject to (9.12) and (9.13). Let

$$\psi = \alpha'\Sigma_{12}\gamma - \frac{1}{2}\lambda(\alpha'\Sigma_{11}\alpha - 1) - \frac{1}{2}\mu(\gamma'\Sigma_{22}\gamma - 1) \tag{9.15}$$

where $\lambda \; and \; \mu$ are Lagrange's multipliers. Differentiating ψ w.r.t. α and γ and equating to zero, individually, we get

$$\frac{\partial \Psi}{\partial \alpha} = \Sigma_{12}\gamma - \lambda\Sigma_{11}\alpha = 0 \tag{9.16}$$

$$\frac{\partial \Psi}{\partial \gamma} = \Sigma_{12}'\alpha - \mu\Sigma_{22}\gamma = 0 \tag{9.17}$$

Multiplying (9.16) on the left by α' and (9.17) by γ', we have

$$\alpha'\Sigma_{12}\gamma - \lambda\alpha'\Sigma_{11}\alpha = 0 \tag{9.18}$$

$$\gamma'\Sigma_{12}'\alpha - \mu\gamma'\Sigma_{22}\gamma = 0 \tag{9.19}$$

Since $\alpha'\Sigma_{11}\alpha = 1$ and $\gamma'\Sigma_{22}\gamma = 1$, then $\alpha'\Sigma_{12}\gamma = \lambda$ and $(\alpha'\Sigma_{12}\gamma)' = \mu$. From (9.16) and (9.17), (9.18) and (9.19) can be written as

$$-\lambda\Sigma_{11}\alpha + \Sigma_{12}\gamma = 0 \tag{9.20}$$

$$\Sigma_{21}\alpha - \lambda\Sigma_{22}\gamma = 0 \tag{9.21}$$

In matrix notation,

$$\begin{bmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} = 0 \tag{9.22}$$

In order that there be a non-trivial solution, and then matrix on left should be singular are. i.e.

$$\begin{bmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{bmatrix} = 0 \tag{9.23}$$

The determinant on the left is a polynomial of degree p and has p roots say, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.

We see that $\lambda = \mathrm{a}'\Sigma_{12}\gamma$ is the correlation between $U = \alpha' x^{(1)}$ and $V = \gamma' x^{(2)}$. Since we want the maximum correlation, we take $\lambda = \lambda_1$. Let a solution to (9.6.13) for $\lambda = \lambda_1$ be $\alpha^{(1)}$, $\gamma^{(1)}$ and let $U_1 = \alpha^{(1)'} x^{(1)}$ and $V_1 = \gamma^{(1)'} x^{(2)}$. Then U_1 and V_1 are normalized linear combinations of $x^{(1)}$ and $x^{(2)}$ respectively with maximum correlation.

We now find a second linear combination of $x^{(1)}$, say $U = \alpha' x^{(1)}$ and a second linear combination of $x^{(2)}$, say $V = \gamma' x^{(2)}$ such that of all linear combination uncorrelated with U_1 and V_1, having maximum correlation.

This procedure is continued and at the r^{th} step we have obtained linear combinations $U_1 = \alpha^{(1)'} x^{(1)}, V_1 = \gamma^{(1)'} x^{(2)}, \ldots, U_r = \alpha^{(r)'} x^{(1)},$

$V_r = \gamma^{(r)'} x^{(2)}$ with corresponding correlations $\lambda^{(1)} = \lambda_1, \lambda^{(2)}, \ldots \lambda^{(r)}$. We now seek a linear combination of $x^{(1)}$, $U = \alpha' x^{(1)}$ and a linear combination of $x^{(2)}$ $V = \gamma' x^{(2)}$ that of all linear combinations uncorrelated with $U_1, V_1, \ldots, U_r, V_r$ have maximum correlation. The condition that U be uncorrelated with U_i is

$$0 = E(UU_i) = E\left(\alpha' x^{(1)}\, x^{(1)'} \mathrm{a}^{(\mathrm{i})}\right) = \alpha' E\left(x^{(1)}\, x^{(1)'}\right) \mathrm{a}^{(\mathrm{i})} = \alpha' \Sigma_{11} \mathrm{a}^{(\mathrm{i})} \tag{9.24}$$

$$0 = E(UV_i) = E\left(\alpha' x^{(1)}\, x^{(2)'} \gamma^{(i)}\right) = \alpha' E\left(x^{(1)}\, x^{(2)'}\right) \gamma^{(i)} = \alpha' \Sigma_{12} \gamma^{(i)}$$

Note that

$$E(U) = E\left(\alpha' x^{(1)}\right) = \alpha' E\left(x^{(1)}\right) = 0$$

and similarly

$$E(V) = E\left(\gamma' x^{(2)}\right) = \gamma' E\left(x^{(2)}\right) = 0.$$

$$0 = E[UV_i] = E\left[\alpha' X^{(1)} X^{(2)'} \gamma^{(i)}\right] = \alpha' \Sigma_{12}\, \gamma^{(i)} = \lambda^{(i)}\, \alpha' \Sigma_{11}\, \alpha^{(i)} = 0 \tag{9.25}$$

The condition that V is uncorrelated with V_i is

$$0 = E[VV_i] = \gamma' E\left[X^{(2)} X^{(2)'}\right] \gamma^{(i)} = \gamma' \Sigma_{22}\, \gamma^{(i)} = 0 \tag{9.26}$$

Again, by the same argument,

$$0 = E[VU_i] = E\left[\gamma' X^{(2)} X^{(1)'} \alpha^{(i)}\right] = \gamma' \Sigma_{21}\, \alpha^{(i)} = 0 \tag{9.27}$$

We now maximize $E[U_{r+1}V_{r+1}]$, choosing α and γ to satisfy (9.12), (9.13), (9.25) and (9.26) for $i = 1,2,\ldots,r$. For this consider

$$\psi_{r+1} = \alpha' \Sigma_{12} \gamma - \frac{1}{2}\lambda(\alpha' \Sigma_{11} \mathrm{a} - 1) - \frac{1}{2}\mu(\gamma' \Sigma_{22} \gamma - 1) + \sum_{i=1}^{r} \upsilon_i \alpha' \Sigma_{11} \alpha^{(i)}$$

$$+ \sum_{i=1}^{r} \theta_i\ \gamma' \Sigma_{22}\, \gamma^{(i)}$$

$\lambda, \mu, \gamma_1, \ldots, \gamma_r, \theta_1, \ldots, \theta_r$ are Lagrange's multipliers. The vector of partial derivatives of ψ_{r+1} with respect to α and γ are set equal to zero, giving

$$\frac{\partial \psi_{r+1}}{\partial \alpha} = \Sigma_{12}\gamma - \lambda \Sigma_{11}\alpha + \sum_{i=1}^{r} \upsilon_i\, \Sigma_{11} \alpha^{(i)} = 0 \tag{9.28}$$

$$\frac{\partial \psi_{r+1}}{\partial \gamma} = \Sigma_{12}'\alpha - \mu \Sigma_{22}\gamma + \sum_{i=1}^{r} \theta_i\, \Sigma_{22}\, \gamma^{(i)} = 0 \tag{9.29}$$

Multiplying (9.28) on the left by $\alpha^{(j)'}$ and (9.29) on the left by $\gamma^{(j)'}$, we have

$$0 = \upsilon_j \alpha^{(j)'} \Sigma_{11} \alpha^{(j)} = \upsilon_j$$

$$0 = \theta_j\, \gamma^{(j)'} \Sigma_{22}\, \gamma^{(j)} = \theta_j$$

Equations (9.28) and (9.29) are simply (9.20) and (9.21) or alternatively (9.22). Hence, we take the largest λ, say $\lambda^{(r+1)}$ such that there exists a solution (13) satisfying (9.12), (9.13), (9.14) and (9.16) for $i = 1,2,\ldots,r$. Let this solution be $\alpha^{(r+1)}, \gamma^{(r+1)}$ and let $U_{r+1} = \alpha^{(r+1)'} X^{(1)}$ and $V_{r+1} = \gamma^{(r+1)'} X^{(2)}$.

This procedure is continued step by step as long as a successive solution can be found which satisfy (9.22) for some λ_i, (9.12), (9.13), (9.24) and (9.26). Let m be the number of steps for which this can be done. Now it can be shown that $m = p_1 (\leq p_2)$. Let

$A = \left(\alpha^{(1)}, \alpha^{(2)}, \ldots, \alpha^{(m)}\right)$,

$\Gamma = \left(\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(m)}\right)$

and

$$\Lambda = \begin{pmatrix} \lambda^{(1)} & 0 & \cdots & 0 \\ 0 & \lambda^{(2)} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda^{(\mathrm{m})} \end{pmatrix}$$

Conditions (9.12) and (9.24) can be summarized as $A'\Sigma_{11}A = I$

Here Σ_{11} is of rank p_1 and I is of rank m, $(m \leq p_1)$. Now if $m < p_1$, then there is another vector satisfying the conditions. Since $A'\Sigma_{11}$ is of order $m \times p_1$, there exists a $p_1 \times (p_1 - m)$ matrix E (of rank $p_1 - m$) such that $A'\Sigma_{11}E = 0$. Similarly, there exists a $p_2 \times (p_2 - m)$ matrix F (of rank $p_2 - m$) such that $\Gamma_1'\Sigma_{22}F = 0$. Also $\Gamma_1'\Sigma_{21}E = \Lambda\, A'\Sigma_{11}E = 0$ and $A'\Sigma_{12}F = \Lambda\, \Gamma_1'\Sigma_{22}F = 0$. Since E is of rank $(p_1 - m)$, $E'\Sigma_{11}E$ is non-singular (if $m < p_1$) and similarly $F'\Sigma_{22}F$ is non-singular. Thus, there is at least one root of

$$\begin{vmatrix} -\gamma E'\Sigma_{11}E & E'\Sigma_{12}F \\ F'\Sigma_{21}E & -F'\Sigma_{22}F \end{vmatrix} = 0$$

Because $|E'\Sigma_{11}E||F'\Sigma_{22}F| \neq 0$. From the preceding algebra, there exists vectors a and b such that

$$E'\Sigma_{12}F\, b = \gamma E'\Sigma_{11}E\, a \tag{9.30}$$

$$F'\Sigma_{21}E\, a = \gamma F'\Sigma_{22}F\, b$$

Let $E\, a = g$ and $F\, b = h$. Now, to obtain v, g and h for a new solution $\lambda^{(m+1)}, \alpha^{(m+1)}$ and $\gamma^{(m+1)}$, let $\Sigma_{11}^{-1}\Sigma_{12}h = k$. Since $A'\Sigma_{11}k = A'\Sigma_{12}F\, b = 0$, k is orthogonal to the rows of $A'\Sigma_{11}$ and therefore is a linear combination of E, say Ec.Thus the equation $\Sigma_{12}h = \Sigma_{11}k$ can be written as

$\Sigma_{12}F\, b = \Sigma_{11}E\, c$

Multiplying by E' on the left,

$$E'\Sigma_{12}F\, b = E'\Sigma_{11}E\, c \tag{9.31}$$

Since $E'\Sigma_{11}E$ is non-singular, comparison of (9.30) and (9.31) shows that $c = \upsilon a$ and therefore, $k = \upsilon g$. Thus

$\Sigma_{12}h = \upsilon\Sigma_{11}g$

Similarly, $\Sigma_{21}g = \upsilon\Sigma_{22}h$

Therefore $\upsilon = \lambda^{(m+1)}, g = \alpha^{(m+1)}, h = \gamma^{(m+1)}$ is another solution But this contradicts the fact that $\lambda^{(m)}, \alpha^{(m)}\ and\ \gamma^{(m)}$ was the last possible solution. Hence $m = p_1$.

The conditions on the $\lambda's, \alpha's\ and\ \gamma's$ can be summarized as

$A'\Sigma_{11}A = I$

$A'\Sigma_{12}\Gamma_1' = \Lambda$

$\Gamma_1'\Sigma_{22}\Gamma_1' = \mathrm{I}$

Let $\Gamma_2' = \left(\gamma^{(p_1+1)}, \gamma^{(p_1+2)}, \ldots, \gamma^{(p_2)}\right)$ be a $p_2 \times (p_2 - p_1)$ matrix satisfying

$\Gamma_2'\Sigma_{22}\Gamma_1 = 0$

$\Gamma_2'\Sigma_{22}\Gamma_2 = \mathrm{I}$

This matrix can be formed one column at a time, $\gamma^{(p_1+1)}$ is a vector orthogonal to $\Sigma_{22}\Gamma_1$ and normalized so $\gamma^{(p_1+1)'}\Sigma_{22}\,\gamma^{(p_1+1)} = 1$, $\gamma^{(p_1+2)}$ is a vector orthogonal to $\Sigma_{22}\left(\Gamma_1\gamma^{(p_1+1)}\right)$ and normalized so $\gamma^{(p_1+2)'}\Sigma_{22}\,\gamma^{(p_1+2)} = 1$ and so on. Let $\Gamma' = (\Gamma_1'\ \Gamma_2')$. This is a square matrix and is non-singular as $\Gamma'\Sigma_{22}\Gamma = \mathrm{I}$.

Consider the determinant

$$\begin{vmatrix} A' & 0 \\ 0 & \Gamma_1 \\ 0 & \Gamma_2 \end{vmatrix} \begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} \begin{vmatrix} A & 0 & 0 \\ 0 & \Gamma_1 & \Gamma_2 \end{vmatrix} = \begin{vmatrix} -\lambda I & \Lambda & 0 \\ \Lambda & -\lambda I & 0 \\ 0 & 0 & -\lambda I \end{vmatrix} = (-\lambda)^{p_2-p_1} \begin{vmatrix} -\lambda I & \Lambda \\ \Lambda & -\lambda I \end{vmatrix}$$

$$= (-\lambda)^{p_2-p_1}|-\lambda I||-\lambda I - \Lambda(-\lambda\mathrm{I})^{-1}\Lambda| = (-\lambda)^{p_2-p_1}|\lambda^2 I - \Lambda^2|$$

$$= (-\lambda)^{p_2-p_1}\prod_{i=1}^{p_1}\left(\lambda^2 - \lambda^{(i)^2}\right) \qquad (9.32)$$

Except for a constant factor the above polynomial is same as

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix}$$

Thus, the roots of (9.23) are roots of (9.32) set equal to zero, namely $\lambda = \pm\lambda^{(i)}, i = 1,2, \dots, p_1$ and $\lambda = 0$ (of multiplicity $p_2 - p_1$). Thus

$$(\lambda_1, \dots, \lambda_{p_2}) = (\lambda_1, \dots, \lambda_{p_1}, 0, \dots, 0, -\lambda_{p_1}, \dots, -\lambda_1)$$

The set $\{\lambda^{(i)^2}\}, i = 1,2, \dots, p_1$ is the set $\{\lambda_i^2\}, i = 1,2, \dots, p_1$. To show that $\{\lambda^{(i)}\}$ is $\{\lambda_i\}, i = 1,2, \dots, p_1$, it is sufficient to show that $\lambda^{(i)}$ are non-negative and therefore is one of $\lambda_i, i = 1,2, \dots, p_1$. Observe that

$$\Sigma_{12}\gamma^{(r)} = -\lambda^{(r)}\Sigma_{11}(-\alpha^{(r)})$$

$$\Sigma_{21}(-\alpha^{(r)}) = -\lambda^{(r)}\Sigma_{22}\gamma^{(r)}$$

If $\lambda^{(r)}, \alpha^{(r)}, \gamma^{(r)}$ is a solution, then so is $-\lambda^{(r)}, -\alpha^{(r)}, -\gamma^{(r)}$. If $\lambda^{(r)}$ were negative, then $-\lambda^{(r)}$ would be non-negative and $-\lambda^{(r)} \geq \lambda^{(r)}$. But $\lambda^{(r)}$ was supposed to be maximum, which implies $\lambda^{(r)} \geq -\lambda^{(r)}$ and hence $\lambda^{(r)}$ cannot be negative, i.e., $\lambda^{(r)} \geq 0$.

Since the set $\{\lambda^{(i)}\}$ is same as the set $\{\lambda_i\}, i = 1,2, \dots, p_1$, we must have $\lambda^{(i)} = \lambda_i$, let

$$U = \begin{pmatrix} U_1 \\ \vdots \\ U_p \end{pmatrix} = A'X^{(1)}$$

$$V^{(1)} = \begin{pmatrix} V_1 \\ \vdots \\ V_{p_1} \end{pmatrix} = \Gamma_1'X^{(2)}$$

$$V^{(2)} = \begin{pmatrix} V_{p_1+1} \\ \vdots \\ V_{p_2} \end{pmatrix} = \Gamma_2'X^{(2)}$$

The components of U are one set of canonical variates and the components of $V = \begin{pmatrix} V^{(1)} \\ V^{(2)} \end{pmatrix}$ are the other set. We have

$$E\left\{\begin{pmatrix} U \\ V^{(1)} \\ V^{(2)} \end{pmatrix}\left(U'V^{(1)'}V^{(2)'}\right)\right\}$$

$$= E\left\{\begin{pmatrix} A' & 0 \\ 0 & \Gamma_1' \\ 0 & \Gamma_2' \end{pmatrix}\begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}\left(X^{(1)'}X^{(2)'}\right)\begin{pmatrix} A & 0 & 0 \\ 0 & \Gamma_1 & \Gamma_2 \end{pmatrix}\right\}$$

$$= \begin{pmatrix} A' & 0 \\ 0 & \Gamma_1' \\ 0 & \Gamma_2' \end{pmatrix}\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\begin{pmatrix} A & 0 & 0 \\ 0 & \Gamma_1 & \Gamma_2 \end{pmatrix}$$

$$= \begin{pmatrix} I_{p_1} & \Lambda & 0 \\ \Lambda & I_{p_1} & 0 \\ 0 & 0 & I_{p_1-p_2} \end{pmatrix}$$

where $\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_{p_1} \end{pmatrix}$

Definition: Let $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$, where $X^{(1)}$ has p_1 components and $X^{(2)}$ has p_2 $(= p - p_1 \geq p_1)$ components. The r^{th} pair of canonical variates is the pair of linear combinations $U_r = \alpha^{(r)'}X^{(1)}$ and $V_r = \gamma^{(r)'}X^{(2)}$ each of unit variance and uncorrelated with the first $(r-1)$ pairs of canonical variates and having maximum correlation. The correlation is the r^{th} canonical correlation.

Theorem 9.6.1: Let $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ be a random vector with covariance matrix Σ. The r^{th} canonical correlation between $X^{(1)}$ and $X^{(2)}$ is the r^{th} largest root of

$$\begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} = 0$$

The coefficients $U_r = \alpha^{(r)}X^{(1)}$ and $V_r = \gamma^{(r)}X^{(2)}$ defining the r^{th} pair of canonical variates satisfy

$$\begin{bmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{bmatrix}\begin{bmatrix} \alpha \\ \gamma \end{bmatrix} = 0.$$

For $\lambda = \lambda_r$, $\alpha'\Sigma_{11}\alpha = 1$ and $\gamma'\Sigma_{22}\gamma = 1$

Proof: First, we show that U_1, V_1 have maximum correlation.

Let

$U' = \left(U_1, U_2, \dots, U_{p_1}\right)'$,

$V' = \left(V_1, V_2, \dots, V_{p_2}\right)'$,

$A = \left(\alpha^{(1)}, \alpha^{(1)}, \dots, \alpha^{(p_1)}\right)$,

$\Gamma = \left(\gamma^{(1)}, \gamma^{(2)}, \dots, \gamma^{(p_2)}\right)$

Then A and Γ are non-singular, and $U = A'X^{(1)}, V = \Gamma'X^{(2)}$. The linear combinations are

$a'U = (a'A')X^{(1)}$,

$b'V = (b'\Gamma')X^{(2)}$

These linear combinations are normalized by $a'a = 1$ and $b'b = 1$.

Since A and Γ are non-singular, any vector α can be written as Aa, i.e. $\alpha = Aa$ and any vector γ can be written as Γb, i.e. $\gamma = \Gamma b$. Hence any linear combination $\alpha'X^{(1)}$ and $\gamma'X^{(2)}$ can be written as $a'U \; and \; b'V$. The correlation between them is

$E(a'U\,V'b) = a'E(U\,V')b$

$= a'E\left(A'X^{(1)}X^{(2)'}\Gamma\right)b$

$= a'E(A'\Sigma_{12}\Gamma)b$

$= a'[\Lambda \quad 0]b$

$= \sum_{i=1}^{p_1} \lambda_i a_i b_i$

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_{p_1} \end{pmatrix}$$

Let

$$C_i = \frac{\lambda_i a_i}{\sqrt{\sum_{i=1}^{p_1}(\lambda_i a_i)^2}}$$

Then the maximum of $a'[\Lambda \quad 0]b = \sum_{i=1}^{p_1} C_i b_i \sqrt{\sum_{i=1}^{p_1}(\lambda_i a_i)^2}$ with respect to b is for $b_i = C_i$, since $\sum_{i=1}^{p_1} C_i b_i$ is the cosine of the angle between the vector b and $C = (C_1, C_2, \dots, C_n, 0,0, \dots, 0)$. Then

$$\sum_{i=1}^{p_1} C_i^2 \sqrt{\sum_{i=1}^{p_1}(\lambda_i a_i)^2}$$

$$= \sqrt{\sum_{i=1}^{p_1} \lambda_i^2 a_i^2}$$

$$= \sqrt{\sum_{i=2}^{p_1}(\lambda_i^2 - \lambda_1^2)a_1^2 + \lambda_1^2 a_1^2}$$

And this is maximized by taking $a_i = 0, i = 2,3, \dots, p_1$. Thus, maximized linear combination are U_1, and V_1. To verify $U_2, and\ V_2$ from the second pair of canonical variates, we note that U_1 will be uncorrelated with a linear combination $a'U$ if

$$0 = E(U_1 a'U) = E\left(U_1 \sum_{i=1}^{p_1} a_i U_i\right) = a_1$$

Similarly, V_1 will be uncorrelated with a linear combination $b'V$ if

$$0 = E(V_1 b'V) = E\left(V_1 \sum_{i=1}^{p_2} b_i V_i\right) = b_1$$

This proves the theorem.

Theorem 9.6.2: The canonical correlations are invariant with respect to transformations $X^{(i)^*} = C_i X^{(i)}$, where C_i is non-singular, $i = 1,2$ and any function of Σ that is invariant is a function of canonical correlation.

Proof: $X^{(i)^*} = C_i X^{(i)} \Rightarrow X^* = \begin{pmatrix} X^{(1)^*} \\ X^{(2)^*} \end{pmatrix} = \begin{pmatrix} C_1 X^{(1)} \\ C_2 X^{(2)} \end{pmatrix}$

$\Sigma = E(XX') = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$

$E(X^* X^{*\prime}) = E \begin{pmatrix} X^{(1)^*} \\ X^{(2)^*} \end{pmatrix} (X^{(1)^{*\prime}} \quad X^{(2)^{*\prime}}) = \begin{pmatrix} \Sigma_{11}^* & \Sigma_{12}^* \\ \Sigma_{21}^* & \Sigma_{22}^* \end{pmatrix} = \begin{pmatrix} C_1\Sigma_{11}C_1' & C_1\Sigma_{12}C_2' \\ C_2\Sigma_{21}C_1' & C_2\Sigma_{22}C_2' \end{pmatrix}$

Hence

$-\lambda\Sigma_{11}\alpha + \Sigma_{12}\gamma = 0$

$\Sigma_{21}\alpha - \lambda\Sigma_{22}\gamma = 0$

The above equations can be written as

$-\lambda C_1\Sigma_{11}C_1'\alpha + C_1\Sigma_{12}C_2'\gamma = 0$

$C_2\Sigma_{21}C_1'a - \lambda C_2\Sigma_{22}C_2'b = 0$

In order that there is non-trivial solution, we should have

$0 = \begin{vmatrix} -\lambda C_1\Sigma_{11}C_1' & C_1\Sigma_{12}C_2' \\ C_2\Sigma_{21}C_1' & -\lambda C_2\Sigma_{22}C_2' \end{vmatrix}$

$= \begin{vmatrix} C_1 & 0 \\ 0 & C_2 \end{vmatrix} \begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} \begin{vmatrix} C_1' & 0 \\ 0 & C_2' \end{vmatrix}$

$\Rightarrow 0 = \begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix}$

$\Rightarrow$ Roots are unchanged.

Conversely, let $f(\Sigma_{11}, \Sigma_{12}, \Sigma_{22})$ be a vector valued function of Σ such that

$f(C_1\Sigma_{11}C_1',\ C_1\Sigma_{12}C_2',\ C_2\Sigma_{22}C_2')$

$= f(\Sigma_{11}, \Sigma_{12}, \Sigma_{22})\ \forall$ non singular C_1 and C_2

If $C_1 = A$ and $C_2 = \Gamma' = (\Gamma_1' \quad \Gamma_2')$, then

$$0 = \begin{vmatrix} A' & 0 \\ 0 & \Gamma_1 \\ & \Gamma_2 \end{vmatrix} \begin{vmatrix} -\lambda\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & -\lambda\Sigma_{22} \end{vmatrix} \begin{vmatrix} A & 0 \\ 0 & \Gamma_1'\Gamma_2' \end{vmatrix}$$

$$= \begin{vmatrix} -\lambda I & \Lambda & 0 \\ \Lambda & -\lambda I & 0 \\ 0 & 0 & -\lambda I \end{vmatrix}$$

This depends only on the canonical correlation. Then $f = f(I, (\Lambda \quad 0), I)$.

Suppose $p \leq q$ and let the p −dimensional random vectors $X^{(1)}$ and q −dimensional $X^{(2)}$ have

$Cov\left(X^{(1)}\right) = \Sigma_{11}$,

$Cov\left(X^{(2)}\right) = \Sigma_{22}$,

$Cov(X^{(1)}, X^{(2)}) = \Sigma_{12}$

The covariance matrix of $X = \left(X^{(1)'} X^{(2)'}\right)'$ is

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

We assume that Σ has full rank. For coefficients $p \times 1$ vector a and $q \times 1$ vector b, consider the linear combination $U = a'X^{(1)}$and $V = b'X^{(2)}$. Then

$Var(U) = E(UU')$

$= E\{a'X^{(1)}X^{(1)'}a\}$

$= a'E\{X^{(1)}X^{(1)'}\}a = a'\Sigma_{11}a$,

$Var(V) = E(VV')$

$= E\{b'X^{(2)}X^{(2)'}b\}$

$= b'E\{X^{(2)}X^{(2)'}\}b = b'\Sigma_{22}b.$

Since

$E(U) = E\left(a'X^{(1)}\right) = a'E\left(X^{(1)}\right) = 0$ and $E(V) = E\left(b'X^{(2)}\right) = b'E\left(X^{(2)}\right) = 0$

We have

$$Corr\ (U,V) = Cov(U,V) = E\{U - E(U)\}\{V - E(V)\} = E(U\,V) = E\left\{a'X^{(1)}X^{(2)'}b\right\}$$

$$= a'\Sigma_{12}b,$$

We shall seek coefficient vectors a and b such that

(i) The first pair of canonical variables is the pair linear combination U_1 and V_1 having unit variances, which maximize the correlation $Corr\ (U,V)$.

(ii) The second pair of canonical variables is the pair of linear combinations U_1 and V_2 having unit variances, which maximize the correlation $Corr\ (U,V)$ among all choices that are uncorrelated with the first pair of canonical variables.

(iii) In general, the k^{th} pair of canonical variables is the pair of linear combinations U_k and V_k having unit variances, which maximize the correlation $Corr(U,V)$ among all choices uncorrelated with the previous $(k\ -\ 1)$ canonical variable pairs.

The correlation between the k^{th} pair of canonical variables is called the k^{th} canonical correlation.

The maximum of correlation, say,

$\max Corr\ (U,V) = \rho_1\ (a,b) = \rho_1^*\ (say)$

is attained by the linear combinations (first canonical variate pair), when

$U_1 = e_1'\Sigma_{11}^{-\frac{1}{2}}X^{(1)}$,

$V_1 = f_1'\Sigma_{22}^{-\frac{1}{2}}X^{(2)}$

In general, the k^{th} pair of canonical variates, $k\ =\ 2,3,\ldots,p$

$U_k = e_k'\Sigma_{11}^{-\frac{1}{2}}X^{(1)}$,

$V_k = f_k'\Sigma_{22}^{-\frac{1}{2}}X^{(2)}$,

maximize

$max\ Corr(U,V) = \rho_k(a,b) = \rho_k^*\ (say)$

among all those linear combinations uncorrelated with the preceding $1,2,\ldots,k-1$ canonical variables.

Here

(i) $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_p^{*2}$ are the eigen values of $\Sigma_{11}^{-\frac{1}{2}}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}$ and $e_1, \ldots, e_p$ are corresponding $p \times 1$ eigenvectors.

(ii) $\rho_1^{*2}, \rho_2^{*2}, \ldots \geq \rho_p^{*2}$ are also the p largest eigenvalues of the matrix $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ with corresponding $q \times 1$ eigenvectors $f_1, f_2, \ldots, f_p$.

(iii) f_i is proportional to $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-\frac{1}{2}}e_i\ \forall i$.

(iv) The canonical variates have the properties

$$Var(U_k) = Var(V_k) = 1, k = 1, \ldots, p,$$

$$Corr(U_k, U_l) = Corr(V_k, V_l) = Corr(U_k, V_l) = 0 \forall\ k, l = 1,2,\ldots,p; k \neq l$$

9.7 Sample Canonical Variates and Sample Canonical Correlation

Consider a sample of random vectors

$$X_i = \begin{pmatrix} X_i^{(1)} \\ X_i^{(2)} \end{pmatrix}; i = 1,2, \ldots, N.$$

Let

$$S = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(X_i - \bar{X})' = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}.$$

Finding canonical variables and canonical correlations, replace population distribution by empirical distribution, replace Σ by S, and replace ρ by R. Then the results the same as above.

Let $\hat{\rho}_1^{*2} \geq \hat{\rho}_2^{*2} \geq \cdots \geq \hat{\rho}_p^{*2}$ be the p ordered eigen-values of $S_{11}^{-\frac{1}{2}} S_{12} S_{22}^{-1} S_{21} S_{11}^{-\frac{1}{2}}$ with corresponding eigenvectors $\hat{e}_1, \ldots, \hat{e}_p$, $(p \leq q)$. Further, $\hat{f}_1, \ldots, \hat{f}_p$ be the eigenvectors of $S_{22}^{-\frac{1}{2}} S_{21} S_{11}^{-1} S_{12} S_{22}^{-\frac{1}{2}}$. Then the k^{th} sample canonical variate pair is

$$\hat{U}_k = \hat{e}_k' \, S_{11}^{-\frac{1}{2}} x^{(1)}, \hat{V}_k = \hat{f}_k' \, S_{22}^{-\frac{1}{2}} x^{(2)}$$

where $x^{(1)}$and $x^{(2)}$are specific values of $X^{(1)}$and $X^{(2)}$ respectively. Further, for the k^{th} pair, $(k = 1, \ldots, p), r_{\hat{U}_k, \hat{V}_k} = \hat{\rho}_k^*$.

$\hat{\rho}_1^*, \hat{\rho}_2^*, \ldots, \hat{\rho}_p^*$ are the sample canonical correlations.

9.7.1 Difference between Multiple Correlation and Canonical Correlation

Generally, we study the relationship between one dependent and independent variable in a simple correlation. Similarly, we study the relationship between one dependent variable and multiple independent variables in Multiple Correlations. In other words, it investigates the relationship between a variable Y and a set of variables $(X_1, X_2, \cdots, X_n)$. In Canonical Correlation, we study the relationship between two sets of variables. It is like simple correlation coefficient r. However, we have more than one dependent variable.

9.7.2 Application of Canonical Correlation

1. **Psychology:** Canonical Correlation Analysis can be used to explore the relationship between personality traits and job performance. It can also be used to understand the relationship between mental health factors and academic achievement.
2. **Economics:** It can help to analyse the relationship between various economic indicators (like GDP, inflation, etc.) and social indicators (like education levels, healthcare access, etc.) to understand their interdependencies.

3. **Medicine:** In medical research, it can be applied to study the relationship between genetic factors and disease outcomes, or to explore the relationship between different treatment methods and patient outcomes.
4. **Ecology:** It is useful for studying the relationship between environmental variables (like temperature, humidity, etc.) and biological variables (like species diversity, population sizes, etc.) to understand ecological processes.
5. **Neuroscience:** It can be used to analyse brain imaging data (like fMRI or EEG) to understand the relationship between brain activity patterns and cognitive processes.
6. **Marketing and Customer Relationship Management:** It can help to identify the underlying factors that govern customer behaviour and preferences, which can be useful for targeted marketing strategies.
7. **Social Sciences:** It can be used to explore the relationship between different social factors (like income, education, etc.) and outcomes (like happiness, well-being, etc.) to understand societal trends.
8. **Climate Science:** It can be applied to study the relationship between climate variables (like temperature, precipitation, etc.) and their impacts on ecosystems and human populations.

9.7.3 Advantages of Canonical Correlation

1. **Identifying Relationships:** Canonical Correlation Analysis can reveal underlying relationships between two sets of variables, even when the variables within each set are highly correlated.
2. **Dimensionality Reduction:** It can reduce the dimensionality of the data by identifying the most important linear combinations of variables in each set.
3. **Interpretability:** The results of canonical correlation analysis are often easy to interpret, as the canonical variables represent the most correlated pairs of variables between the two sets.
4. **Multivariate Analysis:** Canonical correlation analysis allows for the analysis of multiple variables simultaneously, making it suitable for studying complex relationships.

5. **Robustness:** Canonical correlation analysis is robust to violations of normality assumptions and can handle small sample sizes.

9.7.4 Limitations of Canonical Correlation

1. **Linear Relationships:** Canonical correlation analysis assumes that the relationships between variables are linear, which may not always be the case in real-world data.
2. **Sensitivity to Outliers:** It can be sensitive to outliers, which can affect the estimation of the canonical correlations and vectors.
3. **Interpretation of Canonical Variables:** While the canonical variables are easy to interpret, interpreting the original variables in terms of these canonical variables can be challenging.
4. **Assumption of Equal Covariances: It** assumes that the two sets of variables have same population covariance matrices, which may not hold true in practice.

9.8 Summary

In this unit, we have covered the following points.

- Principal Component Analysis is a dimensional reduction technique in which we derive a small number of linear combinations (principal components) of a set of variables that retain as much information in the original variables as possible.
- Principal Components can be derived from covariance matrix or correlation matrix.
- For obtaining principal components, one must know the eigenvalues of the sample covariance/ correlation matrix.
- Canonical correlation is a technique to identify and quantify the association between two sets of variables.
- Canonical correlation analysis focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in the second set.

9.9 Self-Assessment Exercises

1. Find the variance of the first principal component of the covariance matrix Σ defined by

$$\Sigma = (1-\rho)I + \rho ee'$$

Where $\underline{e'} = (1 \quad 1 \quad \cdots \quad 1)$

2. Find the characteristic vector of $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ corresponding to the characteristic roots $(1+\rho)$ and $(1-\rho)$.

3. What is the main objective of principal component? If $X \sim N_p(0, \Sigma)$ then prove that there exists a linear transformation $U = BX$, such that the covariance matrix of U is $E(UU') = \Lambda = diag(\lambda_1, \ldots, \lambda_p)$ with $\lambda_1 \geq \cdots \geq \lambda_p$, where $\lambda_1, \ldots, \lambda_p$ are roots of $|\Sigma - \lambda I| = 0$, and the r[th] column of B, say $\beta^{(r)}$ satisfies $(\Sigma - \lambda_r I)\beta^{(r)} = 0$. How do we select the appropriate number of principal components.

4. Define canonical correlation and canonical variables. How do we estimate different canonical correlations.

5. Show that square of canonical correlations is invariant under nonsingular linear transformation.

9.10 References

- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.
- Giri Narayan, C. (1995): Multivariate Statistical Analysis.
- Khatri, C. G.: Multivariate Analysis
- Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York

9.11 Further Reading

- Hotelling, H. (1933): Analysis of a complex of statistical variables into principal components: Journal of Educational Psychology.
- Hotelling, H. (1936): Relations between two sets of variates: Biometrika.
- Mardia, K. V.: Multivariate Analysis.

- Rencher, Alvin C.: Multivariate Statistical Inference and Applications. John Wiley, New York.
- Seber, G.A.F.: *Multivariate Observations.* Wiley, New York.

UNIT: 10 FACTOR ANALYSIS

Structure

10.1 Introduction

The objective of Factor Analysis is to model the observed variables, and their covariance structure, in terms of a smaller number of underlying latent factors. It is a descriptive method like the principal components. Factor analysis may be considered as an inversion of principal components. In principal components, we create new variables that are linear combinations of the observed variables whereas in factor analysis we model the observed variables as linear functions of the factors. Both Principal Component Analysis and Factor Analysis aim at reducing the dimension of the data. Different areas where factor analysis is frequently used are psychology, health sciences, medicine, sociology, ecology, blind source separation in data mining, etc.

10.2 Objectives

Upon completion of this unit, you should be able to:

- Understand the terminology of factor analysis, including the interpretation of factor loadings, specific variances, and commonalities;
- Understand factor rotation, and interpret rotated factor loadings.

10.3 Factor Analysis

Factor analysis is used to uncover the latent structure of a set of variables. It reduces attribute space from a large no. of variables to a smaller no. of factors and as such is a non-dependent procedure.

Factor analysis could be used for any of the following purpose

1. To reduce a large number of variables to a smaller number of factors for modelling purposes, where the large number of variables precludes modelling all the measures, individually. As such factor analysis is integrated with structural equation modelling, helping create the latent variables modelled by SEM (structure equation model).

2. To select a subset of variables from a large set based on which original variables have the highest correlations with the principal component factors.

3. To create a set of factors to be treated as uncorrelated variables as one approach to handling multicollinearity regression.

10.3.1 Purpose of Factor Analysis

The main purpose of factor analysis is:

1. To identify the underlying structure of the relationship among variables and classify them into homogeneous groups or clusters, that is referred to as factors.

2. Data reduction as it decreases the number of variables and clusters them under factors.

3. To summarize and understand the data by identifying the relationship among the variables.

4. To understand and confirm the latent variables of Structural equation modeling.

5. To remove redundancy or duplicity from a set of correlated variables.

6. To identify and distinguish between Latent variables (that are called factors) and Observed variables within the data set.

7. To identify orthogonal factors that are independent of each other.

10.3.2 Assumptions of Factor Analysis

There is a simple list of fundamental assumptions that underlie factor analysis and distinguish it from principal component analysis.

1) The correlations and covariances that exist between m variables are a result of p underlying, mutually uncorrelated factors, i.e. $p < m$.

2) Usually, p is known in advance. The number of factors, hidden in the data set, is one of the pieces of a priori knowledge that is brought to the table to solve the factor analysis problem.

3) The rank of a matrix and the number of eigenvectors are interrelated, the eigenvalues are the square of the non-zero singular values. The eigenvalues are ordered by the amount of variance accounted for.

Factor analysis starts with the basic principal component approach, but differs in two important ways. First of all, factor analysis is always done with standardized data. This implies that we want the individual variables to have equal weight in their influence on the underlying

variance-covariance structure. In addition, this requirement is necessary for us to be able to convert the principal component vectors into factors. Secondly, the eigenvectors must be computed in such a way that they are normalized, i.e. of unit length or orthonormal.

10.3.3 Uses of Factor Analysis

1. Scale Construction: Factor analysis could be used to develop concise multiple-item scales for measuring various constructs. For example: a 15-item scale to measure job satisfaction.

• At the first step-Generate large number of statements, numbering say 100 or so as a part of exploratory research.

• Assume that we get 3 factors out of it.

• We want to construct a 15-item scale to measure job satisfaction.

• Separate 5 items from each factor having the highest factor loading.

2. Establish Antecedents: This method reduces multiple input variables into grouped factors. Thus, the independent variables can be grouped into broad factors. For example, the variables that measure safety clauses in mutual funds could be reduced to a factor called SAFETY CLAUSE.

3. Psychographic Profiling: Different independent variables are grouped to measure independent factors. These are then used for identifying personality types.

▪ Psychographics can be defined as a quantitative methodology used to describe consumers on psychological attributes.

▪ When a relatively complete profile of a person or group's psychographic makeup is constructed, this is called a "psychographic profile".

▪ Some categories of psychographic factors used in market segmentation include activity, interest, opinion (AIOs), attitudes, values, behaviour, etc.

4. Segmentation Analysis: Factor analysis could also be used for segmentation. For example: There could be different sets of two-wheeler customers owning two-wheelers because of the different importance they give to factors like prestige, economy consideration, Traffic, Time, functional features, etc.

5. Marketing Studies: The technique has extensive use in the field of marketing and can be successfully used for new product development; product acceptance research, development of advertising copy, pricing studies, and for branding studies. For example: It can be used to:

- Identify the attributes of brands that influence consumer's choice;
- get an insight into the media habits of various consumers;
- identify the characteristics of price-sensitive customers etc.

10.3.4 Types of Factor Analysis

The factor analysis is of two types:

1. Exploratory Factor Analysis (EFA): It is the most common factor analysis method used in multivariate statistics to uncover the underlying structure of a relatively large set of variables. It assumes that any indicator or variable may be associated with any factor to identify the underlying relationship between measured variables. It is not based on any prior theory and uses Multiple Regression and partial correlation theory to model sets of manifest or observed variables.

2. Confirmatory Factor Analysis (CFA): It is the second most preferred method to extract the common variance and put them into factors. It determines the factor and factor loading of measured variables. It also confirms what is expected from the basic or pre-established theory by assuming that each factor is associated with a specified subset of measured variables.

Steps in Exploratory Factor Analysis

1. Collect Data: choose relevant variables.
2. Extract initial factors (via principal component).
3. Choose the number of factors to retain.
4. Choose estimation method, estimate model.
5. Rotate and interpret.
6. (a) Decide on changes that need to be made (e.g. drop items include items)

   (b) Repeat (4), (5).
7. Construct scales and use them for further analysis.

10.3.5 Methods of Factor Analysis

The basic data used for factor analysis is the same for the correlation matrix when using the different procedures of analysis. Though there may be procedures that make use of the matrix of covariance. The major methods of factor analysis used are:

1. Principal Component Method
2. Principal Axes Method
3. Summation Method
4. Centroid Method

Principal Component Method: In this method, factors are selected one at a time such that each factor best fits the data. The first fraction is created such that it represents the most highly correlated set of variables. Each subsequent selected factor explains less variance than its predecessor. This procedure is continued till all the factors are selected. All the factors selected explain the largest amount of residual variance in the entire set of standardized response scores.

Principal Axis Method: This is a method that tries to find the lowest number of factors that can account for the variability in the original variables that are associated with these factors (this is in contrast to the principal components method which looks for a set of factors which can account for the total variability in the original variables). These two methods will tend to give similar results if the variables are quite highly correlated and/or the number of original variables is quite high. Whichever method is used, the resulting factors at this stage will be uncorrelated.

Centroid Method: It is the method that extracts the largest sum of absolute loadings for each factor in turn.

It is defined by linear combinations in which all weights are either $+1.0$ or -1.0.

The purpose of this method is to maximize the sum of loadings, disregarding signs.

Example-Suppose the average student's aptitude in the field of astronomy is

$\{10 \times \text{the student's verbal intelligence}\} + \{6 \times \text{the student's mathematical intelligence}\}$.

The numbers 10 and 6 are the factor loadings associated with verbal intelligence and mathematical ability in aptitude towards astronomy.

Steps involving in the Centroid Method:

1) Obtain the correlation matrix.

2) Obtain grand matrix sum, row sum, and column sum.

3) Calculate

$$N = \frac{1}{\sqrt{Grand\ Total}}$$

4) Multiply each column sum with N, which gives the first-factor loading.

5) To find the second-factor loading, find the cross-product matrix of the factor 1 by testing the first factor loading horizontally and vertically and then multiplying corresponding rows and columns.

6) Find the first-factor residual matrix is given as,

Residual Matrix=Correlation Matrix-Cross Product Matrix

7) **Reflection:** Reflection means that each test vector retains its length but extends in opposite directions. The major purpose of reflection is to get a reflected cost matrix having the highest possible total. This step is taken due to the reason that some of the factors loading are with total. The method of reflection is by trial and error. This can be done by changing the signs of the variables from positive to negative and negative to positive column wise and row wise. The outcome we get is a reflective residual correlation matrix.

8) Repeat steps from (3) to (7).

Example 10.3.5(1): In the Centroid method of factor analysis, the correlation matrix is replaced by the factor matrix. A correlation matrix is square in nature where rows and columns are denoted by variables. Theoretically factor matrix can be square or rectangular, but practically, a factor matrix is mostly rectangular. The columns represent the factors and the rows represent the variables.

To understand the Centroid condensation method to extract factors, consider the correlation matrix (r_{ij}) given below:

Tests	1	2	3	4	5
1	(0.54)	0.50	0.23	0.39	0.28
2	0.50	(0.49)	0.31	0.47	0.37
3	0.23	0.31	(0.54)	0.60	0.39
4	0.39	0.47	0.60	(0.74)	0.59
5	0.28	0.37	0.39	0.59	(0.54)

STEP 1: Add values column-wise and row-wise and denote by symbol C and R, respectively to calculate the total

Tests	1	2	3	4	5	Row Total (R)
1	(0.54)	0.50	0.23	0.39	0.28	1.94
2	0.50	(0.49)	0.31	0.47	0.37	2.14
3	0.23	0.31	(0.54)	0.60	0.39	2.07
4	0.39	0.47	0.60	(0.74)	0.59	2.79
5	0.28	0.37	0.39	0.59	(0.54)	2.40
Column Total (C)	1.94	2.14	2.07	2.79	2.40	Grand Total (GT) =11.34

STEP 2: Find the value of N by the following formula,

$$N = \frac{1}{\sqrt{Grand\ Total}} = \frac{1}{\sqrt{11.34}} = \frac{1}{3.37} = 0.297$$

STEP 3: Multiply each column sum with N to get the first factor loadings L_i of each test i.e.

$$L_i = C \times N$$

First-factor loadings for

(i) Test 1 is $L_1 = 1.94 \times 0.297 = 0.58$

(ii) Test 2 is $L_2 = 2.14 \times 0.297 = 0.64$

(iii) Test 3 is $L_3 = 2.07 \times 0.297 = 0.62$

(iv) Test 4 is $L_4 = 2.79 \times 0.297 = 0.83$

(v) Test 5 is $L_5 = 2.40 \times 0.297 = 0.71$

STEP 4: To obtain the second-factor loadings, one must find the cross-product matrix. List all the first factor loadings on the horizontal and vertical sides of a table. Multiply corresponding rows and columns to obtain the cross-product matrix (L_{ij}).

L_i	0.58	0.64	0.62	0.83	0.71
0.58	0.34	0.37	0.36	0.48	0.41
0.64	0.37	0.41	0.40	0.53	0.45
0.62	0.36	0.40	0.38	0.52	0.44
0.83	0.48	0.53	0.52	0.69	0.60
0.71	0.41	0.45	0.44	0.60	0.50

STEP 5: First-factor Residual

To find the first factor residual subtract the first factor cross-product matrix (L_{ij}) from the original correlation matrix (r_{ij})

$Residual\ matrix\ =\ correlation\ matrix - cross\ product\ matrix$

Tests	1	2	3	4	5
1	0.20	0.13	−0.13	−0.09	−0.13
2	0.13	0.08	−0.09	−0.06	−0.08

3	−0.13	−0.09	0.16	0.08	−0.05
4	−0.09	−0.06	0.08	0.05	−0.01
5	−0.13	−0.08	−0.05	−0.01	0.27

STEP 6: Add values column-wise and row-wise and denote by symbol C and R, respectively to calculate the total.

Tests	1	2	3	4	5	Row Total (R)
1	0.20	0.13	−0.13	−0.09	−0.13	0.02
2	0.13	0.08	−0.09	−0.06	−0.08	−0.02
3	−0.13	−0.09	0.16	0.08	−0.05	−0.03
4	−0.09	−0.06	0.08	0.05	−0.01	−0.03
5	−0.13	−0.08	−0.05	−0.01	0.27	0.00
Column Total (C)	0.02	−0.02	−0.03	−0.03	0.00	GT=−0.10

STEP 7: Reflection In the above matrix, the total is negative indicating the need of reflection i.e. the test vectors need to be extended in the opposite direction but maintaining their length. For each factor to account for as much variance as possible, the reflection of vectors is done so as to maximize the GT. The method of reflection works best by inspecting the factor loadings in the residual matrix. The variables with the most negative factor loadings are reflected by changing the sign from positive to negative and vice-versa both row- wise and column-wise. Hence, a reflected residual matrix is obtained. Since the variables 3, 4 and 5 in the above residual matrix needs reflection, therefore, the reflected residual matrix is:

Tests	1	2	3	4	5	Row Total (R)
1	0.20	0.13	0.13	0.09	0.13	0.68

2	0.13	0.08	0.09	0.06	0.08	0.44
3	0.13	0.09	0.16	0.08	−0.05	0.41
4	0.09	0.06	0.08	0.05	−0.01	0.27
5	0.13	0.08	−0.05	−0.01	0.27	0.42
Column Total (C)	0.68	0.44	0.41	0.27	0.42	GT=2.22

STEP 8: Second Factor Loadings

The second factor loadings are calculated similarly as for the first factor loadings i.e. Step 1 to Step 3.

Here

$$N = \frac{1}{\sqrt{Grand\ Total}} = \frac{1}{\sqrt{2.22}} = \frac{1}{1.49} = 0.67$$

Second-factor loadings for

(i) Test 1 is $L_1 = 0.68 \times 0.67 = 0.46$

(ii) Test 2 is $L_2 = 0.44 \times 0.67 = 0.30$

(iii) Test 3 is $L_3 = 0.41 \times 0.67 = 0.28$

(iv) Test 4 is $L_4 = 0.27 \times 0.67 = 0.18$

(v) Test 5 is $L_5 = 0.42 \times 0.67 = 0.28$

Once the second-factor loadings are obtained, the variables that were reflected are reflected to their original signs. Hence, the second-factor loadings are $0.46, 0.30, -0.28, -0.18$ and -0.28.

A similar method could be continued with the residual matrix and the reflected residual matrix for $3^{rd}, 4^{th}, \ldots, n^{th}$ factor extraction.

Hence, the obtained factor matrix is:

Tests	Factor I	Factor II
1	0.58	0.46
2	0.64	0.30
3	0.62	−0.28
4	0.83	−0.18

10.3.6 Methods for Finding the Number of Factors to be Extracted

In theory, the maximum number of factors that can be extracted from a set of correlation coefficients is equal to the number of variables/ tests involved. For example, for a 10×10 correlation matrix, the maximum number of factors that can be extracted are 10. However, in factor analysis, these latent factors are constructed in such a way that they account for as much variance in the observed variables as possible. Hence, it becomes vital to decide the number of factors that can be extracted for a specific research problem. The following methods are:

1) Thumb Rule
2) Eigen Value Index
3) Fruckter Formula
4) Residual correlation matrix
5) Scree Plot

1) Thumb Rule: All the interrelated factors must explain at least as much as variances as an average variable. Check, if a variable is under a factor, then the percentage of variable explaining variance should be less than the percentage of factor explaining.

2) Eigen Value Index: The number of factors that have to be extracted can be determined by calculating the Eigen Value Index for each factor till an Eigen value of 1 is obtained i.e. those factors are to be extracted whose Eigen values are either 1 or more than 1. Factors with Eigen value less than 1 are not considered for contributing to the variance and hence are not given importance.

Example 10.3.6(1): For the following factor matrix, determine the number of factors that can be extracted on the basis of the Eigen Value of the factors.

	Factor I	Factor II	Factor III	Factor IV
Variable I	0.81	0.64	0.64	0.01
Variable II	0.80	0.69	0.39	0.06
Variable III	0.92	0.57	0.17	0.11
Variable IV	0.79	0.04	0.13	0.12
Variable V	0.17	0.72	0.11	0.16
Variable VI	0.12	0.11	0.05	0.31
Variable VII	0.81	0.23	0.04	0.49

Solution:

Eigen value of Factor I

$$= (0.81)^2 + (0.80)^2 + (0.92)^2 + (0.79)^2 + (0.17)^2 + (0.12)^2 + (0.81)^2 = 3.466$$

Eigen Value of Factor II

$$= (0.64)^2 + (0.69)^2 + (0.57)^2 + (0.04)^2 + (0.72)^2 + (0.11)^2 + (0.23)^2 = 1.80$$

Eigen value of Factor III

$$= (0.64)^2 + (0.39)^2 + (0.17)^2 + (0.13)^2 + (0.11)^2 + (0.05)^2 + (0.04)^2 = 0.70$$

Eigen value of Factor IV

$$= (0.01)^2 + (0.06)^2 + (0.11)^2 + (0.12)^2 + (0.16)^2 + (0.31)^2 + (0.49)^2 = 0.39$$

Since, only those factors are accounted for whose Eigen value is 1 or more than 1, therefore, only factor I and factor II will be extracted. Hence, only two factors are to be extracted.

3) Fruckter Formula: Fruckter proposed the following formula to decipher the number of factors that can be extracted for a research problem:

$$Number\ of\ factors = \frac{(2n+1) - \left(\sqrt{8n+1}\right)}{2}$$

Here n is the number of variables in the correlation matrix.

Example 10.3.6(2): Identify the number of factors that can be extracted in a research problem with 15 variables.

Solution:

$$Number\ of\ factors = \frac{(2n+1) - \left(\sqrt{8n+1}\right)}{2} = \frac{(2 \times 15 + 1) - \left(\sqrt{8 \times 15 + 1}\right)}{2} = 10.525$$

Hence, in a research problem with 15 variables, 11 (rounding up) factors are important to be extracted.

4) Residual Correlation Matrix Method: In this method, it is observed that when most of the correlation coefficients i.e. more than fifty percent of the correlation coefficients in the residual correlation matrix are zero or approaching zero, then one should stop further extraction of factors.

Example 10.3.6(3): Factor extraction will be continued in the following kind of residual correlation matrix

	1	2	3	4
1	0.40	−0.22	−0.23	−0.10
2	−0.22	0.46	−0.26	−0.11
3	−0.23	−0.26	0.69	−0.09
4	−0.10	−0.11	−0.09	0.95

In the following residual correlation matrix, further factor extraction will be stopped as most of the correlation coefficients are either zero or approaching zero.

	1	2	3	4
1	0.00	0.01	−0.02	0.00
2	0.01	0.00	0.00	−0.01
3	−0.02	0.00	0.00	0.01
4	0.00	−0.01	0.01	0.00

5) Scree Plot: A scree plot is a graphical method used to determine the number of factors to retain in a factor analysis. It is named after the shape of the graph, which resembles the steep slope of a scree (rock debris) on a mountain

To create a scree plot, you plot the eigenvalues of the factors in descending order on the y-axis, and the factor numbers on the x-axis. The resulting graph typically shows a steep drop in eigenvalues for the first few factors, followed by a more gradual decline. The point where the slope changes from steep to shallow is known as the elbow point.

The scree plot can help to identify the number of factors to retain in factor analysis, as the elbow point represents the point where the added explanatory power of additional factors diminishes. Factors before the elbow point are considered significant, while those after the elbow point are considered to be of less importance.

10.3.7 Applications of Factor Analysis

1. Multiple Regressions: Using factor scores in place of independent variables in a multiple regression estimation overcomes the problem of multicollinearity.

2. Simplifying the Discrimination Solution: If the discriminant model involves a large number of independent variables, these variables can be replaced by a set of manageable factors before estimation.

3. Simplifying the Cluster Analysis Solution: To make the data manageable, the variables selected for clustering can be reduced to a smaller number using factor analysis, and obtained factor scores can be used to cluster the objects/cases under study.

4. Perceptual Mapping in Multidimensional Scaling: Factors can be used as dimensions with the factor scores as the coordinates to develop attribute-based perceptual maps where one can comprehend the placement of brands or products according to the identified factors under study.

10.3.8 Terminology of Factor Analysis

Factor Loading: The correlation between the factor/component and independent variable is known as factor loading.

The observed variables are thought to be influenced by one or more underlying factors that are not directly observable. Factor loading represents the strength of the relationship between each observed variable and the latent factor.

Factor loading values range from -1.0 to +1.0. The positive values indicate a positive relationship between the observed variable and the latent factor, and the negative values indicate a negative relationship. The closer the factor loading value is to +1.0 or -1.0, the stronger the relationship between the observed variable and the latent factor. Factor loadings close to zero indicate that the observed variable is not strongly related to the latent factor.

Factor loadings provide information about the relative importance of each observed variable in measuring the underlying factor. Variables with high factor loadings on a particular factor are thought to be more closely related to that factor than variables with lower factor loadings. Factor loadings are also used to determine which observed variables should be included in a final factor solution, and to evaluate the reliability and validity of the factor solution.

Eigen Values: The eigenvalue (or characteristics root or latent root) of a factor is obtained by summing the squares of all the factor loadings in that factor. It indicates the amount of variance of the independent variables explained by the factor. Using the magnitude of the eigenvalue, a decision about retaining the factor in the model is made. A higher eigenvalue magnitude indicates more usefulness of the factor in explaining the group characteristics. Eigenvalue indicates the relative importance of each factor in accounting for the particular set of variables being analyzed.

Communality: It indicates the proportion of variance in responses to the statement which is explained by the identified factors.

$$Percentage\ of\ Variance = \frac{eigen\ value\ of\ the\ factor}{sum\ of\ all\ eigen\ values} \times 100$$

If any variable is not combined in any of the groups, then it can be left or can be considered as another factor. To remove a different variable, use rotation, i.e., we are changing its direction.

10.3.9 Advantages of Factor Analysis

1. Both objective and subjective attributes can be used.

2. It can be used to identify the hidden dimensions or constraints, which may or may not be apparent from direct analysis.

3. It is not extremely difficult to do and at the same time it's inexpensive and gives accurate results.

4. There is flexibility in naming and using dimensions.

10.3.10 Disadvantages of Factor Analysis

1. The usefulness depends on the researcher's ability to develop a complete and accurate set of product attributes. If important attributes are missed, the value of the procedure is reduced accordingly.

2. Naming of the factors can be difficult. Multiple attributes can be highly correlated for no apparent reason.

3. If the observed variables are completely unrelated, the factor analysis is unable to produce a meaningful pattern.

4. It is not possible to know what factors represent, only theory can help the user on this.

10.4 Linear Factor Model

General Factor Analysis Problem:

In general, we assume that a $p-$ dimensional random vector X is generated by

$X = f(S) + e$

where $S = (S_1, \dots, S_m)'$is $m \times 1$ vector of unobservable sources of independent latent variables. Here

S_j is the j^{th} latent variable with mean 0.

f: $\mathbb{R}^m \to \mathbb{R}^p$is an unknown mixing function

$e_{(p\times 1)}$ is the measurement noise

We assume that $E(S) = 0, \quad Cov(S) = \mathrm{I}_m$ but the distribution of S is unknown. The problem is to invert f and estimate latent variables S.

Centring and Sphering of Observations:

Consider the random vector

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$$

With $E(X) = \mu, \quad Cov(X) = \Sigma_{XX}$.

Before applying factor analysis, we apply centering so that the mean becomes zero and sphering (whitening) so that components are uncorrelated and have variance 1.

Let U be an orthogonal matrix of eigenvectors of $\Sigma_{XX}(UU' = I, U'U = I)$. Further Λ is a diagonal matrix of eigenvalues of Σ_{XX}. Columns of U and diagonal elements of Λ are ordered by decreasing magnitude of eigenvalues of Σ_{XX}. Then, we have

$$\Sigma_{XX} = U\Lambda U'$$

The centered and sphere version of X is

$$\Lambda^{-\frac{1}{2}}\mathrm{U}'(\mathrm{X} - \mu)$$

In practice μ and Σ_{XX} are unknown. Suppose $X_1, \dots, X_n$ are n observation vectors with

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$\hat{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})' = \hat{U}\hat{\Lambda}\hat{U}' .$$

Then centering and sphering of data is done by using the transformation

$$X \leftarrow \hat{\Lambda}^{-\frac{1}{2}}\hat{\mathrm{U}}'(\mathrm{X_i} - \overline{\mathrm{X}}) \quad , \quad \mathrm{i} = 1, \dots, \mathrm{n}.$$

Linear Factor Model:

The linear factor model is defined as

$$X = AS + e$$

Here S and e are uncorrelated, S has mean 0 and covariance matrix I_m, e has mean 0 and covariance matrix

$$\psi = diag\,(\psi_1, \dots, \psi_p)$$

Input variable $X = (X_1, \dots, X_p)'$ has been standardized to have zero mean and unit variance. Then

$X_j = a_{1j}S_1 + \cdots + a_{mj}S_m + e_j \qquad (j = 1, \dots, p)$

Here $S_1, \dots, S_m$ are the latent variables or common factors and $a_{1j}, \dots, a_{mj}$ are the factor loadings. Further, $e_j's$ are called specific (or unique) factors. If $S_j's$ are uncorrelated, they are called orthogonal factors otherwise oblique factor.

$\Sigma_{XX} = AA' + \psi$

The exploratory factor analysis problem is to estimate A and recover S.

10.4.1 Assumptions of Factor Analysis Model

1) Measurement error has mean zero and constant variance, i.e.,

$E(e_i) = 0$ and $Var\ (e_i) = \sigma_i^2$

2) No association between the factor and measurement error, $Cov(F, e_j) = 0$

3) No association between errors, $Cov(e_j, e_k) = 0$

4) **Local (i.e., conditional independence):** Given factor, observed variables are independent of one another, $Cov(X_j, X_k | F) = 0$

10.5 Estimation of Factor Loadings

Consider the linear factor model

$X = AS + e$

Let

$B = (A'A)^{-1}A'.$

Then

$ABX = AS + ABe$

$= (X - e) + ABe$

or $X = ABX + (I - AB)e$

$X = CX + E$ is the reduced rank model. Here $C = AB,\ E = (I - C)e,$ and C has rank m. From standardized input data (centering and sphering have been already done for data standardization) $\tilde{X}_1, \ldots, \tilde{X}_n$, we find estimate

$$\hat{\Sigma}_{XX} = \frac{1}{n}\sum_{i=1}^{n}\left(\tilde{X}_i - \bar{\tilde{X}}\right)\left(\tilde{X}_i - \bar{\tilde{X}}\right)'$$

Let $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_m$ be m ordered eigenvalues of $\hat{\Sigma}_{XX}$ and $\hat{v}_1 \ldots \hat{v}_m$ are the corresponding eigenvectors. Then, estimates of A and B are given by

$$\hat{A} = (\hat{v}_1, \ldots, \hat{v}_m) = \hat{B}'$$

The m vector of estimated factor scores are

$$\hat{f} = \hat{B}X = (\hat{v}_1'X, \ldots, \hat{v}_m'X)'.$$

10.6 Factor Rotation

The unrotated output maximizes variance accounted for by the first and subsequent factors and forces the factors to be orthogonal. This data compression comes at the cost of having most items load on the early factors, and usually, of having many items load substantially on more than one factor. Rotation serves to make the output more understandable, by seeking the so-called "Simple Structure" which is a pattern of loadings where each item loads strongly on only one of the factors, and much more weakly on the other factors. It is of two types:

1. Orthogonal rotation
2. Oblique rotation

1. ORTHOGONAL ROTATION

It is a transformational system used in factor analysis in which the different underlying or latent variables are required to remain separated from or uncorrelated with one another. Three different methods can be used for orthogonal rotation:

1. Varimax Rotation: It is an orthogonal rotation of the factor axes to maximize the variance of the squared loadings of a factor (column) on all the variables (rows) in a factor matrix, which has the effect of differentiating the original variables by extracted factor. A varimax solution yields

results that make it as easy as possible to identify each variable with a single factor. This is the most common and most frequently used rotation method.

2. Quartimax Rotation: It is an orthogonal alternative that minimizes the number of factors needed to explain each variable. This type of rotation often generates a general factor on which most variables are loaded to a high or medium degree.

3. Equimax Rotation: It is a compromise between Varimax and Quartimax criteria.

2. OBLIQUE ROTATION

It is a transformational system used in factor analysis when two or more factors (i.e., latent variables) are correlated. Oblique rotation reorients the factors so that they fall closer to clusters of vectors representing manifest variables, thereby simplifying the mathematical description of the manifest variables. There are two methods used for the oblique rotation:

1. Direct oblimin rotation
2. Promax Rotation

Note: The Promax method is like the Direct Oblimin method but is computationally faster than it.

10.7 Estimation of Factor Scores

Factors scores are measures of principal components or common factors. Under the principal components model, the factor scores are uniquely determined; under the common factor model, they are not. In the latter situation, the factor scores are indeterminate; potentially having an infinite number of solution sets, and thus their true values can only be estimated. Three methods of factor score estimation are:

(a) Regression method

(b) Ordinary Least Squares

(c) Weighted Least Squares

10.8 Summary

In this lesson, we learned about:

- The interpretation of factor loadings.
- The principal component and maximum likelihood methods for estimating factor loadings and specific variances.
- The Centroid condensation method of factor extraction involves the calculation of the grand-total by adding the correlation coefficients of each column. This total is then used to calculate N which is multiplied with each column sum to obtain the first factor loadings of each test. Following this, a cross-product matrix is obtained that is subtracted from the correlation matrix to obtain the residual correlation matrix.
- The residual correlation matrix may have to be reflected to maximize the total of the matrix. This is done by changing the signs of the factor loadings in the residual matrix from positive to negative or vice-versa for the variables both row-wise as well as column-wise. The second factor loadings are then calculated from the reflected residual factor loadings and they are reflected to their original signs. Further factor loadings may be obtained similarly.
- The decision about the number of common factors to extract and retain must steer between the extremes of losing too much information about the original variables on one hand and being left with too many factors on the other. Various criteria have been suggested to understand the number of factors that can be extracted. These are the Fruckter Formula method, the Eigenvalue index method, and the residual correlation matrix method.
- How commonalities can be used to assess the adequacy of a factor model.
- A likelihood ratio test for the goodness-of-fit of a factor model.
- The methods for estimating common factors.

10.9 Self-Assessment Exercises

1. What is meant by a factor in the context of factor analysis?

2. How many types of Factor analysis are there? Explain.

3. Sketch the three basic matrices involved in the factor analysis procedure: Input data matrix, Correlation matrix, and Factor matrix.

4. Discuss the meaning of factor loading. What is its maximum and minimum value?

5. What is accomplished by rotating a factor-loading matrix?

10.10 References

- Brenner, D., Bilodeau, M.: Theory of Multivariate Statistics. Germany: Springer.
- Giri Narayan, C.: Multivariate Statistical Analysis.
- Khatri, C. G.: Multivariate Analysis.
- Mardia, K. V.: Multivariate Analysis.
- Rencher, A.C.: Methods of Multivariate Analysis, Second edition, Wiley.
- Seber, G.A.F.:*Multivariate Observations.* Wiley, New York.

10.11 Further Readings

- Anderson, T.W. & Rubin, H. (1956): "Statistical inference in factor analysis" Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability **5**, 111–150.
- Bartlett, M.S. (1937): "The statistical conception of mental factors" British Journal of Psychology **28**, 97–104.
- Cattell, R.B. (1978): "The Scientific Use of Factor Analysis in the Behavioral and Life Sciences" Plenum Press, New York.
- Gorsuch, R.L. (1983): "Factor Analysis" 2^{nd} Edition, Lawrence Erlbaum Associates, Hillsdale.
- Harman, H.H. (1976): "Modern Factor Analysis" 3^{rd} Edition revised, The University of Chicago Press, Chicago.
- Lawley, D.N. & Maxwell, A.E. (1971): "Factor Analysis as a Statistical Method" American Elsevier Publishing, New York.
- McDonald, R.P. (1985): "Factor Analysis and Related Methods" Lawrence Erlbaum Associates, Hillsdale.
- Mulaik, S.A. (1972): "The Foundations of Factor Analysis" McGraw-Hill, New York.
- Thomson, G.H. (1939): "The Factorial Analysis of Human Ability" Houghton Mifflin, Boston.

- Yates, A. (1987): “Multivariate Exploratory Data Analysis” A Perspective on Exploratory Factor Analysis, State University of New York Press, Albany.

UNIT - 11: TESTS OF HYPOTHESIS

Structure

11.1 Introduction

In multivariate statistical analysis, understanding the structure of relationships among multiple variables is often of paramount importance. One key aspect of this structure is captured by the covariance matrix, which provides a comprehensive summary of the variances and covariances between pairs of variables. Testing hypotheses about the equality of covariance matrices is a fundamental problem in this context, as it has broad applications in various fields, including finance, biology, and social sciences.

This unit explores the problems and methods associated with testing the equality of covariance matrices. In many cases, these problems are multivariate generalizations of simpler univariate problems, extending concepts that are well understood in single-variable contexts to the

more complex scenarios involving multiple variables. Among the various tests employed, likelihood ratio tests (LRTs) and their modifications play a significant role due to their powerful statistical properties.

The consideration of invariance, which refers to properties of statistical tests that remain unchanged under certain transformations, also leads to the development of alternative test procedures. These procedures are often designed to be robust to specific assumptions or to provide more powerful tests under specific conditions.

Initially, we will focus on testing the equality of multiple covariance matrices without assuming a specific form for the common covariance matrix. This analysis is closely related to the multivariate analysis of variance (MANOVA) in scenarios involving random factors. Subsequently, we will explore the problem of testing whether a covariance matrix is equal to a specified matrix, and, further, the simultaneous testing of the equality of a covariance matrix to a given matrix and the equality of a mean vector to a specified vector.

This exploration provides a comprehensive framework for understanding and applying hypothesis tests in the context of covariance matrices, equipping researchers and practitioners with the tools needed to address complex multivariate problems.

11.2 Objectives

After reading this unit, you should be able to: use the tests for the following situations:

1. Equality of covariance matrices,
2. Sphericity for covariance matrix
3. Testing $H_0: \Sigma = \sigma^2[(1-\rho)I + \rho J]$
4. Multivariate Tests of Equality of Several Covariance Matrices
5. Mean vector and covariance matrix are equal to given vector and matrix

11.3 Tests of Hypothesis

There is always some contention about the values of a parameter or the relationship between parameters. When parametric values are unknown, we estimate them through sample values. If the sample value is exactly the same as per our contention, there is no hitch in accepting it. And if it is far from our contention, there is no reason to accept it. But the problem arises when

the sample provides a value which is neither exactly equal to the parametric value, nor too far. In that situation one has to develop some procedures which enables one to decide whether to accept a contended(hypothetical) value or not on the basis to sample values. Such procedure is known as Testing of hypothesis.

11.4 Tests of Hypothesis of Equality of Covariance Matrices

Test the hypothesis $\mathbf{H_0}: \boldsymbol{\Sigma} = \boldsymbol{\Sigma_0}$ against $\mathbf{H_1}: \boldsymbol{\Sigma} \neq \boldsymbol{\Sigma_0}$: Let $\mathrm{X} = \left(\mathrm{X}_1, \ldots, \mathrm{X}_\mathrm{p}\right)'$ be a random vector from the p-variate normal distribution $N_p(\mu, \Sigma)$. Consider the hypothesis $H_0: \Sigma = \Sigma_0$ against $H_1: \Sigma \neq \Sigma_0$. For testing H_0, we obtain a random sample of n observation vectors $X_1, X_2, \ldots, X_n$. Let

$$S = \frac{1}{\nu}\sum_{j=1}^{n}(X_j - \bar{X})(X_j - \bar{X})',$$

where $\nu = n - p$ represents the degrees of freedom.

For observing if S is significantly different from Σ_0, we use the following test statistic:

$$u = \nu[\ln|\Sigma_0| - \ln|S| + tr(S\Sigma_0^{-1}) - p] \tag{11.1}$$

For a single sample of size n, the degrees of freedom are $\nu = n - 1$. Notice that for $S = \Sigma_0$, we obtain $u = 0$ and u increases with the distance between S and Σ_0. For large ν, under H_0, the statistic u is approximately distributed as $\chi^2\left(\frac{1}{2}p(p+1)\right)$. Notice that the degrees of freedom $\frac{1}{2}p(p+1)$ of χ^2-distribution is the number of distinct parameters in Σ.

For moderate ν,

$$u' = \left[1 - \frac{1}{6\nu - 1}\left(2p + 1 - \frac{2}{p+1}\right)\right]u$$

is a better approximation to the $\chi^2\left(\frac{1}{2}p(p+1)\right)$ distribution.

We reject H_0 at 100 α% level of significance if u *or* u' is greater than $\chi^2\left(\alpha; \frac{1}{2}p(p+1)\right)$.

We can express u in terms of the eigenvalues $\lambda_1, \ldots, \lambda_p$of $S\Sigma_0^{-1}$. We have

$$tr(S\Sigma_0^{-1}) = \sum_{i=1}^{p} \lambda_i$$

$$ln|\Sigma_0| - ln|S|$$
$$= -ln|\Sigma_0^{-1}| - ln|S|$$
$$= -ln|\Sigma_0^{-1}S|$$
$$= -ln\left(\prod_{i=1}^{p} \lambda_i\right)$$

Thus

$$u = v\left[-ln\left(\prod_{i=1}^{p} \lambda_i\right) + \sum_{i=1}^{p} \lambda_i - p\right]$$
$$= v\left[\sum_{i=1}^{p} (\lambda_i - ln\lambda_i) - p\right]$$

For testing the hypothesis that the variables are independent and have unit variance, i.e., $H_0: \Sigma = I_p$, we simply set $\Sigma_0 = I_p$.

11.5 Sphericity Tests for Covariance Matrix

Suppose we are interested in testing the hypothesis that the individual variables of $X = (X_1, .., X_p)'$ are independent and have common variance σ^2. The hypothesis can be expressed as $H_0: \Sigma = \sigma^2 I_p$ against $H_1: \Sigma \neq \sigma^2 I_p$, where σ^2 is the unknown common variance. Under H_0, the ellipsoid $(X - \mu)'\Sigma^{-1}(X - \mu) = c^2$ reduces to $(X - \mu)'(X - \mu) = \sigma^2 c^2$, which is the equation of a sphere. Thus, the covariance structure $\sigma^2 I_p$ is called spherical. Based on a random sample $X_1, \ldots, X_n$ of size n, the likelihood ratio test for testing $H_0: \Sigma = \sigma^2 I_p$ against $H_1: \Sigma \neq \sigma^2 I_p$ is

$$LR = \left[\frac{|S|^p}{\left(\frac{tr\,(S)}{p}\right)^p}\right]^{\frac{n}{2}}$$

We have, for large n

$-2ln(LR)$ is approximately χ^2_v.

The degrees of freedom v is the total number of parameters minus the number estimated under the restrictions imposed by H_0.

Here

$-2\ln(LR)$

$$= -nln\left[\frac{|S|^p}{\left(tr\left(\frac{S}{p}\right)\right)^p}\right]$$

$$= -nln(u)$$

where

$$u = (LR)^{\frac{2}{n}}$$

$$= \frac{p^p|S|^p}{(trS)^p}$$

$$= \frac{p^p \prod_{i=1}^{p} \lambda_i}{\left(\sum_{i=1}^{p} \lambda_i\right)^p}.$$

Here $\lambda_1, \ldots, \lambda_p$ are the eigen values of S.

An improved test statistic is given by

$$u' = -\left(v - \frac{(2p^2 + p + 2)}{6p}\right) ln\, u \qquad (11.2)$$

where ν is the degrees of freedom for S. The statistic u' has an approximate χ^2-distribution with $\frac{1}{2}p(p+1)-1$ degrees of freedom. We reject H_0 if $u' > \chi^2\left(\alpha, \frac{1}{2}p(p+1)-1\right)$. For obtaining the degrees of freedom in the χ^2-approximation, notice that the total number of distinct parameters under H_1 is $p(p+1)$ and under H_0 is 1. Thus $\nu = \frac{1}{2}p(p+1)-1$. We can easily verify that if all the eigen values $\lambda_i's$ are equal, say λ, then $u = 1$, and $u' = 0$. Hence, this statistic can be used to test the hypothesis of equality of the population eigenvalues.

Example 11.5.1: John reported the results of an experiment where subjects responded to "probe words" at five positions in a sentence. The variables are response times for the i^{th} probe word, X_i ; $i = 1,2, \ldots ,5$. The data are given in below table:

Subject Number	X_1	X_2	X_3	X_4	X_5
1	51	36	50	35	42
2	27	20	26	17	27
3	37	22	41	37	30
4	42	36	32	34	27
5	27	18	33	14	29
6	43	32	43	35	40
7	41	22	36	25	38
8	38	21	31	20	16
9	36	23	27	25	28
10	26	31	31	32	36
11	29	20	25	26	25

The hypothesis is $H_0: \mu_1 = \mu_2 = \cdots = \mu_5$.

First test $H_0: \Sigma = \sigma^2 I$. The sample mean $\bar{X}$ and sample covariance matrix S are

$$\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i}$$

$$= \frac{397}{11}$$

$$= 36.09$$

$$\bar{X} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \bar{X}_3 \\ \bar{X}_4 \\ \bar{X}_5 \end{bmatrix} = \begin{bmatrix} 36.09 \\ 25.55 \\ 34.09 \\ 27.27 \\ 30.73 \end{bmatrix}$$

$$s_{11} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{1i} - \bar{X}_1)(X_{1i} - \bar{X}_1)'$$

$$= \frac{1}{10}(650.90)$$

$$= 65.09$$

$$S = \begin{bmatrix} 65.09 & 33.65 & 47.59 & 36.77 & 25.43 \\ 33.65 & 46.07 & 28.95 & 40.34 & 28.36 \\ 47.59 & 28.95 & 60.69 & 37.37 & 41.13 \\ 36.77 & 40.34 & 37.37 & 62.82 & 31.68 \\ 25.43 & 28.36 & 41.13 & 31.68 & 58.22 \end{bmatrix}$$

The trace of a matrix S is the sum of its diagonal elements, so the trace S is

$$65.09 + 46.07 + 60.69 + 62.82 + 58.22 = 292.89$$

Determinant of S is 27230647.96

Then

$$u = \frac{p^p|S|}{(Trace\ S)^p}$$

$$= \frac{5^5 \times 27230647.96}{(292.89)^5}$$

$$= 0.0395$$

Using (11.1), we have

$$u' = -\left(v - \frac{(2p^2 + p + 2)}{6p}\right) ln\, u$$

$$= -\left(10 - \frac{(2(5)^2 + 5 + 2)}{6 \times 5}\right) ln(0.0395) = -(10 - 1.9) \times (-3.2315)$$

$$= 8.1 \times 3.2315 = 26.175$$

The approximate χ^2 −test has

$$\frac{1}{2}p(p+1) - 1 = \frac{1}{2}(5 \times 6) - 1$$
$$= \frac{30}{2} - 1$$
$$= 15 - 1 = 14$$

degrees of freedom. Therefore compare $u' = 26.175$ with $\chi^2_{0.05,14} = 23.68$ and reject $H_0: \Sigma = \sigma^2 I$. To test $H_0: C\Sigma C' = \sigma^2 I$, we use the following matrix of orthonormalized contrasts:

$$C = \begin{bmatrix} 4/\sqrt{20} & -1/\sqrt{20} & -1/\sqrt{20} & -1/\sqrt{20} & -1/\sqrt{20} \\ 0 & 3/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} & -1/\sqrt{12} \\ 0 & 0 & 2/\sqrt{6} & -1/\sqrt{6} & -1/\sqrt{6} \\ 0 & 0 & 0 & 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}$$

Then

$$u = \frac{(p-1)^{(p-1)}|CSC'|}{(Trace\ CSC')^{(p-1)}}$$
$$= \frac{4^4 \times 146204.65}{(94.03)^5}$$
$$= \frac{256 \times 146204.65}{78174613.8}$$
$$= 0.48$$

Hence

$$u' = -\left[\nu - \frac{\{2(p-1)^2 + (p-1) + 2\}}{6(p-1)}\right] ln\, u$$

$$= -\left[10 - \frac{\{2(4)^2 + 4 + 2\}}{6 \times 4}\right] ln(0.48) = -(10 - 1.58) \times (-0.73)$$

$$= 8.42 \times 0.73 = 6.15$$

For degrees of freedom, we now have $\frac{1}{2}(4)(5) - 1 = 9$, and the critical value is $\chi^2_{0.05,9} = 16.92$. Hence, we do not reject $H_0: C\Sigma C' = \sigma^2 I$, and a univariate F-test of $H_0: \mu_1 = \mu_2 = \cdots = \mu_5$ may be justified.

11.6 Testing $H_0: \Sigma = \sigma^2[(1-\rho)I + \rho J]$

The univariate ANOVA approach has been found to be appropriate under less stringent conditions than $\Sigma = \sigma^2 I$. Wilks (1946) showed that the ordinary F-tests of ANOVA remain valid for a covariance structure of the form

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix} \tag{11.3}$$

$$= \sigma^2[(1-\rho)I + \rho J] \tag{11.4}$$

Here J is a square matrix of $1's$, as defined in below:

$$J = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$

And ρ is the population correlation between two variables.

Consider the hypothesis

$$H_0: \Sigma = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \cdots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \cdots & \sigma^2\rho \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho & \sigma^2\rho & \cdots & \sigma^2 \end{bmatrix}$$

We obtain the sample covariance matrix S and also estimates of σ^2 and $\sigma^2\rho$ under H_0 are given by

$$s^2 = \frac{1}{p}\sum_{j=1}^{p} s_{jj}$$

And

$$s^2r = \frac{1}{p(p-1)}\sum_{j\neq k} s_{jk}$$

where s_{jj} and s_{jk} are the j^{th}diagonal and $(j,k)^{th}$ off-diagonal elements of S.

s^2 is an average of the variances on the diagonal of S, and s^2r is an average of the off-diagonal covariances in S. An estimate of ρ can be obtained as $r = \frac{s^2r}{s^2}$. The estimate of Σ under H_0 is then

$$\mathrm{S}_0 = s^2 \begin{bmatrix} s^2 & s^2r & \cdots & s^2r \\ s^2r & s^2 & \cdots & s^2r \\ \vdots & \vdots & \ddots & \vdots \\ s^2r & s^2r & \cdots & s^2 \end{bmatrix}$$
$$= s^2[(1-r)I + rJ]$$

To compare S and S_0, use the following function of the Likelihood Ratio

$$u = \frac{|S|}{|S_0|}$$

Alternative form is

$$u = \frac{|S|}{(s^2)^p(1-r)^{p-1}[1+(p-1)r]} \qquad (11.5)$$

Using (11.2), the test statistic is given by

$$u' = -\left[\nu - \frac{p(p+1)^2(2p-3)}{6(p-1)\,(p^2+p-4)}\right] ln\, u \qquad (11.6)$$

where ν is the degrees of freedom of S. The statistic u' is approximately $\chi^2\left[\frac{1}{2}p(p+1)-2\right]$, and we reject H_0 if $u' > \chi^2\left[\alpha, \frac{1}{2}p(p+1)-2\right]$.

Note that 2 degrees of freedom are lost due to estimation of σ^2 and ρ.

Alternative approximate test that is more precise when p is large and ν is relatively small is given by

$$F = \frac{-(\gamma_2 - \gamma_2 c_1 - \gamma_1)\nu}{\gamma_1 \gamma_2} ln\, u$$

where

$$c_1 = \frac{p(p+1)^2(2p-3)}{6\,\nu(p-1)(p^2+p-4)}$$

$$c_2 = \frac{p(p^2-1)(p+2)}{6\,\nu^2(p^2+p-4)}$$

$$\gamma_1 = \frac{1}{2}p(p+1) - 2$$

$$\gamma_2 = \frac{\gamma_1 + 2}{c_2 - c_1^2}$$

We reject the null hypothesis $H_0: \Sigma = \sigma^2[(1-\rho)I + \rho J]$ at α level of significance if $F > F_{(\alpha,\gamma_1,\gamma_2)}$.

Example 11.6.1.: Rao (1948) measured the weight of cork borings taken from the north (N), east (E), south (S), and west (W) directions of 28 trees. A comparison is made of average thickness, and hence weight, in the four directions. A standard ANOVA approach to these repeated measures design would be valid if (11.1) holds. To Test $H_0: \Sigma = \sigma^2[(1-\rho)I + \rho J]$ for the data of below table:

Tree	N	E	S	W
1	72	66	76	77
2	60	53	66	63
3	56	57	64	58
4	41	29	36	38
5	32	32	35	36
6	30	35	34	26
7	39	39	31	27
8	42	43	31	25
9	37	40	31	25
10	33	29	27	36
11	32	30	34	28
12	63	45	74	63
13	54	46	60	52
14	47	51	52	43
15	91	79	100	75
16	56	68	47	50
17	79	65	70	61
18	81	80	68	58
19	78	55	67	60
20	46	38	37	38
21	39	35	34	37
22	32	30	30	32
23	60	50	67	54
24	35	37	48	39
25	39	36	39	31
26	50	34	37	40
27	43	37	39	50
28	48	54	57	43

The sample mean $\bar{X}$ is

$$\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{1i}$$

$$= \frac{1415}{28} = 50.54$$

$$\bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i}$$

$$= \frac{1293}{28} = 46.18$$

$$\bar{X}_3 = \frac{1}{n}\sum_{i=1}^{n} X_{3i}$$

$$= \frac{1391}{28} = 49.68$$

$$\bar{X}_4 = \frac{1}{n}\sum_{i=1}^{n} X_{4i}$$

$$= \frac{1265}{28} = 45.18$$

And sample covariance matrix S is

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{13} & s_{14} \\ s_{21} & s_{22} & s_{23} & s_{24} \\ s_{31} & s_{32} & s_{33} & s_{34} \\ s_{41} & s_{42} & s_{43} & s_{44} \end{bmatrix}$$

$$s_{11} = \frac{1}{n-1}\sum_{i=1}^{n} (X_{i1} - \bar{X}_1)^2$$

$$= \frac{1}{27}(7840.96) = 290.41$$

$s_{21} = s_{12}$

$$= \frac{1}{n-1}\sum_{i=1}^{n}(X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)$$

$$= \frac{1}{27}(6041.32) = 223.75$$

Similarly,

$s_{13} = s_{31} = 288.44,$

$s_{41} = s_{14} = 226.27,$

$s_{22} = 219.93,$

$s_{33} = 350.00,$

$s_{44} = 226.00,$

$s_{43} = s_{34} = 259.54,$

$s_{42} = s_{24} = 171.37,$

$s_{32} = s_{23} = 229.06$

$$S = \begin{bmatrix} 290.41 & 223.75 & 288.44 & 226.27 \\ 223.75 & 219.93 & 229.06 & 171.37 \\ 288.44 & 229.06 & 350.00 & 259.54 \\ 226.27 & 171.37 & 259.54 & 226.00 \end{bmatrix}$$

The determinant of S is

$|S| = 25617563.28$

And

$$s^2 = \frac{1}{p}\sum_{j=1}^{p} s_{jj}$$

$$= \frac{1}{4}(290.41 + 219.93 + 350.00 + 226.00)$$

$$= \frac{1086.34}{4} = 271.59$$

$$s^2r = \frac{1}{p(p-1)} \sum_{j \neq k} s_{jk}$$

$$= \frac{1}{4 \times 3}(223.75 + 223.75 + 288.44 + 288.44 + 226.27 + 226.27 + 229.06 + 229.06$$

$$+ 171.37 + 171.37 + 259.54 + 259.54)$$

$$= \frac{2796.86}{12} = 233.072$$

$$r = \frac{s^2r}{s^2} = \frac{233.072}{271.59} = 0.858$$

Using (11.4) and (11.5), we have

$$u = \frac{|S|}{(s^2)^p(1-r)^{p-1}[1 + (p-1)r]}$$

$$= \frac{25{,}617{,}563.28}{(271.59)^4(1 - 0.858)^{4-1}[1 + (4-1)0.858]}$$

$$= \frac{25{,}617{,}563.28}{(5440704019)(0.003)(3.574)}$$

$$= \frac{25{,}617{,}563.28}{55676853.2}$$

$$= 0.461$$

$$u' = -\left[v - \frac{p(p+1)^2(2p-3)}{6(p-1)\,(p^2+p-4)}\right] ln\, u$$

$$= -\left[27 - \frac{4(4+1)^2(2\times 4-3)}{6(4-1)\,(4^2+4-4)}\right] ln\, 0.461$$

$$= -\left[27 - \frac{4(5)^2(5)}{6(3)\,(16)}\right](-0.774)$$

$$= (25.26)(0.774)$$

$$= 19.55$$

Since $\chi^2_{(0.05,8)} = 15.5$

Hence, $\chi^2_{cal} = 19.55 > 15.5 = \chi^2_{tab(0.05,8)}$, we don't accept the null hypothesis, i.e., Σ does not have pattern (11.3).

11.7 Multivariate Tests of Equality of Several Covariance Matrices

Suppose there are k multivariate populations each of dimension p, with covariance matrices $\Sigma_1, \Sigma_2, \ldots, \Sigma_k$. The hypothesis of equality of covariance matrices is

$H_0: \Sigma_1 = \Sigma_2 \ldots = \Sigma_k$

For $k = 2$, the test reduces to $H_0: \Sigma_1 = \Sigma_2$. Suppose we have independent samples of size $n_1, n_2, \ldots, n_k$ from multivariate normal distributions, $\nu_i = n_i - 1$ and covariance matrix of the i^{th} sample is S_i, which is an an unbiased estimator of Σ_i. Further

$$S_{pl} = \frac{\sum_{i=1}^{k} \nu_i S_i}{\sum_{i=1}^{k} \nu_i}$$

$$= \frac{E}{\nu_E}$$

Where S_{pl} is the pooled sample covariance matrix, $E = \sum_{i=1}^{k} \nu_i S_i$,

$$\nu_E = \sum_{i=1}^{k} \nu_i$$

$$= \sum_{i=1}^{k} n_i - k.$$

Then the test statistic is

$$M = \frac{|S_1|^{\frac{\nu_1}{2}} \dots |S_k|^{\frac{\nu_k}{2}}}{|S_{pl}|^{(\sum_{i=1}^{k} \nu_i)/2}} \tag{11.7}$$

Obviously $\nu_i > p \; \forall \; i$ otherwise $|S_i| = 0$ for some i leading to $M = 0$.

The statistic M is a modification of the likelihood ratio with $0 \leq M \leq 1$. Let us write

$$M = \left(\frac{S_1}{S_{pl}}\right)^{\frac{\nu_1}{2}} \dots \left(\frac{S_k}{S_{pl}}\right)^{\frac{\nu_k}{2}} \tag{11.8}$$

If $S_1 = S_2 = \cdots = S_k = S_{pl}$, then $M = 1$. Further, as the disparity among $S_1, S_2, \dots, S_k$ increases, M approaches to zero. Thus, the values near 1 favor H_0 while values near 0 leading to rejection of H_0. If we assume $\nu_1 = \nu_2 = \nu_3 = \nu$, then for the first set,

$$M_1 = \left\{\left(\frac{1}{3}\right)\left(\frac{2}{3}\right)\left(\frac{6}{3}\right)\right\}^{\frac{\nu}{2}}$$

$$= \{(0.33)(0.67)(2.00)\}^{\frac{\nu}{2}}$$

$$= \{0.44\}^{\frac{\nu}{2}}$$

For the Second set,

$$M_2 = \left\{\left(\frac{3}{3}\right)\left(\frac{2}{3}\right)\left(\frac{4}{3}\right)\right\}^{\frac{\nu}{2}}$$

$$= \{(1)(0.67)(1.33)\}^{\frac{\nu}{2}}$$

$$= \{0.89\}^{\frac{\nu}{2}}$$

In M_1, the smallest value, 0.33 reduces the product proportionally more than the largest value, 2, increase it.

Box M Test: Box has given χ^2 and F-approximations for the distribution of M. Both the approximations referred as Box's M-test. First, we consider the χ^2 approximation. Consider

$$c_1 = \left[\sum_{i=1}^{k} \frac{1}{\nu_i} - \frac{1}{\sum_{i=1}^{k} \nu_i}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right]$$

Then, approximately

$$u = -2(1 - c_1) lnM \sim \chi^2(\eta)$$

where

$$\eta = \frac{1}{2}(k-1)p(p+1)$$

Taking ln on both sides in (11.7), we have

$$ln\ M = \frac{1}{2}\sum_{i=1}^{k} \nu_i \ln\ |S_i| - \frac{1}{2}\left(\sum_{i=1}^{k} \nu_i\right) \ln\ |S_{pl}|$$

We reject H_0 if $u > \chi_\alpha^2(\eta)$, where $\chi_\alpha^2(\eta)$, is the upper $100\alpha\%$ point of the χ^2 distribution with η degrees of freedom. For $\nu_1 = \cdots = \nu_k = \nu$

$$c_1 = \left[\frac{k}{\nu} - \frac{1}{k\nu}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(k-1)}\right]$$

$$= \frac{(k+1)(2p^2 + 3p - 1)}{6k\nu(p+1)}$$

For the F-approximation, we define

$$c_2 = \frac{(p-1)(p+2)}{6(k-1)}\left[\sum_{i=1}^{k} \frac{1}{\nu_i^2} - \frac{1}{\left(\sum_{i=1}^{k} \nu_i\right)^2}\right]$$

$$a_1 = \frac{1}{2}(k-1)p(p+1),$$

$$a_2 = \frac{a_1 + 2}{|c_2 - c_1^2|},$$

$$b_1 = \frac{1 - c_1 - \left(\frac{a_1}{a_2^2}\right)}{a_1},$$

$$b_2 = \frac{\left(1 - c_1 + \left(\frac{2}{a_2}\right)\right)}{a_2}.$$

If $c_2 > c_1^2$, $F = -2b_1 \ln(M)$ is approximately $F(a_1, a_2)$.

If $c_2 < c_1^2$,

$$F = -\frac{2a_2 b_2 \ln(M)}{a_1(1 + 2b_2 \ln(M))}$$

is approximately $F(a_1, a_2)$. In both cases we reject H_0 if $F > F(\alpha; a_1, a_2)$. If $\nu_1 = \cdots = \nu_k = \nu$, we have

$$c_1 = \frac{(k+1)(2p^2 + 3p - 1)}{6k\nu(p+1)}$$

$$c_2 = \frac{(p-1)(p-2)(k^2 + k + 1)}{6k^2\nu^2}$$

11.8 Mean Vector and Covariance Matrix are Equal to Given Vector and Matrix

Lemma 11.8.1: Let Y be an observation vector on a random vector with density $f(y, \theta)$, where θ is a parameter vector in a space Ω. Let H_a be the hypothesis $\theta \in \Omega_a \subset \Omega$, let H_b be the hypothesis $\theta \in \Omega_b \subset \Omega_a$, and let H_{ab} be the hypothesis $\theta \in \Omega_b \subset \Omega$. If λ_a, the likelihood ratio criterion for testing H_a, λ_b, the likelihood ratio criterion for testing H_b and λ_{ab}, the likelihood ratio criterion for testing H_{ab} are uniquely defined for the observation vector Y, then $\lambda_{ab} = \lambda_a \lambda_b$.

Criteria: Suppose Y is a q-component random vector with mean vector $\xi Y = \nu$ covariance matrix is $\xi(Y-\nu)(Y-\nu)' = \Psi$, then

$(Y-\nu)'\Psi^{-1}(Y-\nu) = q+2$

This is called concentration ellipsoid of Y. If Ψ is known, then the statistic $(\bar{Y}-\nu_0)'\Psi^{-1}(\bar{Y}-\nu_0)$ is used for testing hypothesis $H_0: \nu = \nu_0$.

Combine the hypothesis $H_1: \Psi = \Psi_0$ and $H_2: \nu = \nu_0$ and test $H: \nu = \nu_0, \Psi = \Psi_0$

Where $\Psi_0$is positive definite matrix.

Given a sample $Y_1, Y_2, \dots, Y_N$ from $N(\nu, \Psi)$.

Let $X = C(Y-\nu_0)$

where

$C\Psi_0 C' = I$

Then $X_1, X_2, \dots, X_N$ constitute a sample from $N(\mu, \Sigma)$ and the hypothesis is $H: \mu = 0, \Sigma = I$.

The likelihood ratio criterion for $H: \mu = 0$ given $\Sigma = I$ is $\lambda_2 = exp\left(-\frac{1}{2}N\bar{X}'\bar{X}\right)$.

The likelihood ratio criterion for H is (by Lemma 11.8.1)

$\lambda = \lambda_1\lambda_2$

$$= \left(\frac{e}{N}\right)^{pN/2} |A|^{N/2} exp\left(-\frac{1}{2}tr\ A\right) exp\left(-\frac{1}{2}N\bar{X}'\bar{X}\right)$$

$$= \left(\frac{e}{N}\right)^{pN/2} |A|^{N/2} exp\left\{-\frac{1}{2}tr(A + N\bar{X}'\bar{X})\right\}$$

$$= \left(\frac{e}{N}\right)^{pN/2} |A|^{N/2} exp\left\{-\frac{1}{2}\sum X_\alpha' X_\alpha\right\}$$

The two factors λ_1 and λ_2 are independent because λ_1 is a function of A and λ_2 is a function of $\bar{X}$ and A and $\bar{X}$ are independent.

$$\xi\lambda_2^h = \xi \exp\left(\frac{1}{2}hN\sum \bar{X}_i^2\right)$$

$$= \xi \exp\left(\frac{1}{2}h\,\chi_p^2\right)$$

$$= (1+h)^{-p/2}$$

The h^{th} moment of λ is

$$\xi\lambda^h = \xi\lambda_1^h\xi\lambda_2^h$$

$$= \left(\frac{2e}{N}\right)^{pNh/2} \frac{1}{(1+h)^{pN(1+h)/2}} \frac{\Gamma_p\left\{\frac{1}{2}(n+Nh)\right\}}{\Gamma_p\left(\frac{n}{2}\right)}$$

Under the null hypothesis, then

$$-2\log\lambda = -2\log\lambda_1 - 2\log\lambda_2$$

An asymptotic expansion of the distribution of $-2\log\lambda$ is

$$P[-2\rho\log\lambda \le Z]$$

$$= P\left[\chi_f^2 \le Z\right] + \frac{\gamma_2}{\rho^2 N^2}\left[P\left(\chi_{f+4}^2 \le Z\right) - P\left(\chi_f^2 \le Z\right)\right] + O(N^{-3})$$

$$\rho = 1 - \frac{2p^2+9p-11}{6N(p+3)}$$

$$\gamma_2 = \frac{p(2p^4+18p^3+49p^2+36P-13)}{288(p-3)}$$

Let us define $X_\alpha = C(Y_\alpha - \nu_0), \alpha = 1, \ldots, N$, then

$$\sum X_\alpha' X_\alpha = \sum (Y_\alpha - \nu_0)'\, C'C(Y_\alpha - \nu_0)$$

$$= \sum (Y_\alpha - \nu_0)'\, \Psi_0^{-1}(Y_\alpha - \nu_0)$$

$$= tr\ A + N\bar{X}'\bar{X} \quad \left(A = \sum (X_\alpha - \bar{X})(X_\alpha - \bar{X})' \right)$$

$$= tr\ (B\Psi_0^{-1}) + N(\bar{Y} - \nu_0)'\Psi_0^{-1}(\bar{Y} - \nu_0)$$

where $|A| = |B\Psi_0^{-1}|$.

11.9 Summary

In this unit, we have covered the concepts of Testing of Hypothesis under following situations:

- We have test of equality of covariance matrices,
- We have discussed Sphericity tests for covariance matrix,
- We have explained $H_0: \Sigma = \sigma^2[(1-\rho)I + \rho J]$,
- We have discussed Multivariate Tests of Equality of Several Covariance Matrices,
- We have derived Mean vector and covariance matrix are equal to given vector and matrix.

11.10 Self-Assessment Exercises

1. Show that if $S = \Sigma_0$ in (11.1), then $u = 0$.
2. (i) Calculate M as given in (11.5) for

   $$S_1 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}, S_2 = \begin{bmatrix} 4 & 3 \\ 3 & 6 \end{bmatrix}$$

   Assume $\nu_1 = \nu_2 = 5$.

   (ii) Calculate M for

   $$S_1 = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}, S_2 = \begin{bmatrix} 10 & 15 \\ 15 & 30 \end{bmatrix}$$

   Assume $\nu_1 = \nu_2 = 5$.

   In (ii), S_1 and S_2 differ more than in (i) and M is accordingly much smaller.

3. Show that M in (11.7) can be expressed in the form given in (11.8).

4. Rao (1948) measured the weight of cork borings taken from the north (N), east (E), south (S), and west (W) directions of 28 trees. Test $H_0: \Sigma = \sigma^2 I$ and $H_0: C\Sigma C' = \sigma^2 I$ for the data in Example 11.6.1.
5. Test $H_0: \Sigma = \sigma^2 I$ and $H_0: C\Sigma C' = \sigma^2 I$ for the calculator speed data of below table.

Calculator Speed Data

Subjects	A_1		A_2	
	B_1	B_2	B_1	B_2
S_1	30	21	21	14
S_2	22	13	22	5
S_3	29	13	18	17
S_4	12	7	16	14
S_5	23	24	23	8

6. Test $H_0: \Sigma = \sigma^2[(1-\rho)I + \rho J]$ for the calculator speed data in above Table. Use both χ^2 and F approximations.

11.11 References

- Rencher, A.C.: Methods of Multivariate Analysis, Second edition, Wiley.
- Johnson, R. A., Wichern, D. W. (2019): Applied Multivariate Statistical Analysis. United Kingdom: Pearson.
- Härdle WK, Simar L (2015): Applied Multivariate Statistical Analysis. Springer-Verlag, Berlin.
- Muirhead, R. J. (2009): Aspects of Multivariate Statistical Theory. Germany: Wiley.
- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Giri Narayan C. (1995): Multivariate Statistical Analysis
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.

- Mardia, K. V., Bibby, J. M., Kent, J. T. (1979): Multivariate Analysis. United Kingdom: Academic Press.
- Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York.

11.12 Further Readings

- Kotz, S., Balakrishnan, N. and Johnson, N.L.: Continuous Multivariate Distribution Models and Applications (Second Edition). Volume 1, Wiley - Inter science, New York.
- Khatri, C. G.: Multivariate Analysis.
- Mardia, K. V.: Multivariate Analysis.
- Seber, G.A.F.: Multivariate Observations. Wiley, New York.
- Rencher, Alvin C.: Multivariate Statistical Inference and Applications. John Wiley. New York, New York.
- Brenner, D., Bilodeau, M.: Theory of Multivariate Statistics. Germany: Springer.

UNIT: 12 LINEAR REGRESSION MODEL

Structure

12.1 Introduction

In many fields of study, it is common to explore the relationship between one or more dependent (or response) variables and one or more independent (or predictor) variables. This is often achieved using linear models, which provide a mathematical framework for understanding how changes in the predictor variables influence the response variables.

In this unit, we will focus on constructing multivariate linear models, estimating their parameters, and determining which predictor variables to include when building the model. We will explore the different scenarios where one or multiple response and predictor variables are involved, and examine methods for selecting the most relevant variables. This is important to create models that are both interpretable and capable of making accurate predictions.

By the end of this unit, you will understand how to apply linear regression techniques, evaluate the significance of predictors, and refine models to ensure they are useful for real-world applications.

We can distinguish three cases according to the number of variables:

1. Simple linear regression: One y and one x. For example, suppose we wish to predict college grade point average (GPA) based on an applicant's high school GPA.

2. Multiple linear regression: One y and several $x's$. We could attempt to improve our prediction of college GPA by using more than one independent variable, for example, high school GPA, standardized test scores (such as ACT or SAT), or rating of high school.

3. Multivariate multiple linear regression: Several $y's$ and several $x's$. In the preceding illustration, we may wish to predict several $y's$ (such as number of years of college the person will complete or GPA in the sciences, arts, and humanities). As another example, suppose the Air Force wishes to predict several measures of pilot efficiency. These response variables could be regressed against independent variables (such as math and science skills, reaction time, eyesight acuity, and manual dexterity).

Multivariate Analysis of Variance (MANOVA) is a statistical technique used when we are interested in understanding the relationship between one or more independent variables and multiple dependent (outcome) variables. Unlike ANOVA, which focuses on a single outcome, MANOVA allows us to analyse several outcomes simultaneously, taking into account the potential interrelationships between them.

The principles of the linear model naturally extend to MANOVA, making it a versatile tool for examining complex data.

12.2 Objectives

After going through this unit, you will be able to:

- Multivariate linear regression model
- Estimation of parameters and their properties
- Multivariate analysis of variance [MANOVA] of one-way classified data
- Wilk's lambda criterion

12.3 Multivariate Linear Regression Model

Multivariate regression is a sophisticated technique used to determine the extent to which various independent variables are linearly related to multiple dependent variables. This linear relationship is established through the correlation between the variables. By applying multivariate regression to the dataset, researchers can then predict the behaviour of the response variable based on its corresponding predictor variables.

12.3.1 Characteristics of Multivariate Regression

- Multivariate regression allows one to have a different view of the relationship between various variables from all the possible angles.
- It helps to predict the behaviour of the response variables depending on how the predictor variables move.
- Multivariate regression can be applied to various machine learning fields, including economics, science, and medical research studies.

12.3.2 Example of Multivariate Regression

1. In a hypothetical scenario, a doctor has meticulously gathered data on individuals' blood pressure, weight, and red meat consumption to investigate the correlation between health

and dietary habits. This extensive dataset offers valuable insights into how choices such as red meat intake may impact physiological factors like blood pressure and weight.

2. An agricultural expert is determined to uncover the reasons behind the destruction of crops in a particular area. By examining recent weather patterns, water availability, irrigation methods, chemical usage, and other relevant factors, the expert aims to elucidate why the crops have been wilting and failing to produce fruit.

12.3.3 Advantages of Multivariate Regression

- The multivariate regression method helps to find a relationship between multiple variables or features.
- It also defines the correlation between independent variables and dependent variables.

12.3.4 Disadvantages of Multivariate Regression

- Multivariate regression technique requires high-level mathematical calculations.
- It is complex.
- The output of the multivariate regression model is difficult to analysis.
- The loss can use errors in the output.
- Multivariate regression yields better results when used with larger datasets rather than small ones.

12.4 Estimation of Parameters and their Properties

The multivariate regression model with p independent variables is

$$Y = XB + \epsilon \tag{12.1}$$

where

Y is $n \times p$ matrix of observations on dependent or response variable given by

$$Y = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

X is $n \times q$ matrix of observations on independent variables with q predictors and given by

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1q-1} & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q-1} & x_{2q} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nq-1} & x_{nq} \end{bmatrix}$$

B is $q \times p$ matrix of regression parameters given by

$$B = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1p} \\ \beta_{21} & \beta_{22} & \cdots & \beta_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \beta_{q1} & \beta_{q2} & \cdots & \beta_{qp} \end{bmatrix}$$

ε is $n \times p$ matrix of error term, defined as

$$\epsilon = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1p} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{np} \end{bmatrix}$$

12.4.1 Assumptions

(1) $E[Y] = XB$ or $E(\epsilon) = 0$

(2) $Cov(y_i) = \Sigma$ for all $i = 1,2, \ldots, n$

(3) $Cov(y_i, y_j) = 0$ for all $i \neq j$

The covariance matrix Σ in assumption (2) contains the variances and covariances of $y_{i1}, y_{i2}, \ldots, y_{ip}$ in any y_i, the i^{th} column of Y and given by

$$Cov\ (y_i) = \Sigma$$
$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

12.4.2 An Estimator for *B*

The least squares estimate of B minimize the sum of squares of deviations of the n observed $y's$ from their "modeled" values, that is, from their values $\hat{y}_i$ predicted by the model. We start with the equation corresponding to the i^{th} response variable:

$Y_{(i)} = Z\beta_{(i)} + \epsilon_{(i)}; i = 1, \dots, p.$

Here $Y_{(i)}$ is the i^{th} column of Y, $\beta_{(i)}$ is the i^{th} column of B, and $\epsilon_{(i)}$ is the i^{th} column of ϵ.

Thus, the least squares estimator of $\beta_{(i)}$, $\hat{\beta}_{(i)}$, is obtained by minimizing

$SSE_i = (Y_{(i)} - X\beta_{(i)})'(Y_{(i)} - X\beta_{(i)})$

The value of $\widehat{\beta_{(i)}}$ that minimizes SSE_i in above equation is given by

$\hat{\beta}_{(i)} = (X'X)^{-1}X'Y_{(i)}; \ \ i = 1, \dots, p.$

Then, the least squares estimator of B, denoted by $\hat{B}$ is

$$\begin{aligned}\hat{B} &= \left(\hat{\beta}_{(1)}, \dots, \hat{\beta}_{(p)}\right) \\ &= (X'X)^{-1}X'Y\end{aligned}$$

The predicted value of Y is

$$\begin{aligned}\hat{Y} &= X\hat{B} \\ &= X(X'X)^{-1}X'Y\end{aligned}$$

Further, the matrix of estimated residuals is

$$\begin{aligned}\hat{\epsilon} &= Y - \hat{Y} \\ &= Y - X\hat{B} \\ &= [I - X(X'X)^{-1}X']Y\end{aligned}$$

We observe that

$X'[I - X(X'X)^{-1}X'] = 0$

This implies that

$\hat{Y}'\hat{\epsilon}$

$= Y'X'(X'X)^{-1}X'\left[I-X(X'X)^{-1}X'\right]$

$= 0.$

12.4.3 Properties of Least Squares Estimators $\widehat{B}$

$\hat{B}$ has the following properties:

1. The estimator $\hat{B}$ is unbiased, that is, $E(\hat{B}) = B$. This means that if repeated random samples from the same population, the average value of $\hat{B}$ would be B.

2. The least squares estimators $\hat{\beta}_{jk}$ in $\hat{B}$ have minimum variance among all possible linear unbiased estimators. This result is known as the Gauss–Markov theorem. The restriction to unbiased estimators is necessary to exclude trivial estimators such as a constant, which has variance equal to zero, but is of no interest. This minimum variance property of least squares estimators is remark able for its distributional generality; normality of the $y's$ is not required.

3. All $\hat{\beta}_{jk}{}'s$ in $\hat{B}$ are correlated with each other. This is due to the correlations among the $x's$ and among the $y's$. The $\hat{\beta}'s$ within a given column of $\hat{B}$ are correlated because $x_1, x_2, \dots, x_q$ are correlated. If $x_1, x_2, \dots, x_q$ were orthogonal to each other, the $\hat{\beta}'s$ within each column of $\hat{B}$ would be uncorrelated. Thus, the relationship of the $x's$ to each other affects the relationship of the $\hat{\beta}'s$ within each column to each other. On the other hand, the $\hat{\beta}'s$ in each column are correlated with $\hat{\beta}'s$ in other columns because $y_1, y_2, \dots, y_p$ are correlated.

12.4.4 An Estimator for Σ

The matrix of error sum of squares and cross products matrices is

$\hat{\epsilon}'\,\hat{\epsilon} = \left(Y - X\hat{B}\right)'\left(Y - X\hat{B}\right).$

It can be shown that

$E(\hat{\epsilon}'\,\hat{\epsilon}) = (n-q)\Sigma$

An unbiased estimator of Σ is given by

$$S_e = \frac{\hat{\epsilon}'\,\hat{\epsilon}}{n-q}$$
$$= \frac{\left(Y - X\hat{B}\right)'\left(Y - X\hat{B}\right)}{(n-q-1)}$$
$$= \frac{Y'Y - \hat{B}'X'Y}{(n-q-1)}$$

Further

$$Y'Y = (\hat{Y} + \widehat{\epsilon})'(\hat{Y} + \widehat{\epsilon})$$
$$= \hat{Y}'\hat{Y} + \hat{\epsilon}'\,\hat{\epsilon} + 0 + 0$$

$Y'Y$: Total Sum of squares and cross products

$\hat{Y}'\hat{Y}$: predicted sum of squares and cross products

$\hat{\epsilon}'\,\hat{\epsilon}$: residual (error) sum of squares and cross products

12.5 Multivariate Analysis of Variance [MANOVA] of One-Way Classified Data

In the univariate case, the one-way ANOVA investigates the effects of a categorical variable (the classes or groups or treatments, i.e., independent variables) on a continuous outcome variable, i.e., the dependent variable. We have, m random variables $x_1, \dots, x_m$(groups or treatments). For j^{th} group sample is, say, $\{x_{1j}, x_{2j}, \dots, x_{nj}\}$. Group j is said to have $n_j$observations with $n = \sum_{j=1}^{m} n_j$.

Our objective is to test the null hypothesis of equality of means of all the groups $H_0: \mu_1 = \mu_2 = \cdots = \mu_m$.

We use the ANOVA for one-way classification and then the F-test statistic is

$$F_{Cal} = \frac{MSB}{MSW} \sim F(m-1, n-m)(under\ H_o).$$

Here MSB denotes the mean sum of squares between classes (or groups or treatments) and MSW denotes the mean sum of squares within classes (or mean error sum of squares).

We reject the null hypothesis at α level of significance if

$F_{Cal} > F_{crit}(\alpha; m-1, n-m)$.

Multivariate analysis of variance (MANOVA) considers the effects of a categorical variable (the groups, i.e., independent variables) on a vector of dependent variables. One of the options is to perform multiple ANOVA one for each dependent variable. However, the problems in performing multiple ANOVA, one for each dependent variable, are

(i) it would introduce additional experiment-wise error and

(ii) it would not consider the correlations between the dependent variables.

It is, therefore, possible that MANOVA shows a significant difference between the means while the individual ANOVA do not.

MANOVA can also be used when one has repeated measures. In this case, the repeated levels are taken as dependent variables.

12.5.1 Assumption

- **Observation Independence:** Each observation should be independent of one another. For example, one student's performance should not influence another's.
- **Multivariate Normality:** The combined dependent variables should be approximately normally distributed for each group of the independent variable.
- **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.
- **Linear Relationships:** There should be a linear relationship between each pair of dependent variables for each group of the independent variable.
- **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.

12.5.2 Notations

In One-way MANOVA, suppose we have m random vectors $X_1, \ldots, X_m$ (representing groups or treatments). Each X_j is a $k \times 1$ column vector of form

$$\begin{pmatrix} x_{j1} \\ \vdots \\ x_{jk} \end{pmatrix}$$

where each $x_{jp}, j = 1, \dots, m; p = 1, \dots, k$ is a random variable.

For each random vector X_j we collect a sample $\{X_{1j}, \dots, X_{n_j j}\}$ of size n_j with $n = \sum_{j=1}^{m} n_j = n$. Each X_{ij} is a $k \times 1$ vector

$$\begin{pmatrix} x_{ij1} \\ \vdots \\ x_{ijk} \end{pmatrix}; i = 1, \dots, n_j; j = 1, \dots, m; p = 1, \dots, k.$$

Here index i refers to the subject in the experiment, index j refers to the group and index p refer to the position (i.e., dependent variable) within the random vector.
Our objective is to test the null hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_m$, where

$$\mu_j = \begin{pmatrix} \mu_{j1} \\ \vdots \\ \mu_{jk} \end{pmatrix}, j = 1, \dots, m$$

Thus, the null hypothesis is equivalent to $H_0: \mu_{1p} = \mu_{2p} = \cdots = \mu_{mp} \; \forall \; p = 1, \dots, k$. The alternative hypothesis is $H_1: \mu_r \neq \mu_j$ for some $r \neq j, 1 \leq r, j \leq m$, or equivalently, $\mu_{rp} \neq \mu_{jp}$ for some $r \neq j$, and $p, 1 \leq r, j \leq m, 1 \leq p \leq k$.

We define various means as in the univariate case, except that now these means become $k \times 1$ vectors.

The total (or grand) mean vector is

$$\bar{X}_T = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_k \end{pmatrix},$$

$$\bar{x}_p = \frac{1}{n} \sum_{j=1}^{m} \sum_{i=1}^{n_j} x_{ijp}$$

The sample group mean vector for group j is

$$\bar{X}_j = \begin{pmatrix} \bar{x}_{j1} \\ \vdots \\ \bar{x}_{jk} \end{pmatrix},$$

$$\bar{x}_{jp} = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ijp}$$

We define the following total cross products:

$$CP_{pq} = \begin{cases} \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ijp} - \bar{X}_p\right)\left(x_{ijq} - \bar{X}_q\right) \;\; if p \neq q \\ \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ijp} - \bar{X}_p\right)^2 \text{ if } p = q \end{cases}$$

For $p = q, CP_{pp}$ is the total sum of squares and measures the total variation in the p^{th} dependent variable.

For $p \neq q, C_{pq}$, is the total cross-product term, which measure the dependence between the p^{th} and q^{th} variables across all observations.

The multivariate equivalent of the total sum of squares is the matrix of total sum of squares and cross products, say *T*, and is defined as

$$T = \begin{pmatrix} SS_{11} & \cdots & SS_{1k} \\ \vdots & \ddots & \vdots \\ SS_{k1} & \cdots & SS_{kk} \end{pmatrix}$$

Alternatively, we can write T as

$$T = \sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{X}_T\right)\left(x_{ij} - \bar{X}_T\right)'$$

The diagonal terms of T are $SS_{11}, \ldots, SS_{kk}$.

The hypothesis cross products for *p* and *q* are defined as:

$$CP_{pq} = \sum_{j=1}^{m} n_j\left(\bar{X}_{jp} - \bar{X}_p\right)\left(\bar{X}_{jq} - \bar{X}_q\right)$$

The matrix of hypothesis sum of squares and cross products *H* is defined as

$$H = \begin{pmatrix} CP_{11} & \cdots & CP_{1k} \\ \vdots & \ddots & \vdots \\ CP_{k1} & \cdots & CP_{kk} \end{pmatrix}$$

Alternatively

$$H = \sum_{j=1}^{m} n_j(\bar{X}_j - \bar{X}_T)(\bar{X}_j - \bar{X}_T)$$

The error (or residual) cross products for groups p and q is defined as follows:

$$ECP_{pq} = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ijp} - \bar{x}_{jp})(x_{ijq} - \bar{x}_{jq})'$$

Then the matrix of the error (or residual) sum of squares and cross products, denoted by E, is defined as

$$E = \begin{pmatrix} ECP_{11} & \cdots & ECP_{1k} \\ \vdots & \ddots & \vdots \\ ECP_{k1} & \cdots & ECP_{kk} \end{pmatrix}$$

Alternatively, we can write E as

$$E = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(X_{ij} - \bar{X}_j)(X_{ij} - \bar{X}_j)'$$

Theorem 12.5.1: We have

$$T = H + E$$

Proof: We can write

$$T = \sum_{j=1}^{m}\sum_{i=1}^{n_j}(x_{ij} - \bar{X}_T)(x_{ij} - \bar{X}_T)'$$

$$= \sum_{j=1}^{m}\sum_{i=1}^{n_j}\Big((x_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X}_T)\Big)\Big((x_{ij} - \bar{X}_j) + (\bar{X}_j - \bar{X}_T)\Big)'$$

$$= \sum_{j=1}^{m}\sum_{i-1}^{n_j}\Big[(x_{ij} - \bar{X}_j)(x_{ij} - \bar{X}_j)' + (x_{ij} - \bar{X}_j)(\bar{X}_j - \bar{X}_T)' + (\bar{X}_j - \bar{X}_T)(x_{ij} - \bar{X}_j)'$$

$$+ (\bar{X}_j - \bar{X}_T)(\bar{X}_j - \bar{X}_T)'\Big]$$

$$= \sum_{j=1}^{m} n_j(\bar{X}_j - \bar{X}_T)(\bar{X}_j - \bar{X}_T)' + \sum_{j=1}^{m}\sum_{i-1}^{n_j}(x_{ij} - \bar{X}_j)(x_{ij} - \bar{X}_j)'$$

$$= H + E.$$

The last result follows using the fact that $\sum_{i=1}^{n_j}(x_{ij} - \bar{X}_j) = 0$. Hence the theorem follows. For testing the null hypothesis$H_0: \mu_1 = \mu_2 = \cdots = \mu_m$, the following test statistics can be used.

12.5.3 Applications of MANOVA

- Profile analysis is a specialized application of multivariate ANOVA that allows researchers to examine how different groups differ across multiple dependent variables. Rather than focusing on overall group differences, profile analysis explores the unique patterns of means across the dependent variables. This technique is particularly useful when researchers are interested in understanding how groups differ in terms of their profiles or patterns of performance. For example, imagine a study that investigates the effects of three different teaching methods on students' academic performance across multiple subjects. By using profile analysis, researchers can identify if the groups show different patterns of performance across subjects, highlighting the effectiveness of each teaching method in specific subject areas.

- Multivariate analysis of covariance (MANCOVA) is an extension of MANOVA that incorporates one or more covariates into the analysis. Covariates are additional independent variables that are related to the dependent variables but are not the primary focus of the study. By including covariates in the analysis, researchers can control for their effects and better isolate the relationship between the independent variables and the dependent variables. For example, in a study examining the impact of a new medication on patients' physical and psychological well-being, researchers may include age, gender, and pre-existing medical conditions as covariates. MANCOVA allows them to assess the effects of the medication on the dependent variables while accounting for the potential influence of these covariates.

- In some research designs, it is necessary to measure the dependent variables multiple times over a period to capture changes or trends over time. Multivariate analysis of variance with repeated measures (MANOVA-RM) is a specialized technique that allows researchers to analyze data collected in this manner. This approach is particularly useful when studying longitudinal or within-subject designs. For example, imagine a study that investigates the effects of a new exercise program on individuals' physical fitness across three different

time points: before the program, midway, and after completion. By using MANOVA-RM, researchers can examine how the exercise program influences multiple dependent variables (e.g., cardiovascular endurance, muscular strength) over time, providing valuable insights into the program's effectiveness.

- Multivariate Profile Analysis is an advanced technique that combines the concepts of profile analysis and MANOVA-RM. It allows researchers to examine how different groups differ in their profiles across multiple dependent variables measured repeatedly over time. This technique is particularly useful when studying interventions or treatments that aim to change individual's profiles over time. For example, consider a study that investigates the effects of a mindfulness-based therapy on individual's well-being over a 12-week period. By using multivariate profile analysis, researchers can examine if the therapy leads to different profiles of well-being across the treatment group compared to the control group, providing insights into the therapy's efficacy.

12.5.4 Advantages of MANOVA

- It effectively condenses intricate correlations between several independent and dependent variables, assisting in the identification of interactions that univariate testing would overlook.
- By enabling to compare many dependent variables at once, MANOVA can help lower the possibility of Type I errors that might arise from running individual univariate tests for each variable.
- It maintains statistical power by controlling experiment-wise error rates more efficiently by considering all dependent variables collectively.
- A deeper comprehension of the data and underlying patterns can be facilitated by using MANOVA, which can shed light on the linkages and interactions between variables.

12.5.5 Disadvantages of MANOVA

- **Complexity:** Performing and interpreting a MANOVA can be challenging, particularly for researchers who are not familiar with multivariate statistics. It necessitates a solid grasp of the data and the methodology.

- **Assumption Stringency:** The assumptions of MANOVA are linearity, homogeneity of variance-covariance matrices between groups, and multivariate normality. Results that are not trustworthy may arise from breaking these presumptions.

12.6 Wilk's Lambda Criterion

Wilk's lambda distribution is a probability distribution used in multivariate hypothesis testing. It is defined from two independent Wishart distributed variables as the ratio distribution of their determinants, it is given by

$$\Lambda = \frac{|\mathrm{H}|}{|\mathrm{H} + \mathrm{E}|}$$

Wilks' lambda is a test statistic used in multivariate analysis of variance (MANOVA) to test whether there are differences between the means of identified groups of subjects on a combination of dependent variables.

Wilk's lambda is a direct measure of the proportion of variance in the combination of dependent variables that is unaccounted for by the independent variable. If a large proportion of the variance is accounted for by the independent variable, then it suggests that there is an effect from the grouping variable and that the groups have different mean values. Wilk's lambda statistic can be transformed to a statistic which has approximately an F distribution. This makes it easier to calculate the P-value.

There are a number of alternative statistics that can be calculated to perform a similar task to that of Wilk's lambda, such as Pillai's trace criterion and Roy's criterion.

Here H is large compared to E when the numerator of Λ is small compared to the denominator. We reject the null hypothesis when Wilk's Lambda is close to zero.

Hotelling-Lawley Trace:

The Hotelling Trace coefficient (also called Lawley-Hotelling or Hotelling-Lawley Trace) is a statistic for a multivariate test of mean differences between two groups.

$$T_0^2 = trace(HE^{-1})$$

H is large compared to E when Hotelling-Lawley Trace is large. Thus, we reject the null hypothesis when Hotelling-Lawley trace is large.

Pillai-Bartlett Trace:

$$V = tr(H(H+E)^{-1})$$

If H is large compared to Ethen statistic V will be large. Thus, we reject the null hypothesis when V is large.

Roy's Largest Root:

$$\Theta = \text{largest eigenvalue of } HE^{-1}$$

We reject the null hypothesis when Θ is large. If λ_p is the largest eigenvalue of HE^{-1}, the following alternative version can also be used:

$$\frac{\lambda_p}{1+\lambda_p}$$

Now we prove the following results:

Result 12.6.1: We can write

$$\Lambda = \frac{1}{|I+HE^{-1}|}$$

Proof: Since E and H are symmetric matrices, we have

$$HE^{-1} = E^{-1}H$$

Then

$$E(I+HE^{-1}) = E(I+E^{-1}H) = E+H$$

Hence, taking determinant, we obtain

$$|E|\ |I\ +\ HE^{-1}| = |E(I+HE^{-1})| = \ |E\ +\ H|.$$

Therefore

$$\Lambda = \frac{|E|}{|E+H|}$$

$$= \frac{1}{|I+HE^{-1}|}$$

Which leads to the required result.

Result 12.6.2.: Let $\lambda_1, \ldots, \lambda_k$ be the eigen values of HE^{-1}. Then

Wilk's Lambda:

$$\Lambda = \prod_{p=1}^{k} \frac{1}{1+\lambda_p}$$

Hotelling-Lawley Trace:

$$T_0^2 = \sum_{p=1}^{k} \lambda_p$$

Pillai-Bartlett Trace:

$$V = \sum_{p=1}^{k} \frac{\lambda_p}{1+\lambda_p}$$

Proof: The eigenvalues of $I + HE^{-1}$ are

$1 + \lambda_1, \ldots, 1 + \lambda_k$

Hence

$|I + HE^{-1}| = \prod_{p=1}^{k}(1 + \lambda_p)$.

Therefore

$$\Lambda = \prod_{p=1}^{k} \frac{1}{1+\lambda_p}$$

For any matrix *A*, trace(*A*)=sum of its eigenvalues. Hence $T_0^2 = tr(HE^{-1}) = \sum_{p=1}^{k} \lambda_p$.

Finally, we have

$$H(H+E)^{-1} = (H^{-1}H + H^{-1}E)^{-1} = (I + H^{-1}E)^{-1}$$

Now, if λ_p is an eigen value of HE^{-1}, then $\frac{1}{1+\lambda_p}$ is an eigen value of $(I + HE^{-1})^{-1}$. Thus

$$V = tr(H(H+E)^{-1}) = \sum_{p=1}^{k} \frac{1}{1+\lambda_p}.$$

Hence the result follows.

The Pillai-Barlett Trace is like multiple correlation coefficient $R^2 = SSB/SST$, which and is the proportion of the variance explained by the model. It is the most robust in cases of violation of the assumptions at least for balanced models.

Wilk's Lambda is like $R^2 = SSE/SST$.

The Hotelling-Lawley Trace is like F-test used in ANOVA $F = SSB/SSE$.

We state the following results without proof:

Result 12.6.3.:

(i) Let

$$a = n - m - \frac{k-m+2}{2},$$

$$b = \begin{cases} \sqrt{\dfrac{k^2(m-1)^2 - 4}{k^2 + (m-1)^2 - 5}}, & \text{if} k^2 + (m-1)^2 - 5 > 0 \\ 1, & \text{otherwise} \end{cases}$$

$$c = \frac{k(m-1) - 2}{2}$$

Then, under the null hypothesis

$$F = \frac{1 - \Lambda^{\frac{1}{b}}}{\Lambda^{\frac{1}{b}}} \frac{df_2}{df_1} \sim F(df_1, df_2)$$

where $df_1 = k(m-1), df_2 = ab - c$.

(ii) Let

$s = \min(k, m-1) =$ number of non − zero eigenvalues in HE^{-1}

$$t = \frac{|k-m+1|-1}{2}$$

$$u = \frac{n-m-k-1}{2}$$

Under the null hypothesis

$$F = \frac{T_0^2}{s} . \frac{df_2}{df_1} \sim F(df_1, df_2)$$

with $df_1 = s(2t+s+1) = s.\max(k, m-1)\,, df_2 = 2(su+1)$

(iii) We have under the null hypothesis

$$F = \frac{V}{s-V} \frac{df_2}{df_1} \sim F(df_1, df_2)$$

with $df_1 = s(2t+s+1), df_2 = s(2u+s+1)$.

The above distributions of various statistics can be used to form the critical regions for testing the null hypothesis $H_0: \mu_1 = \mu_2 = \cdots = \mu_m$.

12.7 Summary

In this unit, we have covered the concepts of Linear Regression Model under following situations:

- We have discussed Multivariate linear regression model.
- We have derived Estimation of parameters and their properties.
- We have explained Multivariate analysis of variance [MANOVA] of one-way classified data.

- We have discussed Wilk's lambda criterion

12.8 Self-Assessment Exercises

1. Show that $E\left(Y - X\hat{B}\right)'\left(Y - X\hat{B}\right) = (n - q)\Sigma$
2. Show that

   $Y'Y = \hat{Y}'\hat{Y} + \hat{\epsilon}'\,\hat{\epsilon}$
3. Discuss the multivariate analysis of variance for one-way classified data. How can we test the equality of means of several groups using MANOVA?
4. Show that Wilks' Λ can be expressed in terms of the eigenvalues of $E^{-1}H$ as in

   $$\Lambda = \prod_{p=1}^{k}\left(\frac{1}{1+\lambda_p}\right)$$
5. Show that

   $$\Lambda = \frac{1}{|I + HE^{-1}|}$$

12.9 References

- Anderson, T. W. (2003): An Introduction to Multivariate Statistical Analysis. United Kingdom: Wiley.
- Brenner, D., Bilodeau, M. (1999): Theory of Multivariate Statistics. Germany: Springer.
- Dillon William R & Goldstein Mathew (1984): Multivariate Analysis: Methods and Applications.
- Everitt B.S. & Dunn G. (1991): Applied Multivariate Data Analysis. Edward Arnold. London. pp. 219-220.
- Giri Narayan C. (1995): Multivariate Statistical Analysis
- Härdle WK, Simar L (2015): Applied Multivariate Statistical Analysis. Springer-Verlag, Berlin.
- Johnson, R. A., Wichern, D. W. (2019): Applied Multivariate Statistical Analysis. United Kingdom: Pearson.
- Kshirsagar A. M. (1979): Multivariate Analysis, Marcel Dekker Inc. New York.

- Mardia, K. V., Bibby, J. M., Kent, J. T. (1979): Multivariate Analysis. United Kingdom: Academic Press.
- Muirhead, R. J. (2009): Aspects of Multivariate Statistical Theory. Germany: Wiley.
- Rencher, A.C.: Methods of Multivariate Analysis, Second edition, Wiley.

12.10 Further Readings

- Brenner, D., Bilodeau, M.: Theory of Multivariate Statistics. Germany: Springer.
- Kotz, S., Balakrishnan, N. and Johnson, N.L.: Continuous Multivariate Distribution Models and Applications (Second Edition). Volume 1, Wiley - Inter science, New York.
- Khatri, C. G.: Multivariate Analysis.
- Mardia, K. V.: Multivariate Analysis.
- Seber, G.A.F.: *Multivariate Observations*. Wiley, New York.
- Rencher, Alvin C.: Multivariate Statistical Inference and Applications. John Wiley. New York, New York.