RAJALAKSHMI ENGINEERING COLLEGE
(AUTONOMOUS)
DEPARTMENT OF HUMANITIES & SCIENCES - MATHEMATICS

Computational Statistics + Lab

Introduction:

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

A major reason for basing statistical analysis on the normal distribution is that this probabilistic model approximates well the distribution of continuous measurements in many sampled populations.

Statistical theory based on the normal distribution has the advantage that the multivariate methods based on it are exten-sively developed and can be studied in an organized and systematic way. This is due not only to the need for such methods because they are of practical use, but also to the fact that normal theory is amenable to exact mathematical treatment.

The availability of modern computer facilities makes possible the analysis of large data sets and that ability permits the application of multivariate methods to new areas, such as image analysis and more effective analysis of data such as meterorological

Random Vectors and Matrices

A random vector is a vector whose elements are random variables. Similarly, a random matrix whose elements are random variables.

Let $X = \{X_{ij}\}$ be an $n \times p$ random matrix.

i.e) $$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix}$$

Expected value of a random matrix

The expected value of a random matrix (or vector) is the matrix (vector) consisting of the expected values of each of the elements.

$$E[X] = \begin{bmatrix} E[X_{11}] & E[X_{12}] & \cdots & E[X_{1p}] \\ E[X_{21}] & E[X_{22}] & \cdots & E[X_{2p}] \\ \cdots & \cdots & \cdots & \cdots \\ E[X_{n1}] & E[X_{n2}] & \cdots & E[X_{np}] \end{bmatrix}$$

$$E[X_{ij}] = \begin{cases} \int_{-\infty}^{\infty} x_{ij} f_{ij}(x_{ij})\, dx_{ij} & \text{if } X_{ij} \text{ is a continuous r.v.} \\ \sum_{\forall x_{ij}} x_{ij} P_{ij}(x_{ij}) & \text{if } X_{ij} \text{ is a discrete r.v.} \end{cases}$$

Mean Vectors

The marginal means are defined as

$$\mu_i = E[X_i] = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i)\, dx_i & \text{if } X_i \text{ is a continuous r.v} \\ \sum_{\forall x_i} x_i P_i(x_i) & \text{if } X_i \text{ is a discrete r.v} \end{cases}$$

The marginal variances are defined as

$$\sigma_i^2 = E[X_i - \mu_i]^2 = \begin{cases} \int_{-\infty}^{\infty} (x_i - \mu_i)^2 f_i(x_i)\, dx_i, & \text{if } X_i \text{ is a continuous r.v} \\ \sum_{\forall x_i} (x_i - \mu_i)^2 P_i(x_i), & \text{if } X_i \text{ is a discrete r.v.} \end{cases}$$

Covariance Matrices

$$\sigma_{ik}^2 = E(X_i - \mu_i)(X_k - \mu_k)$$

$$= \begin{cases} \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x_i - \mu_i)(x_k - \mu_k) f_i(x_i, x_k)\, dx_i\, dx_k, & \text{if } X_i, X_k \text{ is a continuous r.v} \\ \displaystyle\sum_{\forall x_i}\sum_{\forall x_k}(x_i - \mu_i)(x_k - \mu_k) P_i(x_i, x_k), & \text{if } X_i, X_k \text{ is a discrete r.v.} \end{cases}$$

Note

1. Covariance matrix captures the variance and linear correlation in multivariate/multidimensional data.
2. If data is an $n \times p$ matrix, the covariance matrix is a $p \times p$ square matrix.
3. Think of n, as the number of data instances (rows) and p the number of attributes (columns).

4. The Covariance of the return is

$$\sigma_{AB} = \frac{1}{T} \sum_{t=1}^{T} \left[x_{At} - \bar{x}_A\right]\left[x_{Bt} - \bar{x}_B\right]$$

5. It is always true that

(i) $\sigma_{AB} = \sigma_{BA}$

(ii) $\sigma_{ii} = \sigma_i^2$

Mean matrix

Let $X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$ be the $p \times 1$ random Vector.

$$E[X] = \begin{bmatrix} E[X_1] \\ E[X_2] \\ \vdots \\ E[X_p] \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \mu.$$

Covariance matrix

$$Cov(X] = \Sigma = E[(X-\mu)(X-\mu)']$$

$$= E\left(\begin{bmatrix} x_1-\mu_1 \\ \vdots \\ x_p-\mu_p \end{bmatrix}\begin{bmatrix} x_1-\mu_1 & \cdots & x_p-\mu_p \end{bmatrix}\right)$$

$$= E\begin{bmatrix} (x_1-\mu_1)^2 & \cdots & (x_1-\mu_1)(x_p-\mu_p) \\ \vdots & \cdots & \vdots \\ (x_p-\mu_p)(x_1-\mu_1) & & (x_p-\mu_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(x_1-\mu_1)^2 & \cdots & E[(x_1-\mu_1)(x_p-\mu_p)] \\ \vdots & \vdots & \vdots \\ E[(x_p-\mu_p)(x_1-\mu_1)] & \cdots & E(x_p-\mu_p)^2 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & & \vdots & \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{bmatrix}$$

Example

1. Find the mean and covariance matrix for the two r.v X_1 and X_2 for the given joint probability function $P_{12}(X_1, X_2)$ is

X_1 \ X_2	0	1
-1	0.24	0.06
0	0.16	0.14
1	0.40	0

Soln:- Marginal Distribution of X

X_1	-1	0	1
$P(X_1)$	0.3	0.3	0.4

$$\mu_1 = E[X_1] = \sum x_1 P(x_1) = 0.1$$

Marginal Distribution of Y

X_2	0	1
$P(X_2)$	0.8	0.2

$$\mu_2 = E[X_2] = \sum x_2\, p(x_2) = 0.2$$

$$E[X] = \begin{bmatrix} E(X_1) \\ E(X_2) \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.2 \end{bmatrix}$$

$$\sigma_{11} = E[X_1 - \mu_1)^2 = \sum_{x_1} (x_1 - 0.1)^2\, P_1(x_1)$$

$$= 0.69$$

$$\sigma_{22} = E(X_2 - \mu_2)^2 = \sum_{x_2} (x_2 - 0.2)^2\, P_2(x_2)$$

$$= 0.16$$

$$\sigma_{12} = E[(X_1 - \mu_1)(X_2 - \mu_2)]$$

$$= \sum_{(x_1, x_2)} (x_1 - 0.1)(x_2 - 0.2)\, P_{12}(x_1, x_2)$$

$$= -0.08$$

$$\therefore \ \mathrm{Cov}(X) = \Sigma = E[(X-\mu)(X-\mu)']$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} 0.69 & -0.08 \\ -0.08 & 0.16 \end{bmatrix}$$

(2) Example: The table provides the returns on three assets over three years

	year 1	year 2	year 3
A	10	12	11
B	10	14	12
C	12	06	09

Mean returns $\bar{x}_A = 11$, $\bar{x}_B = 12$, $\bar{x}_C = 9$

Covariance

$$\sigma_{AB} = \frac{1}{3}\left\{ [10-11][10-12] + [12-11][14-12] + [11-11][12-12] \right\}$$

$$\sigma_{AB} = 1.333$$

likewise, $\sigma_{AC} = -2$ and $\sigma_{BC} = -4$

The matrix is symmetric

$$\Sigma = \begin{bmatrix} \sigma_A^2 & \sigma_{AB} & \sigma_{AC} \\ \sigma_{AB} & \sigma_B^2 & \sigma_{BC} \\ \sigma_{AC} & \sigma_{BC} & \sigma_C^2 \end{bmatrix}$$

$$\therefore \Sigma = \begin{bmatrix} 0.666 & 1.333 & -2 \\ 1.333 & 2.66 & -4 \\ -2 & -4 & 6 \end{bmatrix}$$

Correlation Coefficient:

Let the population correlation coefficient matrix be the $p \times p$ symmetric matrix.

$$\rho = \begin{bmatrix} \frac{\sigma_{11}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{11}}} & \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}} & \cdots & \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} \\ \vdots & \vdots & & \vdots \\ \frac{\sigma_{1p}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}}} & \frac{\sigma_{2p}}{\sqrt{\sigma_{22}}\sqrt{\sigma_{pp}}} & \cdots & \frac{\sigma_{pp}}{\sqrt{\sigma_{pp}}\sqrt{\sigma_{pp}}} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & & & \\ \rho_{1p} & \rho_{2p} & \cdots & 1 \end{bmatrix}$$

Standard Deviation,

Let the $p \times p$ standard deviation be

$$V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

Then it is verified that

$$V^{1/2} \rho V^{1/2} = \Sigma$$

$$\rho = (V^{1/2})^{-1} \Sigma (V^{1/2})^{-1}$$

Example

1. Suppose $\Sigma = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix}$ obtain $V^{1/2}$ and ρ.

Soln: Given $\Sigma = \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{bmatrix}$

$$\therefore V^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & 0 \\ 0 & \sqrt{\sigma_{22}} & 0 \\ 0 & 0 & \sqrt{\sigma_{33}} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{bmatrix}$$

$$(V^{1/2})^{-1} = \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/5 \end{bmatrix}$$

$$\rho = (V^{1/2})^{-1} \Sigma (V^{1/2})^{-1}$$

$$= \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/5 \end{bmatrix} \begin{bmatrix} 4 & 1 & 2 \\ 1 & 9 & -3 \\ 2 & -3 & 25 \end{bmatrix} \begin{bmatrix} 1/2 & 0 & 0 \\ 0 & 1/3 & 0 \\ 0 & 0 & 1/5 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 1/6 & 1/5 \\ 1/6 & 1 & -1/5 \\ 1/5 & -1/5 & 1 \end{bmatrix}$$

Multivariate Normal Distribution

The multivariate normal density function is defined as

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

where $F(x_1, \ldots, x_p) = \Pr(X_1 \le x_1, \ldots, X_p \le x_p)$ defined for every set of real numbers $x_1, \ldots, x_p$ and $F(x_1, \ldots, x_p)$ is absolutely continuous.

Thus

$$* \quad F(x_1, \ldots, x_p) = \int_{-\infty}^{x_p} \cdots \int_{-\infty}^{x_1} f(u_1, \ldots, u_p)\, du_1 \ldots du_p$$

The probability of falling in any (measurable) set R in the p-dimensional Euclidean space is.

$$\Pr\{(X_1, \ldots, X_p) \in R\} = \int \cdots \int_R f(x_1, \ldots, x_p)\, dx_1 \ldots dx_p$$

* The joint moments are defined as

$$E\left(X_1^{h_1}, \ldots, X_p^{h_p}\right) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} x_1^{h_1} \ldots x_p^{h_p} f(x_1, \ldots, x_p)\, dx_1 \ldots dx_p.$$

* The marginal distribution:

The marginal density of $X_1, \dots, X_r$ is $\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, \dots, x_r, u_{r+1}, \dots, u_p)\, du_{r+1} \dots du_p$.

* Statistical Independence:

If the cdf of $X_1, \dots, X_p$ is $F(x_1, \dots, x_p)$, the set of random variables is said to be mutually independent if

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

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

* Conditional Distribution.

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

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

Definition:- Multivariate Normal Distribution

The univariate normal density function can be written

$$k e^{-\frac{1}{2}\alpha(x-\beta)^2} = k e^{-\frac{1}{2}(x-\beta)\alpha(x-\beta)} \quad \text{--- (1)}$$

where α is positive and k is chosen so that the integral of (1) over the entire x-axis is unity. The density function of a multivariate normal distribution of $X_1, \ldots, X_p$ has an analogous form. The scalar variable x is replaced by a vector.

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} \quad \text{--- (2)}$$

the scalar constant β is replaced by a vector

$$b = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} \quad \text{--- (3)}$$

and the positive constant α is replaced by a positive definite (symmetric) matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & & & \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix} \quad \text{——— (4)}$$

The Square $\alpha(x-\beta)^2 = (x-\beta)\,\alpha\,(x-\beta)$ is replaced by the quadratic form

$$(x-b)' A (x-b) = \sum_{i,j=1}^{p} a_{ij} (x_i - b_i)(x_j - b_j) \quad \text{——— (5)}$$

Thus the density function of a p-variate normal distribution is

$$f(x_1, \ldots, x_p) = K e^{-\frac{1}{2}(x-b)' A (x-b)} \quad \text{——— (6)}$$

where K (>0) is chosen so that the integral over the entire p-dimensional Euclidean Space of $x_1, \ldots, x_p$ is unity.

where $K = \sqrt{|A|}\,(2\pi)^{-\frac{1}{2}p}$

$\therefore$ The normal density function is

$$\frac{\sqrt{|A|}}{(2\pi)^{\frac{1}{2}p}}\, e^{-\frac{1}{2}(x-b)' A (x-b)}$$

Conditional distribution to the multivariate normal distribution

Suppose that we have a random vector Z that is partitioned into component X and Y that is realized from a multivariate normal distribution with mean vector with corresponding components μ_x and μ_y and variance-covariance matrix which has been partitioned into four parts as shown below

$$Z = \begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{pmatrix}\right)$$

Here Σ_X is the variance - Covariance matrix for the random vector X. Σ_Y is the variance co-variance matrix for the random vector Y. And, Σ_{YX} contains the covariances between the elements of X and the corresponding elements of Y.

Then the conditional distribution of Y given that X takes a particular value x is also going to be a multivariate normal with conditional expectation as shown below.

$$E(Y/X=x) = \mu_Y + \Sigma_{YX}\Sigma_X^{-1}(x-\mu_X)$$

The conditional variance - covariance matrix of Y given $X=x$ is equal to the variance - covariance matrix for Y minus the term that involves the co-variance between X and Y and the variance - covariance matrix for X.

$$Var(Y/X=x) = \Sigma_Y - \Sigma_{YX}\Sigma_X^{-1}\Sigma_{XY}.$$

Unit-II Multivariate analysis of Variance and Covariance

One-way Multivariate analysis of variance:

(One-way MANOVA):

Suppose that we have data on p variables which we can arrange in a table such as the one below:

Treatments

	1	2	....	g
Subject 1	$Y_{11} = \begin{pmatrix} y_{111} \\ \vdots \\ y_{11p} \end{pmatrix}$	$Y_{21} = \begin{pmatrix} y_{211} \\ \vdots \\ y_{21p} \end{pmatrix}$	....	$Y_{g1} = \begin{pmatrix} y_{g11} \\ \vdots \\ y_{g1p} \end{pmatrix}$
2	$Y_{12} = \begin{pmatrix} y_{121} \\ \vdots \\ y_{12p} \end{pmatrix}$	$Y_{22} = \begin{pmatrix} y_{221} \\ \vdots \\ y_{22p} \end{pmatrix}$	....	$Y_{g2} = \begin{pmatrix} y_{g21} \\ \vdots \\ y_{g2p} \end{pmatrix}$
$\vdots$	$\vdots$			
n_i	$Y_{1n_1} = \begin{pmatrix} y_{1n_11} \\ \vdots \\ y_{1n_1p} \end{pmatrix}$	$Y_{2n_2} = \begin{pmatrix} y_{2n_21} \\ \vdots \\ y_{2n_2p} \end{pmatrix}$	....	$Y_{gn_g} = \begin{pmatrix} y_{gn_g1} \\ \vdots \\ y_{gn_gp} \end{pmatrix}$

In this multivariate case the scalar quantities Y_{ij}, of the corresponding table in ANOVA are replaced by vectors having p observations.

where Y_{ijk} = Observation for variable k from subject j in group i. These are collected into vectors:

$$Y_{ij} = \begin{pmatrix} Y_{ij1} \\ \vdots \\ Y_{ijp} \end{pmatrix} = \text{Vector of variables for subject } j \text{ in group } i$$

n_i = the number of subjects in group i

$N = n_1 + n_2 + \dots + n_g$ = Total sample size.

Assumptions:

1. The data from group i has common mean vector $\mu_i = \begin{pmatrix} \mu_{i1} \\ \vdots \\ \mu_{ip} \end{pmatrix}$

2. The data from all groups have common variance-covariance matrix Σ.

3. Independence: The subject are independ-ently sampled.

4. Normality: The data are multivariate normally distributed.

Here we are interested in testing the null hypothesis that the group mean vectors are all equal to one another. (that is)

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_g .$$

The alternative hypothesis being:

$$H_1 : \mu_{ik} \neq \mu_{jk} \text{ for atleast one } i \neq j$$

and atleast one variable k.

This says that the null hypothesis is false if atleast one pair of treatments is different on atleast one variable.

Sample mean vector:

$$\bar{Y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij} = \begin{pmatrix} \bar{Y}_{i.1} \\ \vdots \\ \bar{Y}_{i.p} \end{pmatrix} = \text{Sample mean vector}$$

for group i. This sample mean vector is comprised of the group means for each of the p variables. Thus,

$$\bar{y}_{i.k} = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ijk}$$ = sample mean vector for variable k in group i.

Grand Mean Vector:

$$\bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{q} \sum_{j=1}^{n_i} Y_{ij} = \begin{pmatrix} \bar{y}_{..1} \\ \vdots \\ \bar{y}_{..p} \end{pmatrix}$$ = grand mean vector.

This grand mean vector is comprised of the grand means for each of the p variables. Thus,

$$\bar{y}_{..k} = \frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{n_i} Y_{ijk}$$ = grand mean for variable k

Total Sum of Squares and Cross Products:

In the univariate analysis of variance, we defined the total sums of squares, a scalar quantity. The multivariate analog is the Total sum of Squares and Cross Products matrix, a $p \times p$ matrix of numbers. The total

Sum of squares is a cross product matrix defined by the expression below:

$$T = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})(Y_{ij} - \bar{Y}_{..})'$$

Here, the $(k, l)^{th}$ element of T is

$$\sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ijk} - \bar{Y}_{..k})(Y_{ijl} - \bar{Y}_{..l})$$

For $k=l$, this is the total sum of squares for variable k and measures the total variation in the k^{th} variable. For $k \neq l$, this measures the dependence between variables k and l across all of the observations.

We may partition the total sum of squares and cross products as follows

$$T = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{..})(Y_{ij} - \bar{Y}_{..})'$$

$$= \sum_{i=1}^{g} \sum_{j=1}^{n_i} \{(Y_{ij} - \bar{Y}_{i}) + (\bar{Y}_{i} - \bar{Y}_{..})\}$$

$$\{(Y_{ij} - \bar{Y}_{i}) + (\bar{Y}_{i} - \bar{Y}_{..})\}'$$

$$= \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i.})(y_{ij}-\bar{y}_{i.})'}_{E} + \underbrace{\sum_{i=1}^{g} n_i(\bar{y}_{i.}-\bar{y}_{..})(\bar{y}_{i.}-\bar{y}_{..})'}_{H}$$

where E is the Error sum of Squares and Cross products and H is the hypothesis sum of Squares and Cross products.

The $(k,l)^{th}$ element of the error sum of squares and cross products matrix E is,

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ijk}-\bar{y}_{i.k})(Y_{ijl}-\bar{y}_{i.l})$$

For $k=l$, this is the error sum of Squares for variables k, and measures the within treatment variation for the k^{th} variable. For $k \neq l$, this measures the dependence between variables k and l after taking into account the treatment.

The $(k,l)^{th}$ element of the hypothesis sum of squares and cross product matrix H is

$$\sum_{i=1}^{g} n_i(\bar{y}_{i.k}-\bar{y}_{..k})(\bar{y}_{i.l}-\bar{y}_{..l})$$

For $k=l$, this is the treatment sum of squares for variable k, and measures the between treatment variation for the k^{th} variable. For $k \neq l$, this measures dependence of variables k and l across treatments.

The partitioning of the total sum of squares and cross products matrix may be summarized in the multivariate analysis of variance table:

MANOVA

Source	d.f	SSP
Treatments	$g-1$	H
Error	$N-g$	E
Total	$N-1$	T

We wish to reject H_0, if the hypothesis sum of squares and cross products matrix H is large relative to the error sum of squares and cross products matrix E.

* Discriminant Analysis Procedure:

Discriminant analysis is a 7-step procedure.

Step: 1 Collect training data

Training data are data with known group memberships. Here, we actually know which population contains each subject.

Step: 2 Prior Probabilities

The prior probability p_i represents the expected portion of the community that belongs to population π_i. There are three common choices:

1. Equal priors: $\hat{P}_i = \frac{1}{g}$ This is useful if we believe that all of the population sizes are equal

2. Arbitrary priors selected according to the investigators beliefs regarding the relative population sizes

Also, $\hat{P}_1 + \hat{P}_2 + \cdots + \hat{P}_g = 1$

8. Estimated priors:

$$\hat{p}_i = \frac{n_i}{N}$$

where n_i is the number observations from population π_i in the training data, and $N = n_1 + n_2 + \cdots + n_g$.

Step 3: Bartlett's Test

Use Bartlett's test to determine if the variance-covariance matrices are homogeneous for all populations involved. The result of this test will determine whether to use Linear or Quadratic Discriminant analysis.

Case: 1 Linear

Linear discriminant analysis is for homogeneous variance-covariance ~~analysis~~ matrices:

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_g = \Sigma$$

In this case, the variance-covariance matrix does not depend upon the population.

Case 2: Quadratic

Quadratic discriminant analysis is used for heterogeneous variance - Covariance matrices:

$$\Sigma_i \neq \Sigma_j \text{ for some } i \neq j$$

This allows the variance - Covariance matrices to depend on the population.

Step 4: Estimate the parameters of the Conditional probability density functions $f(X/\pi_i)$:

The following standard assumptions are

1. The data from group i has common mean vector μ_i.
2. The data from group i has common variance - Covariance matrix Σ.
3. Independence: The subjects are independently sampled.
4. Normality: The data are multivariate normally distributed.

Step 5: Compute Discriminant functions.

Step 6: Use Cross validation to estimate misclass-ification probabilities.

Step 7: classify observations with unknown group memberships.

* Linear Discriminant Analysis:

We assume that in population π_i, the probability density function of x is multivariate normal with mean vector μ_i and variance - Covariance matrix Σ (same for all populations). As a formula, i.e.)

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

we classify to the population for which $p_i f(x/\pi_i)$ is largest.

Linear discriminant analysis is used when the variance - Covariance matrix does not depend on the population. In this case, our decision rule is based on the Linear

Score function, a function of the population means for each of our g population, μ_i as well as the pooled variance - Covariance matrix.

Linear Score Function:

$$S_i^L(x) = -\tfrac{1}{2}\mu_i' \Sigma^{-1}\mu_i + \mu_i' \Sigma^{-1} x + \log P_i$$

$$= d_{i0} + \sum_{j=1}^{p} d_{ij} x_j + \log P_i$$

where $d_{i0} = -\frac{1}{2}\mu_i' \Sigma^{-1}\mu_i$

d_{ij} = j^{th} element of $\mu_i' \Sigma^{-1}$.

The expression resembles a linear regression with intercept term d_{i0} and regression coefficients d_{ij}.

Linear Discriminant Function:

$$d_i^L(x) = -\frac{1}{2}\mu_i' \Sigma^{-1}\mu_i + \mu_i' \Sigma^{-1} x$$

$$= d_{i0} + \sum_{j=1}^{p} d_{ij} x_j$$

where $d_{i0} = -\frac{1}{2}\mu_i' \Sigma^{-1} \mu_i$

Given a sample unit with measurements $x_1, x_2, \ldots, x_p$ we classify the sample unit into the population that has the largest linear score function. This is equivalent to classifying to the population for which the posterior probability of membership is largest. The linear score function is computed for each population, then we plug in our observation values and assign the unit to the population with the largest score.

However, this is a function of unknown parameters, μ_i and Σ. So, these must be estimated from the data.

Discriminant analysis requires estimates of

$$p_i = P_r(\pi_i) \quad \forall\ i = 1, 2, \ldots, g$$

$$\mu_i = E(X/\pi_i) \quad \forall\ i = 1, 2, \ldots, g$$

$$\Sigma = Var(X/\pi_i) \quad \forall\ i = 1, 2, \ldots, g$$

- Prior probabilities
- The population means are estimated by the sample mean vectors.
- The variance – Covariance matrix is estimated by using the pooled variance-Covariance matrix.

Typically, these parameters are estimated from training data, in which the population membership is known.