# UNIT 24 PRINCIPAL COMPONENT ANALYSIS

**Structure** **Page No.**



## 24.1 INTRODUCTION

Principal component analysis is a useful statistical technique that has found applications in diverse fields. It helps in reducing the dimensionality of the data set and identification of patterns in the data through graphical representation. In nutshell, we can say that principal component analysis is a data reduction technique.

The purpose of principal component analysis is to 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, which are discussed in Section 24.2. Often a small number of principal components can be used in place of the original variables for plotting, regression, clustering and so on. Principal component analysis can also be viewed as a technique to remove multicollinearity in the data. In this technique, we transform the original set of variables to a new set of uncorrelated random variables, called principal components. These new variables are linear combinations of the original variables and are derived in decreasing order of importance so that the first principal component accounts for as much as possible of the variation in the original data. The researcher often hopes that the first few components will account for most of the variation in the original data so that the effective dimensionality of the data can be reduced. The variance of principal components is computed in Section 24.3 In Section 24.4, contribution of variability through spectral decomposition of $\Sigma$ is discussed.

Principal component analysis often reveals relationships that were not previously suspected and thereby allows interpretations that would not ordinarily result. Geometrically, these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with say, $x_1, x_2, \cdots, x_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simple and more parsimonious description of the covariance structure. The identification of important components are done in Section 24.5. Principal components from the correlation matrix are discussed in Section 24.6. In Section 24.7, estimation of principal components is discussed. The scaling convention, geometrical interpretation and applications of principal components are discussed in Section 24.8, 24.9 and 24.10 respectively.

### 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
- identify the important principal components
- describe the geometrical interpretations of principal components
- apply the principal components in diverse fields

## 24.2 DERIVATION OF PRINCIPAL COMPONENTS

Let us start this section by defining and deriving the principal components. Suppose $x' = \left[x_1, \cdots, x_p\right]$ is a p-dimensional random variable with mean vector $\boldsymbol{\mu}$ and covariance matrix $\Sigma = ((\sigma_{ij}))$. The problem is to find a new set of variables, say $z_1, z_2, \cdots, z_p$, which are uncorrelated and whose variances decrease from first to last. Each $z_j$ is taken to be linear combination of the x's, so that

$$z_j = a_{1j}x_1 + a_{2j}x_2 + \cdots + a_{pj}x_p \qquad (1)$$
$$= a_j'x, \qquad \text{where } j = 1, 2, \cdots, p,$$

where $a_j' = \left[a_{1j}, a_{2j}, \cdots, a_{pj}\right]$ is a vector of constants. Equation (1) contains an arbitrary scale factor. We therefore impose the condition that $a_j' a_j = \sum_{k=1}^{p} a_{kj}^2 = 1$. We shall see that this particular normalization procedure ensures that the overall transformation is orthogonal in other words, that distances in p-space are preserved.

The first principal component $z_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1p}x_p$, is found by choosing $\mathbf{a}_1$ such variance of $z_1$ is as large as possible subject to the condition that $a_{11}^2 + a_{12}^2 + \ldots + a_{1p}^2 = 1$ or $\mathbf{a}_1' \mathbf{a}_1 = 1$.
This constraint is introduced because if this is not done, then $\mathrm{Var}(z_1)$ can be increased simply by multiplying each $a_{1j}$'s by a common constant factor.

The second principal component $z_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2p}x_p$, is found by choosing $\mathbf{a}_2$ such that $\mathrm{Var}(z_2)$ is as large as possible (next to $\mathrm{Var}(z_1)$) subject to the constraint that $a_{21}^2 + a_{22}^2 + \ldots + a_{2p}^2 = 1$ or $\mathbf{a}_2' \mathbf{a}_2 = 1$ and $\mathrm{Cov}(z_1, z_2) = 0$. Continuing this process, the remaining *p*–2 principal components $z_3, \cdots, z_p$ can be derived so that $z_i$ is uncorrelated with the first t(i-1) principal components and variance has maximum subject to $a_j' a_j = 1$. Then $z_i$'s have variances in decreasing order.

We begin by finding the first principal component. We want to choose $\mathbf{a}_1$ so as to maximize the variance of $z_1$ subject to the normalization constraint that $\mathbf{a}_1' \mathbf{a}_1 = 1$.

As

$$\begin{aligned}\mathrm{Var}(z_1) &= \mathrm{Var}(a_1' X) \\ &= \mathbf{a}_1' \boldsymbol{\Sigma} \mathbf{a}_1 . \end{aligned} \qquad (2)$$

so we take $\mathbf{a}_1' \boldsymbol{\Sigma} \mathbf{a}_1$ as our objective function.

The standard procedure for maximizing a function of several variables with one or more constraints is the method of Lagrange multipliers. With just one constraint, this method uses the fact that the stationary points of a differentiable function of *p* variables, say $f(x_1, \cdots, x_p)$, subject to a constraint $g(x_1, \cdots, x_p) = c$, are such that there exists a number $\lambda$, called the Lagrange multiplier, so that

$$\frac{\partial f}{\partial x_i} - \lambda \frac{\partial g}{\partial x_i} = 0 \qquad i = 1, 2, \cdots, p \qquad (3)$$

at the stationary points. These p equations, together with the constraint, are sufficient to determine the co-ordinates of the stationary points (and the corresponding values of $\lambda$, which, however, are usually of little interest). Further investigation is needed to see if a stationary point is a maximum, minimum or saddle point. It is helpful to form a new function, $L(x)$, such that

$$L(x) = f(x) - \lambda \left[ g(x) - c \right]$$

where the term in the square brackets is of course zero. Then the set of equations in (3) may be written simply as

$$\frac{\partial L}{\partial x} = 0 .$$

Applying this method to maximization of $\mathbf{a}_1' \boldsymbol{\Sigma} \mathbf{a}_1$, we write

$$L(a_1) = a_1' \Sigma a_1 - \lambda (a_1' a_1 - 1) .$$

Then, we have

$$\frac{\partial L}{\partial x} = 2\Sigma a_1 - 2\lambda a_1 .$$

Setting this equal to 0, we have

$$(\Sigma - \lambda I) a_1 = 0 . \qquad (4)$$

We now come to the crucial step in the argument. If Equation (4) is to have a non trivial solution for $\mathbf{a}_1$, then $(\Sigma - \lambda I)$ must be a singular matrix. Thus $\lambda$ must be chosen that

$$|\Sigma - \lambda I| = 0 .$$

Thus a non-zero solution for Equation (4) exists if and only if $\lambda$ is an eigenvalue of $\boldsymbol{\Sigma}$. But $\boldsymbol{\Sigma}$ will generally have *p* eigenvalues, all of which must be non-negative as $\boldsymbol{\Sigma}$ is positive semi-definite symmetric matrix. Let us denote the eigenvalue by $\lambda_1, \lambda_2, \cdots, \lambda_p$, and assume for the moment that they are distinct, and that $\lambda_1 > \lambda_2 > \cdots > \lambda_p \geq 0$. Which one shall be choose to determine the first principal component? Now,

$$\begin{aligned}\mathrm{Var}(a_1' x) &= (a_1' \Sigma a_1) \\ &= (a_1' \lambda I a_1) \quad \text{(using equation (4))} \\ &= \lambda .\end{aligned}$$

Since we are interested in maximization of this variance, therefore, we choose $\lambda$ to be largest eigenvalue, namely $\lambda_1$. Then, using equation (4), $\mathbf{a}_1$, must be the eigenvector of $\boldsymbol{\Sigma}$ corresponding to the largest eigenvalue.

The second principal component, namely $z_2 = a_2' x$, is obtained by an extension of the above argument. In addition to the scaling constraint that $\mathbf{a}_2' \mathbf{a}_2 = 1,$, we now have a second constraint that $z_2$ should be uncorrelated with $z_1$. Now,

$$\begin{aligned} \text{Cov}(z_2, z_1) &= \text{Cov}\ (a_2' x, a_1' x) \\ &= \text{E}\left[ a_2' (x-\mu)(x-\mu)' a_1 \right] \\ &= a_2' \Sigma\, a_1 . \end{aligned} \tag{5}$$

Since, the first two principal components should be uncorrelated, therefore it is required that $a_2' \Sigma\, a_1$ is equal to zero. But since $\Sigma a_1 = \lambda_1 a_1$, an equivalent simpler condition is that $\mathbf{a}_2' \mathbf{a}_1 = 0.$ In other words, $\mathbf{a}_1$ and $\mathbf{a}_2$ should be orthogonal.

In order to maximize the variance of $z_2$, namely $\mathbf{a}_2' \boldsymbol{\Sigma}\, \mathbf{a}_2$, subject to the two constraints, viz. $\mathbf{a}_2' \mathbf{a}_2 = 1$ and $\mathbf{a}_2' \mathbf{a}_1 = 0,$ we need to introduce two Lagrange multipliers, which may be denoted by $\lambda$ and $\Gamma$. Now consider the function

$$L(a_2) = a_2' \Sigma\, a_2 - \lambda (a_2' a_2 - 1) - \Gamma a_2' a_1 .$$

At the stationary point(s) we must have

$$\frac{\partial L}{\partial a_2} = 2(\Sigma - \lambda\, I)\, a_2 - \Gamma a_1 = 0 . \tag{6}$$

If we pre-multiply this equation by $\mathbf{a}_1'$ we get $2a_1' \Sigma\, a_2 - \Gamma = 0$, since $\mathbf{a}_1'\ \mathbf{a}_2 = 0.$ But from equation (5), $\mathbf{a}_1'\ \boldsymbol{\Sigma}\, \mathbf{a}_2$ required to be zero, so that $\Gamma$ is zero at the stationary point(s). Thus equation (6) becomes $(\Sigma - \lambda\, I) a_2 = 0$.

This time we choose $\lambda = \lambda_2,$ the second largest eigenvalue of $\boldsymbol{\Sigma}$, and $\mathbf{a}_2$ to be the corresponding eigenvector.

Continuing as above, $a_j$ turns out to be the eigenvector associated with the $j^{th}$ largest eigenvalue.

There is no difficulty in extending the above argument to the case where some of the eigenvalues of $\boldsymbol{\Sigma}$ are equal. In that case there is no unique way of choosing the corresponding eigenvectors, but as long as the eigenvectors associated with multiple roots are chosen to be orthogonal, the argument carries through.

## 24.3 VARIANCE OF PRINCIPAL COMPONENTS

Let $\mathbf{A} = \left[ a_1, \cdots, a_p \right]$, denote the $(p \times p)$ matrix of the eigenvectors of $\Sigma$ obtained in the last section and the $(p \times 1)$ vector of principal components by $Z' = (z_1, z_2 . \cdots, z_p)$. Then

$$\mathbf{Z} = \mathbf{A}' \mathbf{x} .$$

The $(p \times p)$ covariance matrix of $\mathbf{Z}$ is then

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ . & . & \cdots & 0 \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}.$$

Note that the matrix is diagonal as the components have been chosen to be uncorrelated.

One can also express $\mathrm{Var}(\mathbf{Z})$ in the form $\mathbf{A}'\mathbf{\Sigma}\mathbf{A}$, so that

$$\mathbf{\Lambda} = \mathbf{A}'\mathbf{\Sigma}\mathbf{A} \tag{7}$$

gives the important relation between the covariance matrix of original variables $\boldsymbol{x}$ and the corresponding principal components. Note that the Equation (7) can be rewritten as

$$\mathbf{\Sigma} = \mathbf{A}\,\mathbf{\Lambda}\,\mathbf{A}'. \tag{8}$$

since $\mathbf{A}$ is an orthogonal matrix.

We have already noted that the eigenvalues can be interpreted as the respective variances of the different principal components. Now the sum of these variances is given by

$$\sum_{i=1}^{p} \mathrm{Var}(z_i) = \sum_{i=1}^{p} \lambda_i = \mathrm{trace}(\mathbf{\Lambda}).$$

But

$$\begin{aligned} \mathrm{trace}(\mathbf{\Lambda}) &= \mathrm{trace}(\mathbf{A}'\mathbf{\Sigma}\mathbf{A}) \\ &= \mathrm{trace}(\mathbf{\Sigma}\mathbf{A}\mathbf{A}') \\ &= \mathrm{trace}(\mathbf{\Sigma}) \\ &= \sum_{i=1}^{p} \mathrm{Var}(x_i) \end{aligned}$$

Thus we have the important result that the sums of the variances of the original variables and of their principal components are the same. It is, therefore, convenient to make statements such as the $i^{th}$ principal component accounts for a proportion $\dfrac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$ of the total variation in the original data, though it should be emphasized that this is not an analysis of variance in the usual sense of the expression. We will also say that the first m principal components account for a proportion $\dfrac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j}$ of the total variation.

## 24.4 CONTRIBUTION OF VARIABILITY THROUGH SPECTRAL DECOMPOSITION OF $\Sigma$

Spectral decomposition of $\mathbf{\Sigma}$ is given by

$$\Sigma AA' = \lambda_1 a_1 a_1' + \lambda_2 a_2 a_2' + \cdots + \lambda_p a_p a_p'.$$

This clearly shows that the contribution of the first principal component to $\boldsymbol{\Sigma}$ is $\lambda_1 a_1 a_1'$ when that is removed, the residual matrix of variances and covariances is

$$\Sigma - \lambda_1 a_1 a_1' = \sum_{j=2}^{p} \lambda_j a_j a_j'$$

The second principal component contributes $\lambda_2 a_2 a_2'$ and so on.

## 24.5 IDENTIFICATION OF IMPORTANT COMPONENTS

The usual procedure is to look at the first few components which, hopefully, account for a large proportion of the total variance. In order to do this, it is necessary to decide which eigenvalues are 'large' and which are 'small' so that components corresponding to the later may be disregarded. When analysing a correlation matrix where the sum of the eigenvalues is p, many researchers use the rule that eigenvalues less than 1 may be disregarded. This arbitrary policy is a useful rule of thumb but has no theoretical justification. It may be better to look at the pattern of eigenvalues and see if there is a natural breakpoint. The fact that there is no objective way of deciding how many components to retain is a serious drawback of the method.

Alternatively, the number of important principal components, (say q < p) can be known, if $\boldsymbol{\Sigma}$ can be approximated by another matrix of order *p*, say **B**, of a smaller rank, say q(< p). The goodness of approximation of the closeness of **B** to $\boldsymbol{\Sigma}$ can be measured by the norm of $\boldsymbol{\Sigma} - \mathbf{B}$ which is defined as

$$\|\boldsymbol{\Sigma} - \mathbf{B}\|_2 = \left( \sum_{i=1}^{p} \sum_{j=1}^{p} (\sigma_{ij} - b_{ij})^2 \right)^{1/2}.$$

Here $\|.\|_p$ denotes the entry wise p-norm of vectors. Entry wise p-norm of a matrix $A_{m\times n} = ((a_{ij}))$ is defined as $\|A\|_p = \left( \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^p \right)^{1/p}$.

Among all possible $p \times p$ matrices B of rank q, we desire to find the one for which the norm is minimum. We see that

$$\begin{aligned} \|\boldsymbol{\Sigma} - \mathbf{B}\|_2^2 &= \text{trace}(\boldsymbol{\Sigma} - \mathbf{B})\ (\boldsymbol{\Sigma} - \mathbf{B})' \\ &= \text{trace}\{\mathbf{A}\mathbf{A}'\ (\boldsymbol{\Sigma} - \mathbf{B})\ \ \mathbf{A}\mathbf{A}'\ (\boldsymbol{\Sigma} - \mathbf{B})'\} \\ &= \text{trace}\{(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} - \mathbf{A}'\mathbf{B}\mathbf{A})(\mathbf{A}'\boldsymbol{\Sigma}\mathbf{A} - \mathbf{A}'\mathbf{B}\mathbf{A})'\} \\ &= \text{trace}(\boldsymbol{\Lambda} - \mathbf{G})(\boldsymbol{\Lambda} - \mathbf{G})' \end{aligned}$$

where $\mathbf{G} = \mathbf{A}'\mathbf{B}\mathbf{A}$.

Since **A** is an orthogonal matrix, **G** is of the same rank as **B**, viz. q. If the elements of **G** are $g_{ij}$, then

$$\|\boldsymbol{\Sigma} - \mathbf{B}\|_2^2 = \sum_{i=1}^{p} (\lambda_i - g_{ii})^2 + \sum_{i=1}^{p} \sum_{j \neq i=1}^{p} g_{ij}^2$$

To minimize this, we must obviously choose $g_{ij} = 0$ for all i, j such that $i \neq j$, i.e., **G** is a diagonal matrix. But **G** is of rank q. Therefore, $p - q$ elements on the diagonal of **G** must be zero and the remaining elements must be different from zero. In order to satisfy this and minimize the quantity $\sum_{i=1}^{p} (\lambda_i - g_{ii})^2$, it is obvious that the choice must

be $g_{ii} = \lambda_i$ $(i = 1, 2, \cdots, q)$. Therefore, it follows that the required matrix **B** is $\lambda_1 a_1 a_1' + \lambda_2 a_2 a_2' + \cdots + \lambda_q a_q a_q'$.

This result shows that, if $\boldsymbol{\Sigma}$ is to be approximated by a matrix of order p and rank q, choose the first q terms in the spectral decomposition of $\boldsymbol{\Sigma}$. These correspond to the first q principal components.

## 24.6 PRINCIPAL COMPONENTS FROM THE CORRELATION MATRIX

It is quite common to calculate the principal components of a set of variables after they have been standardized to have unit variance. This means that one is effectively finding the principal components from the correlation matrix **R** rather than from the covariance matrix $\boldsymbol{\Sigma}$. The mathematical derivation is the same, so that the principal components turn out to be in terms of the eigenvectors of **R**. However, it is important to realize the eigenvalues and eigenvectors of **R** will generally not be the same as those of $\boldsymbol{\Sigma}$. Choosing to analyze **R** rather than $\boldsymbol{\Sigma}$ involves a definite but arbitrary decision to make the variables 'equally important'.

For the correlation matrix, the diagonal terms are all unity. Thus the sum of the diagonal terms (or the sum of the variances of the standardized variables) will be equal to p. Thus the sum of the eigenvalues of **R** will also be equal to p, so that the proportion of the total variation accounted for by the $j^{th}$ component is simply $\lambda_i / p$.

**Example 1 (Principal components from Covariance Matrix) :**
Suppose the random variables $X_1$, $X_2$ and $X_3$ have the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

It may be verified that eigenvalue-eigenvector pairs are

$\lambda_1 = 5.83$; $\quad \mathbf{a}_1' = (0.383,\ -0.924,\ 0)$

$\lambda_2 = 2.00$; $\quad \mathbf{a}_2' = (0,\ 0,\ 1)$

$\lambda_3 = 0.17$; $\quad \mathbf{a}_3' = (-0.924,\ 0.383,\ 0)$.

Therefore, the principal components become

$$z_1 = \mathbf{a}_1' \mathbf{x} = 0.383 x_1 - 0.924 x_2$$

$$z_2 = \mathbf{a}_2' \mathbf{x} = x_3$$

$$z_3 = \mathbf{a}_3' \mathbf{x} = -0.924 x_1 + 0.383 x_2 .$$

The variable $x_3$ is one of the principal components because it is uncorrelated with the other two variables.

$$\begin{aligned} \mathrm{Var}(z_1) &= \mathrm{Var}(0.383 x_1 - 0.924 x_2) \\ &= (0.383)^2 \mathrm{Var}(x_1) + (-0.924)^2 \mathrm{Var}(x_2) + 2(0.383)(-0.924) \\ &\quad \mathrm{Cov}(x_1, x_2) \\ &= 0.147(1) + 0.854(5) - 0.708(-2) \\ &= 5.83 = \lambda_1 . \end{aligned}$$

$$\mathrm{Cov}(z_1, z_2) = \mathrm{Cov}(0.383 x_1 - 0.924 x_2, x_3)$$

$$= 0.383\ \mathrm{Cov}(x_1, x_3) - 0.924\ \mathrm{Cov}(x_2\ x_3)$$
$$= 0.383\ (0) - 0.924\ (0) = 0.$$

It is also readily apparent that
$\sigma_{11} + \sigma_{22} + \sigma_{33} = 1 + 5 + 2 = \lambda_1 + \lambda_2 + \lambda_3 = 5.83 + 2.00 + 0.17.$

The proportion of total variance accounted for by the first principal component is $\lambda_1 / (\lambda_1 + \lambda_2 + \lambda_3) = 5.83/8 = 0.73$. Continuing like this, the first two principal components account for a proportion (5.83+2)/8=0.98 of the population variance. In this case the components $z_1$ and $z_2$ could replace the three variables with little loss of information.

$$\mathrm{B} = \lambda_1 a_1 a_1' + \lambda_2 a_2 a_2' .$$

$$\mathrm{B} = \begin{bmatrix} 0.85519687 & -2.0631787 & 0 \\ -2.0631787 & 4.97751408 & 0 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$
$$= \begin{bmatrix} 0.85519687 & -2.0631787 & 0 \\ -2.0631787 & 4.97751408 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$
$$\approx \mathbf{\Sigma}\ \text{(approx)}.$$

That is, a few principal components will approximate covariance structure of *x*'s.

**Example 2 (Principal components from Covariance and Correlation Matrix) :**
Consider the covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix}.$$

and the derived correlation matrix as

$$\mathbf{R} = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}.$$

The eigenvalue-eigenvector pairs from $\mathbf{\Sigma}$ are

$$\lambda_1 = 100.16, \qquad \mathbf{a}_1' = [0.040,\ 0.999]$$
$$\lambda_2 = 0.84, \qquad \mathbf{a}_1' = [0.999,\ -0.040].$$

Similarly, the eigenvalue-eigenvector pairs from **R** are

$$\lambda_1 = 1 + \rho = 1.4, \quad \mathbf{a}_1' = [0.707,\ 0.707]$$
$$\lambda_2 = 1 - \rho = 0.6, \quad \mathbf{a}_2' = [0.707,\ -0.707].$$

The respective principal components using $\mathbf{\Sigma}$ are

$$z_1 = 0.040 x_1 + 0.999 x_2$$
$$z_2 = 0.999 x_1 - 0.040 x_2$$

and using **R,** these are

$$z_1 = 0.707\left[(x_1 - \mu_1)/1\right] + 0.707\left[(x_2 - \mu_2)/10\right]$$
$$= 0.707(x_1 - \mu_1) + 0.707(x_2 - \mu_2)$$
$$z_2 = 0.707(x_1 - \mu_1) + 0.707(x_2 - \mu_2).$$

Because of its large variance, $x_2$ completely dominates the first principal component determined from $\boldsymbol{\Sigma}$. Moreover, this first principal component explains a proportion $\lambda_1/(\lambda_1+\lambda_2) = 100.16/101 = 0.992$ of the total population variance.

When the variables $x_1$ and $x_2$ are standardized, however, the resulting variables contributed equally to the principal component determined from **R**.

In this case, the first principal component explains a proportion $\lambda_1/p = 1.4/2 = 0.70$ of the total (standardized) population variance.

Most strikingly we see that the relative importance of the variables, for instance, the first principal component is greatly affected by the standardization. When the principal component obtained from **R** are expressed in terms of $x_1$ and $x_2$, the relative magnitudes of the weights 0.707 and 0.707 are in direct opposition to those of the weights 0.040 and 0.999 attached to these variables in the principal components obtained from $\boldsymbol{\Sigma}$.

Example 2 demonstrates that the principal components derived from $\boldsymbol{\Sigma}$ are different from those derived from **R**. Furthermore, one set of principal components is not a simple function of the other. This suggests that the standardization has its own consequences. Variables should probably be standardized if they are measured on scales with widely differing ranges or if the measurement units are not commensurate.

## 24.7 ESTIMATING THE PRINCIPAL COMPONENTS

The above derivation of the principal components of **x** assumes that $\boldsymbol{\Sigma}$ is known. Generally we deal with sample data and $\boldsymbol{\Sigma}$ is unknown. In such situations, $\boldsymbol{\Sigma}$ is replaced by **S**, the sample covariance matrix. The derivation of the principal components of **x** using the sample variances and covariances **S** is same as before. Let us denote the eigenvalues of **S** in descending order of size by $\lambda_1, \lambda_2, \cdots, \lambda_p$ and the corresponding eigenvectors by $a_1, a_2, \cdots, a_p$. Since **S** is positive semi-definite symmetric, the eigenvalues are all non-negative and represent the estimated variances of the different components.
If our sample of 'individuals' a random sample from a larger population, then $\{\lambda_i\}$ and $\{a_i\}$ may be regarded as estimates of the eigenvalues and vectors of $\boldsymbol{\Sigma}$, giving us estimates of the principal components of **x**. But no assumptions have been made about the underlying population, and without such assumptions it is impossible to derive the sampling properties of the estimates. If we are prepared to assume that the observations are taken from a multivariate normal distribution, then some sampling theory is available. But this theory is of limited principal value, partly because many of the results are for the asymptotic case (as $n \to \infty$), and partly because the normality assumption is often questionable. In any case, the 'sample' may be observations for a complete population. Thus the modern tendency is to view Principal Component Analysis as a mathematical technique with no underlying statistical model. The principal components obtained from the sample covariance matrix **S** are seen as the principal components and not as estimates of the corresponding quantities obtained from $\boldsymbol{\Sigma}$. The 'hats' over $\lambda_i$ and $a_i$ are often omitted. Indeed, it is not even necessary to regard **x** and **z** as random variables.

**Example 3 (Principal Components from sample data):**
The data is same as given in Example 2 in Johnson and Wichern, 2002. Perspiration from 20 healthy females was analyzed. Three components, $X_1$ = sweat rate, $X_2$ = sodium content and $X_3$ = potassium content were measured and the results are presented in Table 1.

**Table 1: Sweat Data**

| Individual | $X_1$ (sweat rate) | $X_2$ (sodium content) | $X_3$ (potassium content) |
|---|---|---|---|
| 1 | 3.7 | 48.5 | 9.3 |
| 2 | 5.7 | 65.1 | 8.0 |
| 3 | 3.8 | 47.2 | 10.9 |
| 4 | 3.2 | 53.2 | 12.0 |
| 5 | 3.1 | 55.5 | 9.7 |
| 6 | 4.6 | 36.1 | 7.9 |
| 7 | 2.4 | 24.8 | 14.0 |
| 8 | 7.2 | 33.1 | 7.6 |
| 9 | 6.7 | 47.4 | 8.5 |
| 10 | 5.4 | 54.1 | 11.3 |
| 11 | 3.9 | 36.9 | 12.7 |
| 12 | 4.5 | 58.8 | 12.3 |
| 13 | 3.5 | 27.8 | 9.8 |
| 14 | 4.5 | 40.2 | 8.4 |
| 15 | 1.5 | 13.5 | 10.1 |
| 16 | 8.5 | 56.4 | 7.1 |
| 17 | 4.5 | 71.6 | 8.2 |
| 18 | 6.5 | 52.8 | 10.9 |
| 19 | 4.1 | 44.1 | 11.2 |
| 20 | 5.5 | 40.9 | 9.4 |

The variance-covariance matrix of the above data is

$$\boldsymbol{\Sigma} = \begin{bmatrix} 2.879368 & 10.01 & -1.80905 \\ 10.01 & 199.7884 & -5.64 \\ -1.80905 & -5.64 & 3.627658 \end{bmatrix}.$$

Now find the eigenvalues and eigenvectors of the above matrix. Arrange the eigenvalues in decreasing order. Let the eigenvalues in decreasing order and corresponding eigenvectors are

$$\lambda_1 = 200.462 \quad a_1 = (0.0508, 0.9983, -0.0291)$$
$$\lambda_2 = 4.532 \quad a_2 = (-0.5737, 0.0530, \quad 0.8173)$$
$$\lambda_3 = 1.301 \quad a_3 = (0.8175, -0.0249, 0.5754).$$

The principal components for this data are

$$z_1 = 0.0508x_1 + 0.9983x_2 - 0.0291x_3$$
$$z_2 = -0.5737x_1 + 0.0530x_2 + 0.8173x_3$$
$$z_3 = 0.8175x_1 - 0.0249x_2 + 0.5754x_3.$$

The variance of principal components will be eigenvalues i.e.

$$\mathrm{Var}(z_1) = 200.462, \quad \mathrm{Var}(z_2) = 4.532, \quad \mathrm{Var}(z_3) = 1.301.$$

The total variation explained by principal components is

$$\lambda_1 + \lambda_2 + \lambda_3 = 200.462 + 4.532 + 1.301 = 206.295.$$

As such, it can be seen that the total variation explained by principal components is same as that explained by original variables. It could also be proved mathematically as well as empirically that the principal components are uncorrelated.

The proportion of total variation accounted for by the first principal component is

$$\frac{\lambda_1}{\lambda_1+\lambda_2+\lambda_3}=\frac{200.462}{206.295}=0.9717 \text{ of the total variance,}$$

Where as, the first two components account for a proportion

$$\frac{\lambda_1+\lambda_2}{\lambda_1+\lambda_2+\lambda_3}=\frac{204.994}{206.295}=0.9937 \text{ of it.}$$

Hence, in further analysis, the first or first two principal components $z_1$ and $z_2$ could replace three variables by sacrificing negligible information about the total variation in the system. The scores of principal components can be obtained by substituting the values of $x_i$'s in the equations of $z_i$'s. For above data, the scores from the first two principal components for first observation i.e. for first individual are

$$z_1 = 0.0508\times 3.7 + 0.9983\times 48.5 - 0.0291\times 9.3$$

$$z_2 = -0.5737\times 3.7 + 0.0530\times 48.5 + 0.8173\times 9.3$$

Similarly principal component scores for other individuals can be obtained. Thus the whole data with three variables can be converted to a new data set with two principal components.

Now try the following exercise.

---

**E1):** Weather data given below pertains to Raipur district (M.P.) from 1970 to 1986 for kharif crop season from 21st May to 7th October. The weather variables are average minimum temperature ($x_1$), average relative humidity at 8 hrs ($x_2$), average relative humidity at 14 hrs ($x_3$) and total rainfall in cm ($x_4$).

| S.No. | $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|---|
| 1 | 25 | 86 | 66 | 186.49 |
| 2 | 24.9 | 84 | 66 | 124.34 |
| 3 | 25.4 | 77 | 55 | 98.79 |
| 4 | 24.4 | 82 | 62 | 118.88 |
| 5 | 22.9 | 79 | 53 | 71.88 |
| 6 | 7.7 | 86 | 60 | 111.96 |
| 7 | 25.1 | 82 | 58 | 99.74 |
| 8 | 24.9 | 83 | 63 | 115.2 |
| 9 | 24.9 | 82 | 63 | 100.6 |
| 10 | 24.9 | 78 | 56 | 62.38 |
| 11 | 24.3 | 85 | 67 | 154.4 |
| 12 | 24.6 | 79 | 61 | 112.71 |
| 13 | 24.3 | 81 | 58 | 79.63 |
| 14 | 24.6 | 81 | 61 | 125.59 |
| 15 | 24.1 | 85 | 64 | 99.87 |
| 16 | 24.5 | 84 | 63 | 143.56 |
| 17 | 24 | 81 | 61 | 114.97 |
| **Mean** | **23.56** | **82.06** | **61** | **112.97** |
| **S.D.** | **4.13** | **2.75** | **3.97** | **30.06** |

Find the principal components using variance covariance matrix and correlation matrix.

E2) An experiment was conducted to compare the drying methods of microwave assisted convective drying with those of conventional methods such as microwave, convective and freeze drying of banana. The data obtained for

overall drying rate ( $x_1$ ), rehydration ratio ( $x_2$ ), total sugars ( $x_3$ ), total carbohydrates ( $x_4$ ) and energy use efficiency ( $x_5$ ).

| **odr( $x_1$ )** | **rr( $x_2$ )** | **ts( $x_3$ )** | **tc( $x_4$ )** | **ee( $x_5$ )** |
|---|---|---|---|---|
| 1.208 | 1.936 | 47.765 | 555.980 | 19.539 |
| 0.806 | 1.896 | 47.774 | 557.186 | 13.026 |
| 0.773 | 1.908 | 48.083 | 562.236 | 12.505 |
| 1.706 | 1.810 | 48.521 | 568.761 | 22.436 |
| 1.381 | 1.743 | 48.992 | 574.829 | 13.225 |
| 1.137 | 1.718 | 49.843 | 583.223 | 8.564 |
| 1.706 | 1.698 | 49.299 | 581.713 | 20.521 |
| 1.611 | 1.766 | 50.652 | 595.202 | 12.817 |
| 1.381 | 1.690 | 50.417 | 594.310 | 8.206 |
| 2.231 | 1.603 | 49.905 | 591.923 | 24.725 |
| 1.758 | 1.754 | 50.533 | 598.169 | 11.957 |
| 1.526 | 1.759 | 50.368 | 597.298 | 7.491 |
| 2.762 | 2.130 | 50.103 | 584.927 | 35.729 |
| 2.522 | 2.082 | 50.807 | 591.597 | 32.622 |
| 1.289 | 2.200 | 51.602 | 602.573 | 16.673 |
| 2.900 | 1.839 | 50.899 | 597.215 | 31.696 |
| 2.320 | 1.986 | 51.229 | 600.325 | 19.353 |
| 2.071 | 2.000 | 52.031 | 610.402 | 13.972 |
| 3.625 | 1.770 | 52.546 | 618.794 | 36.769 |
| 3.222 | 1.656 | 51.872 | 616.357 | 22.826 |
| 3.222 | 1.889 | 52.706 | 621.384 | 17.537 |
| 3.625 | 1.683 | 53.323 | 629.932 | 34.300 |
| 3.412 | 1.669 | 51.674 | 612.174 | 21.002 |
| 3.222 | 1.858 | 52.783 | 625.333 | 14.699 |
| 4.833 | 2.234 | 52.421 | 613.470 | 52.104 |
| 4.143 | 2.315 | 53.223 | 622.417 | 44.661 |
| 3.625 | 2.405 | 53.925 | 629.653 | 39.078 |
| 5.273 | 2.368 | 53.566 | 629.317 | 49.299 |
| 4.462 | 2.414 | 54.632 | 642.673 | 32.966 |
| 4.143 | 2.513 | 54.283 | 638.949 | 25.305 |
| 5.273 | 2.785 | 54.300 | 639.864 | 46.232 |
| 5.273 | 2.572 | 55.377 | 652.587 | 33.665 |
| 5.273 | 2.660 | 55.181 | 650.602 | 26.470 |
| 6.444 | 2.812 | 54.685 | 646.854 | 53.196 |
| 5.800 | 2.682 | 54.799 | 648.481 | 32.601 |
| 5.800 | 2.692 | 54.808 | 648.804 | 24.715 |
| 1.209 | 1.945 | 47.686 | 554.018 | 19.550 |
| 0.804 | 1.903 | 47.919 | 559.000 | 13.001 |
| 0.776 | 1.900 | 48.000 | 560.832 | 12.548 |
| 1.708 | 1.800 | 48.864 | 570.624 | 22.459 |
| 1.379 | 1.723 | 49.000 | 576.062 | 13.214 |
| 1.140 | 1.701 | 49.941 | 585.406 | 8.573 |
| 1.704 | 1.662 | 49.412 | 581.003 | 20.504 |
| 1.615 | 1.800 | 50.427 | 592.981 | 12.832 |
| 1.384 | 1.713 | 50.666 | 596.324 | 8.213 |
| 2.228 | 1.602 | 50.000 | 589.258 | 24.704 |
| 1.760 | 1.786 | 50.026 | 596.871 | 11.964 |
| 1.523 | 1.800 | 50.369 | 599.000 | 7.486 |
| 2.766 | 2.119 | 50.232 | 583.048 | 35.782 |
| 2.549 | 2.103 | 51.000 | 590.005 | 32.974 |
| 1.287 | 2.224 | 51.805 | 604.000 | 16.649 |
| 2.896 | 1.799 | 51.000 | 594.961 | 31.659 |
| 3.322 | 2.000 | 51.134 | 603.426 | 24.026 |
| 2.074 | 2.036 | 52.000 | 606.924 | 13.981 |

| | | | | |
|---|---|---|---|---|
| 3.622 | 1.780 | 52.500 | 615.885 | 36.745 |
| 3.226 | 1.686 | 51.919 | 615.109 | 22.841 |
| 3.219 | 1.923 | 52.669 | 624.009 | 17.530 |
| 3.622 | 1.700 | 53.419 | 631.008 | 34.279 |
| 3.410 | 1.700 | 51.428 | 612.887 | 20.997 |
| 3.224 | 1.901 | 52.699 | 628.051 | 14.702 |
| 4.830 | 2.240 | 52.664 | 615.008 | 52.068 |
| 4.140 | 2.299 | 52.172 | 624.097 | 44.630 |
| 3.629 | 2.400 | 54.000 | 630.059 | 39.121 |
| 5.270 | 2.338 | 53.316 | 626.841 | 49.277 |
| 4.460 | 2.451 | 54.761 | 640.984 | 32.958 |
| 4.139 | 2.497 | 54.399 | 640.019 | 25.291 |
| 5.278 | 2.800 | 54.418 | 641.216 | 46.269 |
| 5.276 | 2.586 | 55.400 | 650.000 | 33.677 |
| 5.270 | 2.558 | 55.280 | 652.189 | 26.463 |
| 6.446 | 2.830 | 54.806 | 649.057 | 53.206 |
| 5.780 | 2.598 | 54.900 | 646.189 | 32.542 |
| 5.770 | 2.700 | 54.607 | 650.007 | 24.664 |

Obtain principal components using both covariance matrix and correlation matrix and interpret the interpret the results.

E3) Find the principal components, and the proportion of total population variance explained by each, when the covariance matrix is

$$\sum = \begin{bmatrix} \sigma^2 & \sigma^2\rho & 0 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ 0 & \sigma^2\rho & \sigma^2 \end{bmatrix}, \quad -\frac{1}{\sqrt{2}} < \rho < \frac{1}{\sqrt{2}}.$$

E4) Consider the census-tract data listed below. Suppose the observations on $X_5$ = median value home were recorded in thousands, rather than ten thousands of Rupees; that is, multiply all the numbers listed in the sixth column of the data Table by 10.

| **Tract** | **Total Population (thousands)** | **Median School Years** | **Total employment (thousands)** | **Health services employment (hundreds)** | **Median Value home (Rs.10,000s)** |
|---|---|---|---|---|---|
| 1 | 5.935 | 14.2 | 2.265 | 2.27 | 2.91 |
| 2 | 1.52 | 13.1 | .597 | .75 | 2.62 |
| 3 | 32.599 | 12.7 | 1.237 | 1.11 | 1.72 |
| 4 | 4.009 | 15.2 | 1.649 | .81 | 3.02 |
| 5 | 4.687 | 14.7 | 2.312 | 2.50 | 2.22 |
| 6 | 8.044 | 15.6 | 3.641 | 4.51 | 2.36 |
| 7 | 2.766 | 13.3 | 1.244 | 1.03 | 1.97 |
| 8 | 6.538 | 17.0 | 2.618 | 2.39 | 1.85 |
| 9 | 6.451 | 12.9 | 3.147 | 5.52 | 2.01 |
| 10 | 3.314 | 12.2 | 1.606 | 2.18 | 1.82 |
| 11 | 3.777 | 13.0 | 2.119 | 2.83 | 1.80 |
| 12 | 1.530 | 13.8 | 0.789 | 0.84 | 4.25 |
| 13 | 2.768 | 13.6 | 1.336 | 1.75 | 2.64 |
| 14 | 6.585 | 14.9 | 2.763 | 1.91 | 3.17 |

(a) Construct the sample covariance matrix, **S**, for the census-tract data when $X_5$= median value home is recorded in thousands of Rupees.

(b) Obtain the eigenvalue-eigenvector pairs and take first two sample principal components for the covariance matrix in (a).

(c) Compute the proportion of total variance explained by the first two principal components obtained in (b).

E5) The following data table was published in Time Magazine in January, 1996. It gives the beer, wine and liquor consumption (in liters per year), life expectancy (in years) and heart disease rate (in cases per 100,000 per year) for the 10 countries listed.

| THE WINE DATA SET | | | | | |
|---|---|---|---|---|---|
| **Country** | **Liquor** | **Wine** | **Beer** | **Life Expectancy in years** | **Heart disease rate** |
| France | 2.5 | 63.5 | 40.1 | 78 | 61.1 |
| Italy | 0.9 | 58.0 | 25.1 | 78 | 94.1 |
| Switzerland | 1.7 | 46.0 | 65.0 | 78 | 106.4 |
| Australia | 1.2 | 15.7 | 102.1 | 78 | 173.0 |
| Great Britain | 1.5 | 12.2 | 100.0 | 77 | 199.7 |
| United States | 2.0 | 8.9 | 87.8 | 76 | 176 |
| Russia | 3.8 | 2.7 | 17.1 | 69 | 373.6 |
| Czech Republic | 1.0 | 1.7 | 140.0 | 73 | 283.7 |
| Japan | 2.1 | 1.0 | 55.0 | 79 | 34.7 |
| Mexico | 0.8 | 0.2 | 50.4 | 73 | 36.4 |

(a) Construct the sample covariance matrix, **S**, for the census-tract data when $X_5$= median value home is recorded in thousands of dollars.
(b) Obtain the eigenvalue-eigenvector pairs and the first two sample principal components for the covariance matrix in (a).
(c) Compute the proportion of total variance explained by the first two principal components obtained in (b).
(d) Obtain the principal component score for the first two principal components.
(e) Standardized the data with mean 0 and variance 1 and repeat the steps (a) to (d) on the transformed data.

Now, let us discuss scaling convention in principal component analysis in the following section.

## 24.8 SCALING CONVENTION IN PCA

It is important to realize that the principal components of a set of variables depend critically upon the scales used to measure the variables. For example, suppose that for each of n individuals we measure their weight in pounds, their height in feet and their age in years to give a vector **x**. Denote the resulting sample covariance matrix by $S_x$, its eigenvalues by $\{\lambda_i\}$ and its eigenvectors by $\{a_i\}$.

If we transform to new co-ordinates so that $\mathbf{u}'$ = (weight in kilograms, height in metres, age in months). Then $\mathbf{u} = \mathbf{Kx}$, where **K** is the diagonal matrix

$$\mathbf{K} = \begin{bmatrix} 1/2.2 & 0 & 0 \\ 0 & 1/3.28 & 0 \\ 0 & 0 & 12 \end{bmatrix}.$$

As there are, for example, 2.2 pounds in one kilogram, 3.28 feet in one meter and 12 months in a year. The covariance matrix of the new variables will be given by

$$S_u = KS_xK,$$

since $\mathbf{K}' = \mathbf{K}$. The eigenvalues and eigenvectors of $S_u$ will generally be different from those of $S_x$ and will be denoted by $\{\lambda_i^*\}$ and $\{a_i^*\}$. But will they give the same

principal components when transformed back to the original variables? The answer generally is no. The principal components will change unless:

**a)** All the diagonal element of **K** are the same, so that $K = cI$, where c is a scalar constant. This would mean that all the variables are scaled in the same way.

**b)** Variables corresponding to unequal diagonal elements of **K** are uncorrelated. In particular, if all the elements of **K** are unequal and all variables are uncorrelated, then $\mathbf{S}_x$ must be a diagonal matrix, in that case there is no point in carrying out PCA.

The practical outcome of the above result is that principal component are generally changed by scaling and that they are, therefore, not a unique characteristic of the data. If, for example, one variable has a much larger variance than all the other variables, then this variable will dominate the first principal component of the covariance matrix whatever be the correlation structure, whereas if the variables are all scaled to have unit variance, then the first principal component will be quite different in kind. Because of this, it is generally thought to be of little importance in carrying out a PCA unless the variables have 'roughly similar' variance, as may be the case, for example, if all the variables are percentage, or are measured in the same co-ordinates.

The conventional way of getting rid of the scaling problem is to analyze the correlation matrix rather than the covariance matrix, so that each multivariate observation, **x**, is transformed by

$$\mathbf{u} = \mathbf{K}\mathbf{x}$$

where

$$\mathbf{K} = \begin{bmatrix} 1/s_1 & \cdots & 0 & \cdots & 0 \\ 0 & \cdots & 1/s_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & \cdots & 0 & \cdots & 1/s_p \end{bmatrix}.$$

and $s_i$ is the sample standard deviation for the $i^{th}$ variable. This ensures that all variables are scaled to have unit variance and so in some sense have equal importance. This scaling procedure is still arbitrary to some extent, is data-dependent and avoids rather than solves the scaling problem. If the variables are not thought to be of equal importance, then the analysis of the correlation matrix is not recommended. Analysing the correlation matrix also makes it more difficult to compare the results of Principal Component Analysis on two or more different samples.

Note that the principal components of the correlation matrix will not be orthogonal if the variables are transformed back to their original co-ordinates. This is because a linear transformation of two lines at right angles in Euclidean space will not generally give two new lines at right angles, which is another way of explaining why the scaling problem arise in the first place.

**Example 4:** Using the data given in E 1), discuss the scaling convention in PCA.
**Solution:** For the data given in E1), $x_1$ is measured in $^0$C, $x_2$ and $x_3$ in terms of percentage and $x_4$ in cm. Then the sample covariance matrix $S_x = \Sigma$ and eigenvalue-eigenvector pairs $\{\lambda_i, a_i'\}$, i =1,2,3,4, have been obtained earlier

If we transform $x$ to new coordinates u by means of the transformation

$$\mathbf{u} = \mathbf{K}\,\mathbf{x}$$

with $\mathbf{u}'$=($x_1$ in F−32, $x_2$ in %, $x_3$ in %, $x_4$ in mm), then

$$\mathbf{K} = \begin{bmatrix} 9/5 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}$$

as we have the relation

$F = (9/5)^{o}C + 32$ and 10 mm=1 cm.

The covariance matrix of the new variables is

$$\mathbf{S}_u = \mathbf{K}\,\mathbf{S}_x\mathbf{K} \text{ as } \mathbf{K}' = \mathbf{K}.$$

We get the eigenvalue-eigenvector pairs of $\mathbf{S}_u$ as

$\lambda_1^* = 90400.807$ $\mathbf{a}_1'^* = (0.001,\ 0.006,\ 0.010,\ 1)$

$\lambda_2^* = 56.300$ $\mathbf{a}_2'^* = (0.988,\ 0.150,\ 0.027,\ 0)$

$\lambda_3^* = 8.061$ $\mathbf{a}_3'^* = (0.055,\ 0.521,\ 0.852,\ -0.012)$

$\lambda_4^* = 1.106$ $\mathbf{a}_4'^* = (0.142,\ 0.840,\ -0.678,\ 0).$

These are not the same as $(\lambda_i, a_i'),\ i = 1,2,3,4.$

Seeing the matrix **K**, we find that

(1) At least one of the diagonal elements of **K** is different from other (in our case 3 out of 4 are different).

(2) Variables corresponding, to unequal diagonal elements of **K** are correlated.

Hence, the eigenvalue-eigenvector pairs of $S_u$ give different principal components when transformed back to original **x's** Principal Components therefore do not depict the unique characteristic of the original data due to scaling problem.

The conventional way of getting round the scaling problem is to analyze the correlation matrix rather than the covariance matrix, so that **x** is transformed to $\mathbf{u} = \mathbf{K}_1\mathbf{x}$, where

$$\mathbf{K}_1 = \begin{bmatrix} 1/4.13 & 0 & 0 & 0 \\ 0 & 1/2.75 & 0 & 0 \\ 0 & 0 & 1/3.97 & 0 \\ 0 & 0 & 0 & 1/30.06 \end{bmatrix}.$$

The diagonals being, inverses of the sample Standard deviation of the $i^{th}$ variable and then carrying out Principal Component Analysis.

## 24.9 GEOMETRICAL INTERPRETATION OF PRINCIPAL COMPONENTS

The transformation $\mathbf{Z} = \mathbf{A}'\mathbf{x}$ is an orthogonal transformation and it transforms an ellipsoid $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} = \text{Constant}$ into the ellipsoid $\mathbf{Z}'\boldsymbol{\Lambda}^{-1} = \text{Constant}$, i.e.

$$\sum_{i=1}^{p} \frac{z_i^2}{\lambda_i} = \text{Constant}.$$

This is an ellipsoid referred to its principal axes and $\sqrt{\lambda_i}$ (i = 1,2, ..., p) are the lengths of principal semi-axes. For p = 2, i.e., when **x**'s can be replaced by two

principal components, this is an ellipse in a coordinate system with axes $z_1$ and $z_2$ lying in the directions of $\mathbf{a}_1$ and $\mathbf{a}_2$ respectively as depicted in the figure below.

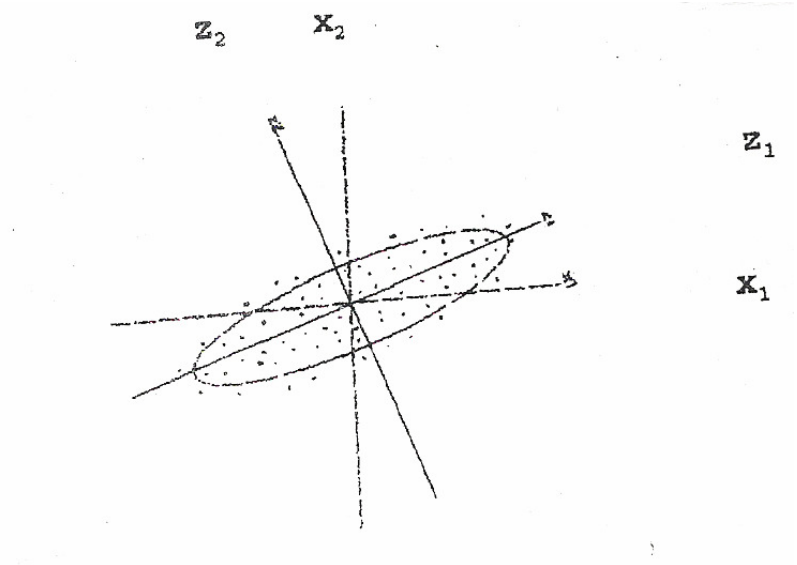


**Fig. 1: Geometrical Interpretation of Principal Components**

If the eigenvalues of $\Sigma$ are equal, the new ellipse is a circle and there is unique transformation. In fact, this only happens if the original variables are uncorrelated and have equal variance.

## 24.10 APPLICATIONS OF PRINCIPAL COMPONENTS

i) An important benefit of Principal Component Analysis is that it provides a quick way of assessing the effective dimensionality of a set of data. If the first few components account for most of the variation in the original data, it is often a good idea to use these first few components scores in subsequent analysis.

ii) Plotting of data becomes difficult with more than three variables. But if the first two components account for a large proportion of the total variation, then it will often be useful to plot the values of the first two components scores for each individual. In other words, Principal Component Analysis may enable us to plot the data in two dimensions. This procedure is known as Biplot. A biplot is a scatter plot that graphically displays both the row factors and the column factors of a two-way data. The concept of biplots were first developed by Gabriel (1971). Since then, biplot has been used in data visualization and pattern analysis in various research fields, from psychology to economics to agronomy. Currently there are over 50000 web pages containing the keyword 'biplot'. This technique has extensively been used in the analysis of multi-environment trials. For generating a biplot, we take the matrix representing the effects of two factors. This matrix is then subjected to singular value decomposition.

iii) In particular, one can then look for outliers or for groups or 'clusters' of individuals. This is an important use of Principal Component Analysis and often groupings of variables which would not be found by other means.

iv) Reduction is dimensionality can help in discriminant analysis when a large number of correlated variables (p) are observed on *n* observation (n<p). As the number of observations is less than the number of variables, there will be unpleasant singularity problem unless the dimensionality is drastically reduced. Hopefully, Principal Component Analysis does that.

v) Multiple regression can be dangerous if the so-called independent variables are in fact highly correlated. Various techniques, such as ridge regression, have been developed to overcome this problem. An alternative approach, which may

well be more fruitful, is to regress, not on the original variables, but on those components which are most highly correlated with the dependent variable. This is called principal components regression.

It is quite likely that first few principal components account for most of the variability in the original data. If so, these few principal components can then replace the initial p variables in subsequent analysis, thus reducing the effective dimensionality of the problem. An analysis of principal components often reveals relationships that were not previously suspected and thereby allows interpretation that would not ordinarily result.

However, Principal Components Analysis is more of a mean to an end rather than end in itself because this frequently serves as intermediate steps in much larger investigations by reducing the dimensionality of the problem and providing easier interpretation. It is a mathematical technique, which does not require user to specify the statistical model or assumption about distribution of original variates. It may also be mentioned that principal components are artificial variables and often it is not possible to assign physical meaning to them. Further, since Principal Components Analysis transforms original set of variables to new set of uncorrelated variables. It is worth stressing that if the original variables are uncorrelated, then there is no point in carrying out the Principal Components Analysis. It is important to note here that the principal components depend on the scale of measurement. Conventional way of getting rid of this problem is to use the standardized variables with unit variances.

Now, let us summarize the unit.

## 24.11 SUMMARY

In this Unit, we covered the following points.

1. 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.
2. Principal Components can be derived from covariance matrix or correlation matrix.
3. For obtaining principal components, one must know the eigenvalues of the sample covariance/ correlation matrix.
4. Principal component scores are obtained by multiplying the row vector of variables with the column eigenvectors.
5. Principal components of a set of variables depend critically upon the scales used to measure the variables.

## 24.12 SOLUTIONS/ANSWERS

E1) The problem is to find first few principal components, which may account for most of the variation in the original data.

**I. Identification of important principal components through covariance matrix of the original variables.**

The variables $\mathbf{X}' = (x_1, x_2, x_3, x_4)$ have the covariance matrix as

$$\boldsymbol{\Sigma} = \begin{bmatrix} 17.02 & -4.12 & 1.54 & 5.14 \\ & 7.56 & 8.50 & 54.82 \\ & & 15.75 & 92.95 \\ & & & 903.87 \end{bmatrix}.$$

The eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}$ are

$\lambda_1 = 916.902 \quad \mathbf{a}_1' = (0.006, \quad 0.061, \quad 0.103, \quad 0.993)$

$\lambda_2 = 18.375 \quad \mathbf{a}_2' = (0.955, \quad -0.296, \quad 0.011, \quad 0.012)$

$\lambda_3 = 7.87 \quad \mathbf{a}_3' = (0.141, \quad 0.485, \quad 0.855, \quad -0.119)$

$\lambda_4 = 1.056 \quad \mathbf{a}_4' = (0.260, \quad 0.820, \quad -0.509, \quad .001).$

The principal component are $\mathbf{Z}_i = \mathbf{a}_i' \mathbf{x}, i = 1,2,3,4.$

The proportion of total variance accounted for by the first principal component is

$$\frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{916.902}{944.203} = 0.97.$$

Continuing, the first two components account for proportion

$$\frac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4} = \frac{935.277}{944.203} = 0.99$$

of the total variance. Hence the components $\mathbf{z}_1 = \mathbf{a}_1' \mathbf{x}$ and $\mathbf{z}_2 = \mathbf{a}_2' \mathbf{x}$ could replace the four variables with negligible loss of information

$$\mathbf{B} = \lambda_1 \mathbf{a}_1 \mathbf{a}_1' + \lambda_2 \mathbf{a}_2 \mathbf{a}_2'$$

$$= \begin{bmatrix} 16.79 & -4.86 & 0.76 & 5.67 \\ & 5.02 & 5.70 & 55.47 \\ & & 9.73 & 93.78 \\ & & & 904.11 \end{bmatrix}$$

$\approx \boldsymbol{\Sigma}$ (approximately)

$\Rightarrow$ a few Principal Components will approximate the covariance structure of x,s

**II. Principal components from the correlation matrix**

The correlation matrix for $\mathbf{x}' = (x_1, x_2, x_3, x_4)$ is

$$\mathbf{R} = \begin{bmatrix} 1 & -0.36 & 0.09 & 0.04 \\ & 1 & 0.78 & 0.66 \\ & & 1 & 0.78 \\ & & & 1 \end{bmatrix}.$$

The eigenvalue-eigenvector pairs of $\mathbf{R}$ are

$\lambda_1 = 2.493 \quad \mathbf{a}_1' = (-0.088, \quad 0.577, \quad 0.588, \quad 0.560)$

$\lambda_2 = 1.137 \quad \mathbf{a}_2' = (0.919, \quad -0.269, \quad 0.214, \quad 0.195)$

$\lambda_3 = 0.270 \quad \mathbf{a}_3' = (0.200, \quad 0.412, \quad 0.388, \quad -0.800)$

$\lambda_4 = 0.100 \qquad \mathbf{a}_4' = (0.329, \quad 0.652, \quad -0.677, \quad 0.090).$

The proportion of total variance accounted for by the first principal component is $\lambda_1 / p = 0.62$ and by the first two components is $(\lambda_1 + \lambda_2)/p = 0.91$.

Hence, the component $\mathbf{z}_1$ and $\mathbf{z}_2$ could replace the four $x_i$'s with little loss of information. Here we get

$$\mathbf{B} = \lambda_1 a_1 a_1' + \lambda_2 a_2 a_2' \text{ as}$$

$$= \begin{bmatrix} 0.980 & -0.410 & 0.095 & 0.081 \\ & 0.912 & 0.780 & 0.750 \\ & & 0.914 & 0.868 \\ & & & 0.825 \end{bmatrix}$$

$\approx \mathbf{R}$ (approximately).

Using $\mathbf{\Sigma}$, the important Principal Components are

$$z_1 = 0.006x_1 + 0.061x_2 + 0.103x_3 + 0.993x_4$$
$$z_2 = 0.955x_1 + -0.296x_2 + 0.011x_3 + 0.012x_4 .$$

Using $\mathbf{R}$, the important Principal Components are

$$z_1 = -0.088x_1 + 0.577x_2 + 0.588x_3 + 0.563x_4$$
$$z_2 = 0.919x_1 + -0.269x_2 + 0.214x_3 + 0.195x_4 .$$

Because of its large variance of 903.87, $x_4$ completely dominates the first Principal Component determined from $\mathbf{\Sigma}$. Moreover, this Principal Component explains a proportion 0.97 of the total variance.

When the variables are standardized, however, the resulting variables contribute equally to the Principal Components determined from **R**. In this case, the first principal component explains only a proportion 0.62 of the total (standardized) variance.

**E2)** Compute Variance-Covariance matrix of the above data

| **Variance -Covariance Matrix** | | | | | |
|---|---|---|---|---|---|
| | **odr** | **rr** | **ts** | **tc** | **ee** |
| **odr** | 2.6137021 | 0.4696316 | 3.2294105 | 40.1744637 | 16.0029870 |
| **rr** | 0.4696316 | 0.1353220 | 0.6144554 | 7.2058641 | 2.8985903 |
| **ts** | 3.2294105 | 0.6144554 | 4.9444081 | 61.0660178 | 16.5357453 |
| **tc** | 40.1744637 | 7.2058641 | 61.0660178 | 770.9029016 | 190.3575371 |
| **ee** | 16.0029870 | 2.8985903 | 16.5357453 | 190.3575371 | 164.1281505 |

Obtain eigenvalue and eigenvectors of the matrix obtained and arrange the eigenvalues and eigenvectors in descending order of eigenvalues.

| | Eigenvalue |
|---|---|
| **1** | 832.839971 |
| **2** | 109.535336 |
| **3** | 0.215563 |
| **4** | 0.106023 |
| **5** | 0.027591 |

| **Eigenvectors** | | | | | |
|---|---|---|---|---|---|
| | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** |
| **odr** | 0.051913 | 0.039006 | 0.964708 | 0.048212 | -.250595 |
| **rr** | 0.009324 | 0.007244 | 0.184747 | 0.543387 | 0.818816 |
| **ts** | 0.076287 | -.009233 | -.179117 | 0.835033 | -.514521 |
| **tc** | 0.956806 | -.277839 | -.025543 | -.069282 | 0.043302 |
| **ee** | 0.275537 | 0.959764 | -.049719 | -.018084 | 0.011590 |

Therefore, the principal components are

pc1 = 0.051913*odr + 0.009324*rr + 0.076287*ts + 0.956806*tc + 0.275537*ee;

pc2 = 0.039006*odr + 0.007244*rr - 0.009233*ts - 0.277839*tc + 0.959764*ee;

pc3 = 0.964708*odr + 0.184747*rr - 0.179117*ts - 0.025543*tc - 0.049719*ee;

pc4 = 0.048212*odr + 0.543387*rr + 0.835033*ts - 0.069282*tc - 0.018084*ee;

pc5 = -0.250595*odr + 0.818816*rr - 0.5145218ts + 0.043302*tc + 0.011590*ee;

Obtain Total variance explained by all the 5 principal components

Total variance = 832.839971+109.535336+0.215563+0.106023+0.027591
=942.7245

Obtain Proportion of Total variance explained by each principal component and cumulative variance explained

| | **Eigenvalue** | **Proportion Explained by Principal components** | **Cumulative proportion of variance explained** |
|---|---|---|---|
| **1** | 832.839971 | 0.8834 | 0.8834 |
| **2** | 109.535336 | 0.1162 | 0.9996 |
| **3** | 0.215563 | 0.0002 | 0.9999 |
| **4** | 0.106023 | 0.0001 | 1.0000 |
| **5** | 0.027591 | 0.0000 | 1.0000 |

One can easily see that first two principal components explain 99.96% of the total variance.

**E3)** Obtain eigenvalues of $\Sigma$. These are $\sigma^2(1+\sqrt{2}\rho)$, $\sigma^2(1-\sqrt{2}\rho)$ and 1. Now proportion of total population variance explained by each principal component can be obtained easily.

**E4)** (a) Multiply median value home by 10 to get median value home in thousand dollars. Now obtain variance-covariance matrix for Total Population (totalpop), median school years (schoolyear), total employment (totalemp), health services employment (healthemp) and median value home in thousand dollars (valhometh)

| **Covariance Matrix** | | | | | |
|---|---|---|---|---|---|
| | **totalpop** | **schoolyear** | **totalemp** | **healthemp** | **valhometh** |
| **totalpop** | 60.63731038 | -1.34907582 | 0.15584659 | -0.29379681 | -19.48018132 |
| **schoolyear** | -1.34907582 | 1.76747253 | 0.58817473 | 0.17797802 | 1.75549451 |
| **totalemp** | 0.15584659 | 0.58817473 | 0.80227255 | 1.06574978 | -1.59582198 |
| **healthemp** | -0.29379681 | 0.17797802 | 1.06574978 | 1.96947473 | -3.56806593 |
| **valhometh** | -19.48018132 | 1.75549451 | -1.59582198 | -3.56806593 | 50.43802198 |

(b) Obtain eigenvalues and eigenvectors of the variance covariance matrix obtained in (a)

| | **Eigenvalues** |
|---|---|
| **1** | 75.8040303 |
| **2** | 35.7208028 |
| **3** | 2.6198422 |
| **4** | 1.4114582 |
| **5** | 0.0584187 |

| **Eigenvectors** | | | | | |
|---|---|---|---|---|---|
| | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** |
| **totalpop** | 0.789183 | 0.612993 | 0.034794 | -.010785 | -.010123 |
| **schoolyear** | -.028727 | 0.015099 | 0.556550 | 0.790274 | -.254293 |
| **totalemp** | 0.014830 | -.035568 | 0.512331 | -.084884 | 0.853713 |
| **healthemp** | 0.026615 | -.089269 | 0.650409 | -.601568 | -.454319 |
| **valhometh** | -.612729 | 0.784077 | 0.059372 | -.079127 | -.000188 |

**(c) Compute the proportion of variance explained by first two components as:**

a. Variance explained by first principal component = ratio of first eigenvalue to that of sum of all eigenvalues; 0.6557
b. Variance explained by first two principal components = ratio of sum of first two eigenvalues to that of sum of all eigenvalues; 0.9646

**E5)** (a) Now obtain variance-covariance matrix for liquor, wine, beer, life expectancy in years (LifeExpyears), heart disease rate (Heartdisrate)

| Covariance Matrix | | | | | |
|---|---|---|---|---|---|
| | **Liquor** | **Wine** | **Beer** | **LifeExpyears** | **Heartdisrate** |
| **Liquor** | 0.83389 | -1.00167 | -16.92111 | -1.09444 | 43.34611 |
| **Wine** | -1.00167 | 621.35656 | -375.77600 | 41.99889 | -1090.38144 |
| **Beer** | -16.92111 | -375.77600 | 1495.50711 | 11.88444 | 1070.23089 |
| **LifeExpyears** | -1.09444 | 41.99889 | 11.88444 | 10.32222 | -249.84778 |
| **Heartdisrate** | 43.34611 | -1090.38144 | 1070.23089 | -249.84778 | 12280.66678 |

(b) Obtain eigenvalues and eigenvectors of the variance covariance matrix obtained in (a)

| | **Eigenvalue** |
|---|---|
| **1** | 12497.0853 |
| **2** | 1465.8137 |
| **3** | 442.5038 |
| **4** | 2.9614 |
| **5** | 0.3224 |

| Eigenvectors | | | | | |
|---|---|---|---|---|---|
| | **Prin1** | **Prin2** | **Prin3** | **Prin4** | **Prin5** |
| **Liquor** | 0.003310 | -.014450 | -.006519 | 0.137392 | 0.990384 |
| **Wine** | -.094152 | -.267864 | 0.957149 | -.056040 | 0.010481 |
| **Beer** | 0.099531 | 0.955489 | 0.275040 | -.032985 | 0.019994 |
| **LifeExpyears** | -.020038 | 0.020858 | 0.063218 | 0.988225 | -.136305 |
| **Heartdisrate** | 0.990362 | -.121022 | 0.064654 | 0.017523 | -.007081 |

(c) Compute the proportion of variance explained by first two components as:

c. Variance explained by first principal component = ratio of first eigenvalue to that of sum of all eigenvalues; 0.8673
d. Variance explained by first two principal components = ratio of sum of first two eigenvalues to that of sum of all eigenvalues; 0.9691

(d) Obtain the mean and arithmetic mean of each variable. Take the deviation of each observation from mean of the corresponding variable and divide the deviation by standard deviation of corresponding variable and repeat steps given above.

## 24.13 PRACTICAL ASSIGNMENTS

**Session 9**

Write a programme in ‘C’ language to find the principal components from a covariance matrix and test it for the data given in Example 1.

**Session 10**

Write a programme in ‘C’ language to find the principal components from a correlation matrix and test it for the data given in Example 2.

# UNIT 25 FACTOR ANALYSIS

**Structure** **Page No**



## 25.1 INTRODUCTION

Factor analysis is a generic name given to a class of multivariate statistical methods whose primary purpose is to define the underlying structure in a data matrix and achieve the objectives of summarization and reduction of data. The essential purpose of factor analysis is to describe, if possible, the covariance relationships among many variables in terms of a few underlying but unobservable random quantities called *factors*. Factor analysis has originated in the field of psychology 100 years ago by Charles Spearman to define the concepts like intelligence, attitude, measures of mathematical skill, vocabulary, other verbal skills, artistic skills, logical reasoning ability, etc.

A frequent source of confusion in the field of factor analysis is the term factor. It sometimes refers to a hypothetical, unobservable variable as in the phrase common factor. In this sense, factor analysis must be distinguished from component analysis since a component is an observable linear combination. Factor is also used in the sense of matrix factor, in that one matrix is a factor of second matrix if the first matrix multiplied by its transpose equals the second matrix. In this sense, factor analysis refers to all methods of data analysis using matrix factors, including component analysis and common factor analysis, which is discussed in Section 25.2.

A common factor is an unobservable hypothetical variable that contributes to the variance of at least two of the observed variables. The unqualified term “factor” often refers to a common factor. A unique factor is an unobservable hypothetical variable that contributes to the variance of only one of the observed variables. The model for common factor analysis posits one unique factor for each observed variable is discussed in Section 25.3. The common factors generate the covariance among the observable responses while the specific factor contribute only to the variances of their particular response. Basically the factor model is motivated by the following argument - suppose variables can be grouped by their correlations, i.e., all variables within a particular group are highly correlated among themselves but have relatively small correlations with variables in a different group. It is conceivable that each group of variables represents a single underlying construct, or factor, that is responsible for the observed correlations. For example, for an individual, marks in different subjects may be governed by aptitudes (common factors) and individual variations (specific factors) and interest may lie in obtaining scores on unobservable aptitudes from observable data on marks in different subjects. The methods of estimation are discussed in Section 25.4 and the factor analysis versus clustering and multidimensional scaling is discussed in Section 25.5.

### Objectives

After reading this unit, you should be able to:

- have an overview of factor analysis and meaning of common factors.
- derive factors.
- define factor loadings.
- identify the number of factors.
- apply factor analysis to data.

## 25.2 ABSOLUTE VERSUS HEURISTIC USES OF FACTOR ANALYSIS

A heuristic is a way of thinking about a topic which is convenient even if not absolutely true. We use a heuristic approach when we talk about the sun rising and setting as if the sun moved around the earth, even though we know it doesn't. Spearman hypothesized that if g could be measured and one could select a subpopulation of people with the same score on *g*, in that subpopulation one would find no correlations among any tests of mental ability. In other words, he hypothesized that *g* was the only factor common to all those measures. Spearman's *g* theory of intelligence, and the activation theory of autonomic functioning, can be thought of as absolute theories which are or were hypothesized to give complete descriptions of the pattern of relationships among variables.

It is well known that Rubenstein (1986) studied the nature of curiosity by analyzing the agreements of junior-high-school students with a large battery of statements such as "I like to figure out how machinery works" or "I like to try new kinds of food." A factor analysis identified seven factors: three measuring enjoyment of problem-solving, learning, and reading; three measuring interests in natural sciences, art and music, and new experiences in general; and one indicates a relatively low interest in money. Rubenstein never claimed that her list of the seven major factors of curiosity offered a complete description of curiosity. Rather those factors merely appear to be the most important seven factors - the best way of summarizing a body of data. Factor analysis can suggest either absolute or heuristic models; the distinction is in how you interpret the output.

## 25.3 THE FACTOR MODEL

Suppose observations are made on p variables, $x_1, x_2, \ldots x_p$. The factor analysis model assumes that there are m underlying factors ($m<p$) $f_1, f_2, \ldots, f_m$ and each observed variables is a linear function of these factors together with a residual variate, so that

$$x_j = a_{j1} f_1 + a_{j2} f_2 + \ldots + a_{jm} f_m \qquad \text{where} \qquad j = 1,2, ..., p$$

where $a_{j1}, a_{j2}, \ldots, a_{jm}$ are called factor loadings i.e. $a_{ji}$ is loading of j-th variable on i-th factor. $y_j$'s are called specific factors.

The proportion of the variance of the j-th variable contributed by the *m* common factors is called the j-th communality and the proportion due to the specific factor is called the uniqueness, or specific variance.

In matrix notation, we can write the model as

$$X = \mathbf{a}F + Y$$

The covariance matrix of X if the population means of $X_i$ are zero, is given as

$$\boldsymbol{\Sigma} = \text{cov}\,(X) = E(X-\mu)(X-\mu') = \mathbf{aa}' + \boldsymbol{\psi}$$

where $\psi = \begin{bmatrix} \psi_1 & \cdots & 0 \\ \vdots & \ddots & \\ 0 & & \psi_p \end{bmatrix}$ and $\psi_i = \text{var}(y_i)$

Also, $\text{var}(X_i) = \sigma_{ii} = a_{i1}^2 + a_{i2}^2 + \cdots + a_{im}^2 + \psi_i$,

$\text{cov}(X_i, X_j) = \sigma_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} + \cdots + a_{im}a_{jm}$

and $\text{cov}(\mathbf{X,F}) = \mathbf{a}$

or $\text{cov}(X_i, F_j) = a_{ij}$.

where we have assumed that $y_j$ are uncorrelated with each other and with factors $f_1, f_2, \cdots, f_m$. Also $f_j$ are standardised variables and are uncorrelated.

**Example 1:** Consider the three random variables $X_1$, $X_2$ and $X_3$ having the following the covariance matrix.

$$\Sigma = \begin{bmatrix} 1 & 0.63 & 0.45 \\ 0.63 & 1 & 0.35 \\ 0.45 & 0.35 & 1 \end{bmatrix}$$

for p = 3 and m = 1, write its factor model.

**Solution:** Comparing the given covariance matrix with $\boldsymbol{\Sigma} = \mathbf{aa}' + \boldsymbol{\psi}$, we get

$$a_{11}^2 + \psi_1 = 1$$
$$a_{11}a_{21} = 0.63$$
$$a_{11}a_{31} = 0.45$$
$$a_{21}^2 + \psi_2 = 1$$
$$a_{21}a_{31} = 0.35$$
$$a_{31}^2 + \psi_3 = 1$$

from these we get

$$|a_{11}| = 0.9, |a_{21}| = 0.7, |a_{31}| = 0.5 \text{ and } \psi_1 = 0.19, \psi_2 = 0.51, \psi_3 = 0.75$$

Now try the following exercise.

E1) Let p = 3 and m = 1 and suppose the random variables $X_1$, $X_2$, and $X_3$ have the positive definite covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.4 \\ 0.7 & 0.4 & 1 \end{bmatrix}$$

Write its factor model.

Factor analysis involves:

- Deciding number of common factors (m)
- Estimating factor loadings ($a_{ji}$)
- Calculating factor scores ($f_i$)

Now let us discuss methods of estimation.

## 25.4 METHODS OF ESTIMATION

Factor analysis is done in two parts, first solution is obtained by placing some restrictions and then final solution is obtained by rotating this solution. There are two most popular methods available in literature for parameter estimation, the principal component (and the related principal factor) method and the maximum likelihood method. After the factors are estimated, it is necessary to interpret them, i.e., to assign a name to each common factor that represents the importance of the factor in predicting each of the observed variables. The interpretation is a subjective process. To simplify the interpretation of factors and to make it less subjective, the solution from either method can be rotated in order i.e. either factor loadings are close to unity or close to zero.

The most popular method for orthogonal rotation is Varimax Rotation method. Varimax rotation method emphasizes on detection of the factors each of which is related to few variables rather than detection of factors influencing all variables. In some specific situations, oblique rotations are also used. It is always prudent to try more than one method of solution. If the factor model is appropriate for the problem at hand, the solutions should be consistent with one another. The estimation and rotation methods require iterative calculations that must be done on a computer. If variables are uncorrelated factor analysis will not be useful. In these circumstances, the specific factors play the dominant role, whereas the major aim of the factor analysis is to determine a few important common factors.

After the initial factor extraction, the common factors are uncorrelated with each other. If the factors are rotated by an orthogonal transformation, the rotated factors are uncorrelated. If the factors are rotated by an oblique transformation, the rotated factors become correlated. Oblique rotations often produce more useful patterns than do orthogonal rotations. However, a consequence of correlated factors is that there is no single unambiguous measure of the importance of a factor in explaining a variable. Thus, for oblique rotations, the pattern matrix doesn't provide all the necessary information for interpreting the factors.

Number of factors is theoretically given by rank of population variance covariance matrix. However, in practice, number of common factors retained in the model is increased until a suitable proportion of total variance is explained. Another convention, frequently encountered in packaged computer programs is to set m equal to the number of eigenvalues greater than one. In fact determining the optimal number of factors to extract is not straightforward. Besides, the criteria of eigenvalue more than one, there are several criteria for the number o factors to be extracted, but these are just empirical guidelines rather than exact solution. Some of the most commonly used guidelines are the Kaiser- Guttman rule (eigenvalues more than the average of initial communality estimates, which is always less than one); percentage of variance (the pre fixed percentage or proportion of the common variance that is explained by successive factors obtained by the sum of communality estimates); the scree test (plot the eigenvalues against the corresponding factor numbers and take maximum number of factors to be extracted as the one less than the factor number at which the curve bends or make an "elbow"), the size of residuals and interpretability.

As in principal component analysis, principal factor method for factor analysis depends upon unit of measurement. If units are changed, the solution will change. However, in this approach estimated factor loadings for a given factor do not change as the number of factors is increased. In contrast to this, in maximum likelihood method, the solution does not change if units of measurements are changed. However, in this method the solution changes if number of common factors is changed.

**Example 2:** What underlying attitudes lead people to respond to the questions on a political survey as they do? Examining the correlations among the survey items reveals that there is significant overlap among various subgroups of items--questions about taxes tend to correlate with each other, questions about military issues correlate with each other, and so on. With factor analysis, you can investigate the number of

underlying factors and, in many cases; you can identify what the factors represent conceptually. Additionally, you can compute factor scores for each respondent, which can then be used in subsequent analyses. For example, you might build a logistic regression model to predict voting behavior based on factor scores.

**Example 3:** A manufacturer of fabricating parts is interested in identifying the determinants of a successful salesperson. The manufacturer has on file the information shown in the following table. He is wondering whether he could reduce these seven variables to two or three factors, for a meaningful appreciation of the problem.

**Data Matrix for Factor Analysis of seven variables (14 sales people)**

| **Sales Person** | **Height** $(x_1)$ | **Weight** $(x_2)$ | **Education** $(x_3)$ | **Age** $(x_4)$ | **No. of Children** $(x_5)$ | **Size of Household** $(x_6)$ | **IQ** $(x_7)$ |
|---|---|---|---|---|---|---|---|
| 1 | 67 | 155 | 12 | 27 | 0 | 2 | 102 |
| 2 | 69 | 175 | 11 | 35 | 3 | 6 | 92 |
| 3 | 71 | 170 | 14 | 32 | 1 | 3 | 111 |
| 4 | 70 | 160 | 16 | 25 | 0 | 1 | 115 |
| 5 | 72 | 180 | 12 | 30 | 2 | 4 | 108 |
| 6 | 69 | 170 | 11 | 41 | 3 | 5 | 90 |
| 7 | 74 | 195 | 13 | 36 | 1 | 2 | 114 |
| 8 | 68 | 160 | 16 | 32 | 1 | 3 | 118 |
| 9 | 70 | 175 | 12 | 45 | 4 | 6 | 121 |
| 10 | 71 | 180 | 13 | 24 | 0 | 2 | 92 |
| 11 | 66 | 145 | 10 | 39 | 2 | 4 | 100 |
| 12 | 75 | 210 | 16 | 26 | 0 | 1 | 109 |
| 13 | 70 | 160 | 12 | 31 | 0 | 3 | 102 |
| 14 | 71 | 175 | 13 | 43 | 3 | 5 | 112 |

Can we now collapse the seven variables into three factors? Intuition might suggest the presence of three primary factors: maturity revealed in age/children/size of household, physical size as shown by height and weight, and intelligence or training as revealed by education and IQ.

The sales people data have been analyzed in the following table. For this the data is given in the original units, and transformed into standard scores. The three factors derived from the sales people data by principal component analysis are presented below:

**Three-factor results with seven variables**

| **Variable** | **Sales People Characteristics** | | | **Communality** |
|---|---|---|---|---|
| | **Factor I** | **Factor II** | **Factor III** | |
| Height | 0.59038 | 0.72170 | -0.30331 | 0.96140 (sum sq I, II and III) |
| Weight | 0.45256 | 0.75932 | -0.44273 | 0.97738 |
| Education | 0.80252 | 0.18513 | 0.42631 | 0.86006 |
| Age | -0.86689 | 0.41116 | 0.18733 | 0.95564 |
| No. of Children | -0.84930 | 0.49247 | 0.05883 | 0.96730 |
| Size of Household | -0.92582 | 0.30007 | -0.01953 | 0.94756 |
| IQ | 0.28761 | 0.46696 | 0.80524 | 0.94918 |
| **Sum of squares** | **3.61007** | **1.85136** | **1.15709** | |
| **Variance summarized** | **0.51572** | **0.26448** | **0.16530** | **Average=0.94550** |

**Factor Loadings**

The coefficients in the factor equations are called "factor loadings". They appear above in each factor column, corresponding to each variable. The equations are:

$$\mathbf{F}_1 = 0.59038x_1 + 0.45256x_2 + 0.80252x_3 - 0.86689x_4 - 0.84930x_5 - 0.92582x_6 + 0.28$$

$$\mathbf{F}_2 = 0.72170x_1 + 0.75932x_2 + 0.18513x_3 + 0.41116x_4 + 0.49247x_5 + 0.30007x_6 + 0.46$$

$$\mathbf{F}_3 = -0.30331x_1 - 0.44273x_2 + 0.80252x_3 + 0.18733x_4 + 0.58830x_5 - 0.01953x_6 + 0.8$$

The factor loadings depict the relative importance of each variable with respect to a particular factor. In all the three equations, education $(x_3)$ and IQ $(x_7)$ have got positive loading factor indicating that they are variables of importance in determining the success of sales person.

**Variance summarized**

Factor analysis employs the criterion of maximum reduction of variance - variance found in the initial set of variables. Each factor contributes to reduction. In our example Factor I accounts for 51.6% of the total variance. Factor II for 26.4% and Factor III for 16.5%. Together the three factors "explain" almost 95% of the variance.

**Communality**

In the ideal solution the factors derived will explain 100% of the variance in each of the original variables; "Communality" measures the percentage of the variance in the original variables that is captured by the combinations of factors in the solution. Thus communality is computed for each of the original variables. Each variables communality might be thought of as showing the extent to which it is revealed by the system of factors. In our example the communality is over 85% for every variable. Thus the three factors seem to capture the underlying dimensions involved in these variables.

**Example 4:** In a consumer - preference study, a random sample of customers were asked to rate several attributes of a new product. The response on a 5-point semantic differential scale were tabulated and the attribute correlation matrix constructed and is given below

| **Attribute** | | **Correlation matrix** | | | | |
|---|---|---|---|---|---|---|
| | | **1** | **2** | **3** | **4** | **5** |
| **Taste** | **1** | 1 | .02 | .96 | .42 | .01 |
| **Good buy for money** | **2** | .02 | 1 | .13 | .71 | .85 |
| **Flavor** | **3** | .96 | .13 | 1 | .5 | .11 |
| **Suitable for snack** | **4** | .42 | .71 | .50 | 1 | .79 |
| **Provides energy** | **5** | .01 | .85 | .11 | .79 | 1 |

It is clear from the correlation matrix that variables 1 and 3 and variables 2 and 5 form groups. Variable 4 is "closer" to the (2,5) group than (1,3) group. Observing the results, one can expect that the apparent linear relationships between the variables can be explained in terms of, at most, two or three common factors.

**Solution:** Using the method of Principal Components we get

Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

| Eigenvalues of the | Correlation Matrix: | Total = 5 | Average | = 1 | |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | |
| Eigenvalue | 2.8531 | 1.8063 | 0.2045 | 0.1024 | 0.0337 |
| Difference | 1.0468 | 1.6018 | 0.1021 | 0.0687 | |
| Proportion | 0.5706 | 0.3613 | 0.0409 | 0.0205 | 0.0067 |
| Cumulative | 0.5706 | 0.9319 | 0.9728 | 0.9933 | 1.0000 |

| | FACTOR1 | FACTOR2 |
|---|---|---|
| TASTE | 0.55986 | 0.81610 |
| MONEY | 0.77726 | -0.52420 |
| FLAVOR | 0.64534 | 0.74795 |
| SNACK | 0.93911 | -0.10492 |
| ENERGY | 0.79821 | -0.54323 |

Variance explained by each factor

| FACTOR1 | FACTOR2 |
|---|---|
| 2.853090 | 1.806332 |

Final Communality Estimates: Total = 4.659423

| TASTE | MONEY | FLAVOR | SNACK | ENERGY |
|---|---|---|---|---|
| 0.979461 | 0.878920 | 0.975883 | 0.892928 | 0.932231 |

Residual Correlations with Uniqueness on the Diagonal

| ENERGY | TASTE | MONEY | FLAVOR | SNACK |
|---|---|---|---|---|
| 0.02054 | 0.01264 | -0.01170 | -0.02015 | 0.00644 |
| 0.01264 | 0.12108 | 0.02048 | -0.07493 | -0.05518 |
| -0.01170 | 0.02048 | 0.02412 | -0.02757 | 0.00119 |
| -0.02015 | -0.07493 | -0.02757 | 0.10707 | -0.01660 |
| 0.00644 | -0.05518 | 0.00119 | -0.01660 | 0.06777 |

Rotation Method: Varimax; Rotated Factor Pattern

| | FACTOR1 | FACTOR2 |
|---|---|---|
| TASTE | 0.02698 | 0.98545 |
| MONEY | 0.87337 | 0.00342 |
| FLAVOR | 0.13285 | 0.97054 |
| SNACK | 0.81781 | 0.40357 |
| ENERGY | 0.97337 | -0.01782 |

Variance explained by each factor

| | FACTOR1 | FACTOR2 |
|---|---|---|
| Weighted | 25.790361 | 58.765959 |
| Unweighted | 2.397426 | 2.076256 |

| | TASTE | MONEY | FLAVOR | SNACK | ENERGY |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| Communality | 0.971832 | 0.762795 | 0.959603 | 0.831686 | 0.947767 |

It can be seen that two factor model with factor loadings shown above is providing a good fit to the data as the first two factors explains 93.2% of the total standardized sample variance, i.e., $\left(\frac{\lambda_1 + \lambda_2}{p}\right)x100$ , where p is the number of variables. It can also be seen from the results that there is no clear-cut distinction between factor loadings for the two factors before rotation but after rotation the same is clear.

Now, in the following section we shall discuss factor analysis versus clustering and multidimensional scaling.

## 25.5 FACTOR ANALYSIS VERSUS CLUSTERING AND MULTIDIMENSIONAL SCALING

Factor analysis is typically applied to a correlation matrix, where as clustering and multidimensional scaling methods can be applied to any sort of matrix of similarity measures, such as ratings of the similarity of faces. But unlike factor analysis, those methods cannot cope with certain unique properties of correlation matrices, such as reflections of variables. For instance, if you reflect or reverse the scoring direction of a measure of "introversion", so that high scores indicate "extroversion" instead of introversion, then you reverse the signs of all that variable's correlations: -.36 becomes +.36, +.42 becomes -.42, and so on. Such reflections would completely change the output of a cluster analysis or multidimensional scaling, while factor analysis would recognize the reflections for what they are; the reflections would change the signs of the "factor loadings" of any reflected variables, but would not change anything else in the factor analysis output.

Another advantage of factor analysis over these other methods is that factor analysis can recognize certain properties of correlations. For instance, if variables A and B each correlate .7 with variable C, and correlate .49 with each other, factor analysis can recognize that A and B correlate zero when C is held constant because $.7^2 = .49$. Multidimensional scaling and cluster analysis have no ability to recognize such relationships, since the correlations are treated merely as generic "similarity measures" rather than as correlations.

We are not saying these other methods should never be applied to correlation matrices; sometimes they yield insights not available through factor analysis. But they have definitely not made factor analysis obsolete.

As discussed earlier, the primary purpose of factor analysis is data reduction and summarization. This technique has been widely used especially in behavioral sciences, to assess and construct the validity of a test or a scale. For example, a psychologist developed a new battery of 15 subtests to measure three distinct psychological constructs and wants to validate that battery. A sample of 300 subjects was drawn from the population and measured on the battery of 15 subsets. The $300 \times 15$ data matrix can be subjected to a factor analysis. The number of factors extracted and the pattern of relationships among the observed variables and the common factors can provide information on the construct validity of the test battery to the researcher.

Factor analysis is an interdependence technique in which all the variables are simultaneously considered each related to all others. It is a highly useful and powerful multivariate statistical technique for effectively extracting information from large databases and making sense of large bodies of interrelated data. Factor analysis has the ability to identify sets of related variables and even develop a single composite

measure to represent the entire set of related variables. Factor analysis has a lot of potential applications in problem solving and decision making in business and research institutions and it will continue to grow as more and more researchers gets familiarized with the benefits of data summarization and data reduction.

Factor analysis is a much more complex and involved subject than given in this brief exposition. The major limitations of this technique are

- Existence of several techniques for performing factor analysis and no clear cut comparison of these techniques
- Subjective assessment of number of factors to be extract, rotation method to be used and judging the significance of factor loadings
- Problems with reliability with the results of a single-factor analytic solution which looks plausible but may change with the changes in sample, data gathering process or the numerous kinds of measurement errors. It is important to emphasize that the plausibility is no guarantee of validity or even stability.

Now try the following exercises.

---

E2) A large number of people were asked to rate their liking of each of the five ready to serve fruit beverages: lemon, aonla, grape, mango and pineapple. The factor loadings of first three factors obtained through factor analysis are

| | Factors | | | |
|---|---|---|---|---|
| | I | II | III | Communality |
| Lemon | -0.219 | 0.363 | -0.338 | 0.2939 |
| Aonla | -0.137 | 0.682 | 0.307 | 0.5781 |
| Grape | 0.514 | -0.213 | -0.227 | 0.3611 |
| Mango | 0.485 | -0.117 | 0.115 | 0.2621 |
| Pineapple | -0.358 | -0.635 | 0.534 | 0.8165 |
| Sum of Squares | 0.6943 | 1.0592 | 0.5584 | |
| Variance Summarized | 0.1389 | 0.2118 | 0.1117 | 0.4624 |

(i) Write the linear equations for all the three factors.
(ii) Interpret the loading coefficients, variance summarized and communality value of this Table.

E3) Department of transportation obtained the data on the following features of 50 bridges: Design time in man days ($X_1$); Deck area of bridge in square metre ($X_2$), Construction cost in rupees ($X_3$), number of structural drawing ($X_4$), length of bridge in metres ($X_5$), number of spans ($X_6$) and degree of difficulty in bridge design($X_7$). The factor loadings of first two factors obtained through factor analysis using principal component estimation method with varimax rotation are

| | Factor | |
|---|---|---|
| | I | II |
| $X_1$ | 0.69732 | 0.47572 |
| $X_2$ | 0.74797 | 0.44545 |
| $X_3$ | 0.83123 | 0.35001 |
| $X_4$ | 0.59594 | 0.64808 |
| $X_5$ | 0.93472 | 0.16039 |
| $X_6$ | 0.86564 | 0.20127 |
| $X_7$ | 0.16549 | 0.93573 |

(i) Compute the communality of each variable and the percentage of its variance that I explained by factors.
(ii) Compute the percentage of total variation explained by each factor separately and both factors jointly.
(iii) Interpret the loading coefficients, variance summarized and communality value obtained in (a) and (b).

---

Now, let us summarize the unit.

## 25.6 SUMMARY

In this unit, we have covered the following points.

1) Factor analysis is a dimensional reduction technique in which the covariance relationships among many variables is descried in terms of a smaller number of unobservable random quantities called factors.
2) Communality of a variable is the part of its variance that is explained by the common factors.
3) Most widely used method for determining a first set of loadings is the principal component method and gives the values of loadings that bring the estimate of total communality as close as possible to the total observed variance.
4) If the variables are measured in different units, then it is advisable to standardize the variables so that all have mean equal to zero and variance equal to one. This is achieved by subtracting the mean from each of the observed vales of the corresponding variable and dividing the difference by its corresponding standard deviation.
5) The initially extracted factors can be rotated either using orthogonal or oblique transformations.

## 27.7 SOLUTIONS/ANSWERS

E1) The covariance structure is

$$\boldsymbol{\Sigma} = \mathbf{a}\mathbf{a}' + \boldsymbol{\psi}$$

or

$$1 = a_{11}^2 + \psi_1$$
$$0.90 = a_{11}a_{21}$$
$$0.70 = a_{11}a_{31}$$
$$1 = a_{21}^2 + \psi_2$$
$$0.40 = a_{21}a_{31}$$
$$1 = a_{31}^2 + \psi_3$$

The pair of equations

$$0.70 = a_{11}a_{31}$$
$$0.40 = a_{21}a_{31}$$

imply

$$a_{21} = \left(\frac{0.40}{0.70}\right)a_{11}$$

Substituting this result for $a_{21}$ in the equation

$$.90 = a_{11}a_{21}$$

Yields $a_{11}^2 = 1.575$ or $a_{11} = \pm 1.255$. Since $\text{Var}(F_1) = 1$ (by assumption) and $\text{Var}(X_1) = 1$, $a_{11} = \text{Cov}(X_1, F_1) = \text{Corr}(X_1, F_1)$. A correlation coefficient cannot be greater than unity (in absolute value) so, from this point of view, $|a_{11}| = 1.255$ is "too large." Also the equation

$$1 = a_{11}^2 + \psi_1 \quad \text{or} \quad \psi_1 = 1 - a_{11}^2$$

given

$$\psi_1 = 1 - 1.575 = -.575$$

Which is unsatisfactory since it gives a negative value for $\text{Var}(y_1) = \psi_1$. Thus, with m = 1,it is possible to get a unique numerical solution to the equations $\boldsymbol{\Sigma} = \mathbf{LL}' + \boldsymbol{\psi}$. However, the solution is not consistent with the statistical interpretation of the coefficients, so it is not a proper solution.

E2) (i) The linear equations of all the three factors are

$$F_1 = -0.219 * \text{lemon} - 0.137 * \text{aonla} + 0.514 * \text{grape} + 0.485 * \text{mango} - 0.358 * \text{pineapple}$$

$$F_2 = 0.363 * \text{lemon} + 0.682 * \text{aonla} - 0.213 * \text{grape} - 0.117 * \text{mango} - 0.635 * \text{pineapple}$$

$$F_3 = -0.338 * \text{lemon} + 0.307 * \text{aonla} - 0.227 * \text{grape} + 0.115 * \text{mango} + 0.534 * \text{pineapple}$$

(ii) Factor 1 accounts for 13.9 % of variance; Factor II accounts for 21.2% and Factor III accounts for 11.2% of variance. Taken together 3 factors account for 46.2% of variance.

Communality for most of the variables is less except pineapple. Therefore, one may conclude that the 3 factors are not able to capture the underlying dimensions in these variables.

E3) Communality of each variable are obtained by taking sum of squares of the loadings for that variable in all the factors. Average variance explained is obtained by taking the average of communalities. For obtaining the variance summarized by each factor, take sum of squares of the loadings of all the variables for that factor and divide it by the sum of such sum of squares for all the factors. Then this ratio is multiplied by the average variance explained. The results are given as

| Variables | Factors | | Communality (Sum squares I and II) |
|---|---|---|---|
| | I | II | |
| $X_1$ | 0.69732 | 0.47572 | 0.712565 |
| $X_2$ | 0.74797 | 0.44545 | 0.757885 |
| $X_3$ | 0.83123 | 0.35001 | 0.81345 |
| $X_4$ | 0.59594 | 0.64808 | 0.775152 |
| $X_5$ | 0.93472 | 0.16039 | 0.899426 |
| $X_6$ | 0.86564 | 0.20127 | 0.789842 |
| $X_7$ | 0.16549 | 0.93573 | 0.902978 |
| Sum of Squares | 3.742223 | 1.909075 | |
| Variance summarized | 0.534603 | 0.2772725 | Average=0.807328 |

## 25.8 PRACTICAL ASSIGNMENTS

**Session 11**

Write a programme in C-language to write the factor model of a covariance matrix given in Example 1.

# UNIT 26 CANONICAL CORRELATION ANALYSIS

**Structure** **Page No.**



## 26.1 INTRODUCTION

**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 actually 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 pair. 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.

Analogous with ordinary correlation, canonical correlation squared is the percent of variance in the dependent set explained by the independent set of variables along a given dimension (there may be more than one). In addition to asking how strong the relationship is between two latent variables, canonical correlation is useful in determining how many dimensions are needed to account for that relationship. Canonical correlation finds the linear combination of variables that produces the largest correlation with the second set of variables. This linear combination, or "root," is extracted and the process is repeated for the residual data, with the constraint that

the second linear combination of variables must not correlate with the first one. The process is repeated until a successive linear combination is no longer significant.

Canonical correlation is a member of the multiple general linear hypothesis (MLGH) family and shares many of the assumptions of mutliple regression such as linearity of relationships, homoscedasticity (same level of relationship for the full range of the data), interval or near-interval data, untruncated variables, proper specification of the model, lack of high multicollinearity, and multivariate normality for purposes of hypothesis testing.

Often in applied research, scientists encounter variables of large dimensions and are faced with the problem of understanding dependency structures, reduction of dimensionalities, construction of a subset of good predictors from the explanatory variables, etc. Canonical correlation Analysis (CCA) provides us with a tool to attack these problems. However, its appeal and hence its motivation seem to differ from the theoretical statisticians to the social scientists. We deal here with the various motivations of CCA as mentioned above and related statistical inference procedures. We shall begin the unit by discussing the canonical correlation in Section 26.2. In Section 26.3, we shall focus on non linear canonical correlation

.

## Objectives

After reading this unit, you should be able to

- understand the meaning and concept of canonical correlation
- interpret the results of a canonical correlation analysis
- make use of canonical correlation

# 26.2 MATHEMATICAL FORMULATION AND COMPUTATIONS

Consider two groups of variables. The first set of *p* variables is represented by a $(p \times 1)$ random vector **X** and the second set of *q* variables is represented by the $(q \times 1)$ random vector **Y**. Without loss of generality, we assume that $p \le q$. For two random vectors **X** and **Y**, let $E(X) = \mu_X$; $D(\mathbf{X}) = \Sigma_{11}$; $E(Y) = \mu_Y$; $D(\mathbf{Y}) = \Sigma_{22}$; Cov (**X**, **Y**) = $\Sigma_{12}$. Here E (.) denotes expectation and D(.) denotes the variance-covariance matrix. For the sake of convenience, let us consider **X** and **Y** jointly as the random vector

$\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \cdots \\ \mathbf{Y} \end{bmatrix}$ with mean vector $\mu = E(Z) = \begin{bmatrix} \mu_X \\ \cdots \\ \mu_Y \end{bmatrix}$ and variance covariance matrix as

$D(Z) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. The pq elements in $\mathbf{\Sigma}_{12}$ gives the measures of association between *p* variables of first set and *q* variables of second set. The canonical correlation analysis summarizes the associations between **X** and **Y** in terms of few correlations rather than the pq elements in $\mathbf{\Sigma}_{12}$.

Consider $U = a'X$ and $V = b'Y$ as the linear combinations of variables within the two sets for some coefficient vectors **a** and **b**. Now, using simple statistical computations, one can see that

$$Var(U) = a'\Sigma_{11}a;\ Var(V) = b'\Sigma_{22}b \quad \text{and} \quad Cov(U,V) = \mathbf{a'\Sigma_{12}b}.$$

As a consequence

$Corr(U,V) = \dfrac{a'\Sigma_{12}b}{\sqrt{a'\Sigma_{11}a}\sqrt{b'\Sigma_{22}b}}$. Through canonical correlation analysis we want to find **a** and **b** such that $Corr(U,V)$ is as large as possible.

The first linear combination of two sets of variables known as first pair of canonical variables is

$$U_1 = a_{11}x_1 + a_{21}x_2 + ... + a_{p1}x_p = \mathbf{a_1'X}$$

$$V_1 = b_{11}y_1 + b_{21}y_2 + ... + b_{q1}y_q = \mathbf{b_1'Y}.$$

The coefficients are so chosen that the two canonical variables, $U_1$ and $V_1$ have unit variances and have the largest possible correlation. This maximized correlation between the two canonical variables is the first canonical correlation. The coefficients of the linear combinations are canonical coefficients or canonical weights. Let the maximum correlation attained in first pair of canonical variables be $\text{Corr}(U_1, V_1) = \rho_1$.

The second set of canonical variables, uncorrelated with the first pair of canonical variables is

$$U_2 = a_{12}x_1 + a_{22}x_2 + ... + a_{p2}x_{p=} \mathbf{a_2'X}$$

$$V_2 = b_{12}y_1 + b_{22}y_2 + ... + b_{q2}y_q = \mathbf{b_2'Y}.$$

The coefficients are so chosen that the two canonical variables, $U_2$ and $V_2$; $U_2$ is uncorrelated with $U_1$ and $V_1$, $V_2$ is uncorrelated with $U_1$ and $V_1$, and $U_2$ and $V_2$ have the largest possible correlation subject to these constraints. Let the maximum correlation attained in second pair of canonical variables be $\text{Corr}(U_2, V_2) = \rho_2$.

This process continues until the number of pairs of canonical variables is equal to the number of variables in the set containing smaller number of variables. In this case number of canonical variables will be p as $p \leq q$. Let the maximum correlation attained in $p^{th}$ pair of canonical variables be $\text{Corr}(U_p, V_p) = \rho_p$.

Now the question arises, how to compute $\rho_1, \rho_2, \ldots, \rho_p$.

Through algebraic manipulations, it can easily be seen that the eigenvalues of $\mathbf{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}}$ are $\rho_1^2 \geq \rho_2^2 \geq \ldots \geq \rho_p^2$ with corresponding $p \times 1$ eigenvectors as $a_1, a_2, \ldots, a_p$. Similarly the eigenvalues of $\mathbf{\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}}$ are $\rho_1^2 \geq \rho_2^2 \geq \ldots \geq \rho_p^2$ with corresponding $q \times 1$ eigenvectors as $b_1, b_2, \ldots, b_p$. Therefore, the maximum correlation, $\rho_1$ is the positive square root of the largest eigenvalue of $\mathbf{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}}$ or $\mathbf{\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}}$. Similarly the maximum correlation between second pair of canonical variables $\rho_2$ is the positive square root of the second largest eigenvalue of $\mathbf{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}}$ or $\mathbf{\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}}$. The canonical variables are obtained by using the coefficient vectors as corresponding ortho-normalized eigenvectors. This process continues till we get $p^{th}$ pair of canonical variables and p canonical correlations. In the above discussion, it has been tacitly assumed that $\Sigma_{ii}$, i = 1,2 is non-singular. If $\Sigma_{ii}$, i=1 or 2 happens to be singular, one can use a g-inverse of $\Sigma_{ii}$ in place of true inverse of $\Sigma_{ii}$.

**Remark 1:** $p = q = 1 \Rightarrow \rho_1$= usual Pearson's product moment correlation coefficient between the scalar random variables X and Y; p = 1, q >1 $\Rightarrow \rho_1$= Multiple correlation coefficient between the scalar X and the vector random variable **Y**. Sample analogues are trivially defined.

**Remark 2:** The canonical correlation also reduces the dimensionality. One feature of dimensionality reduction is that we have only 2p canonical variables rather than p+q original variables. Further, reduction in dimensionality can be achieved by retaining only those pair of canonical variables, whose correlation coefficient is significantly different from zero. If the correlation coefficient between $(k+1)^{th}$ pair of canonical variables is not significantly different from zero, then we can retain only first k pairs of canonical variables. This gives only 2k variables rather than p+q original variables.

**Remark 3:** The canonical correlations measure the linear association between two sets of variables. If the variables are associated in a nonlinear manner, the association can not be captured by canonical correlation. Thus, while canonical correlation analysis is the most generalized multivariate method, it is still constrained to identifying linear associations. Canonical correlation analysis can accommodate any metric variable without the strict assumption of normality. Normality is desirable because it standardizes a distribution to allow for a higher correlation among the variables. But in the strictest sense, canonical correlation analysis can accommodate even non-normal variables if the distributional form (e.g., highly skewed) does not decrease the correlation with other variables. This allows for transformed nonmetric data (in the form of dummy variables) to be used as well. However, multivariate normality is required for the statistical inference test of the significance of each canonical function. Because tests for multivariate normality are not readily available, the prevailing guideline is to ensure that each variable has univariate normality. Thus, although normality is not strictly required, it is highly recommended that all variables be evaluated for normality and transformed if necessary. Homoscedasticity, to the extent that it decreases the correlation between variables, should also be remedied. Finally, multicollinearity among either variable set will confound the ability of the technique to isolate the impact of any single variable, making interpretation less reliable.

Now, we shall discuss nonlinear canonical correlation in the following Section.

## 26.3 NONLINEAR CANONICAL CORRELATION

Nonlinear canonical correlation analysis corresponds to categorical canonical correlation analysis with optimal scaling. The OVERALS procedure in SPSS (part of SPSS Categories) implements nonlinear canonical correlation. Independent variables can be nominal, ordinal, or interval, and there can be more than two sets of variables (more than one independent set and one dependent set). Whereas ordinary canonical correlation maximizes correlations between the variable sets, in OVERALS the sets are compared to an unknown compromise set defined by the object scores

OVERALS makes use of optimal scaling, which quantifies categorical variables and then treats as numerical variables, including applying nonlinear transformations to find the best-fitting model. For nominal variables, the order of the categories is not retained but values are created for each category such that goodness of fit is maximized. For ordinal variables, order is retained and values maximizing fit are created. For interval variables, order is retained as are equal distances between values.

Obtain OVERALS from the SPSS menu by selecting Analyze, Data Reduction, Optimal Scaling; Select Multiple sets; Select either Some variable(s) not multiple nominal or All variables multiple nominal; click Define; define at least two sets of variables; define the value range and measurement scale (optimal scaling level) for each selected variable. SPSS output includes frequencies, centroids, iteration history, object scores, category quantifications, weights, component loadings, single and multiple fit, object scores plots, category coordinates plots, component loadings plots, category centroids plots, and transformation plots.

**Tip:** To minimize output, use the Automatic Recode facility on the Transform menu to create consecutive categories beginning with 1 for variables treated as nominal or ordinal. To minimize output, for each variable scaled at the numerical (integer) level, subtract the smallest observed value from every value and add 1.

**Warning:** Optimal scaling recodes values on the fly to maximize goodness of fit for the given data. As with any atheoretical, post-hoc data mining procedure, there is a danger of overfitting the model to the given data. Therefore, it is particularly appropriate to employ cross-validation, developing the model for a training dataset

and then assessing its generalizability by running the model on a separate validation dataset.

The SPSS manual notes, "If each set contains one variable, nonlinear canonical correlation analysis is equivalent to principal components analysis with optimal scaling. If each of these variables is multiple nominal, the analysis corresponds to homogeneity analysis. If two sets of variables are involved and one of the sets contains only one variable, the analysis is identical to categorical regression with optimal scaling."

Redundancy is the percent of variance in one set of variables accounted for by the variate of the other set. The researcher wants high redundancy, indicating that independent variate accounts for a high percent of the variance in the dependent set of original variables. Note this is not the canonical correlation squared, which the percent of variance in the dependent variate is accounted for by the independent variate.

## Applications of Canonical Correlation Analysis

- There could be a situation where some of variables have high structure correlations even though their canonical weights are near zero. This could happen because the weights are partial coefficients whereas the structure correlations (canonical factor loadings) are not: if a given variable shares variance with other independent variables entered in the linear combination of variables used to create a canonical variable, its canonical coefficient (weight) is computed based on the residual variance it can explain after controlling for these variables. If an independent variable is totally redundant with another independent variable, its partial coefficient (canonical weight) will be zero. Nonetheless, such a variable might have a high correlation with the canonical variable (that is, a high structure coefficient). In summary, the canonical weights have to do with the unique contributions of an original variable to the canonical variable, whereas the structure correlations have to do with the simple, overall correlation of the original variable with the canonical variable.
- Canonical correlation is not a measure of the percent of variance explained in the original variables. The square of the structure correlation is the percent of the variance in a given original variable accounted for by a given canonical variable on a given (usually the first) canonical correlation. Note that the average percent of variance explained in the original variables by a canonical variable (the mean of the squared structure correlations for the canonical variable) is not at all the same as the canonical correlation, which has to do with the correlation between the weighted sums of the two sets of variables. Put another way, the canonical correlation does <u>not</u> tell us how much of the variance in the original variables is explained by the canonical variables. Instead, that is determined on the basis of the squares of the structure correlations.
- Canonical coefficients can be used to explain with which original variables a canonical correlation is predominantly associated. The canonical coefficients are standardized coefficients and (like beta weights in regression) their magnitudes can be compared. Looking at the columns in SPSS output which list the canonical coefficients as columns and the variables in a set of variables as rows, some researchers simply note variables with the highest coefficients to determine which variables are associated with which canonical correlations and use this as the basis for inducing the meaning of the dimension represented by the canonical correlation.

However, Levine (1977) argues against the procedure above on the ground that the canonical coefficients may be subject to multi-collinearity, leading to incorrect judgments. Also, because of suppression, a canonical coefficient may even have a different sign compared to the correlation of the original variable with the canonical

variable. Therefore, instead, Levine recommends interpreting the relations of the original variables to a canonical variable in terms of the correlations of the original variables with the canonical variables - that is, by structure coefficients. This is now the standard approach.

Canonical correlation places the fewest restrictions on the types of data on which it operates. Because the other techniques impose more rigid restrictions, it is generally believed that the information obtained from them is of higher quality and may be presented in a more interpretable manner. For this reason, many researchers view canonical correlation as a last-ditch effort, to be used when all other higher-level techniques have been exhausted. But in situations with multiple dependent and independent variables, canonical correlation is the most appropriate and powerful multivariate technique. It has gained acceptance in many fields and represents a useful tool for multivariate analysis, particularly as interest has spread to considering multiple dependent variables.

Now try the following exercises.

E1) Following data was collected on physiological variables (weight in pounds, waist in inches and pulse rate) and exercise variables (chins, situps and jumps) on middle-aged men in a fitness club. Perform the canonical correlation analysis and interpret your results.

| Weight | Waist | Pulse | Chins | Situps | Jumps |
|---|---|---|---|---|---|
| 191 | 36 | 50 | 5 | 162 | 60 |
| 189 | 37 | 52 | 2 | 110 | 60 |
| 193 | 38 | 58 | 12 | 101 | 101 |
| 162 | 35 | 62 | 12 | 105 | 37 |
| 189 | 35 | 46 | 13 | 155 | 58 |
| 182 | 36 | 56 | 4 | 101 | 42 |
| 211 | 38 | 56 | 8 | 101 | 38 |
| 167 | 34 | 60 | 6 | 125 | 40 |
| 176 | 31 | 74 | 15 | 200 | 40 |
| 154 | 33 | 56 | 17 | 251 | 250 |
| 169 | 34 | 50 | 17 | 120 | 38 |
| 166 | 33 | 52 | 13 | 210 | 115 |
| 154 | 34 | 64 | 14 | 215 | 105 |
| 247 | 46 | 50 | 1 | 50 | 50 |
| 193 | 36 | 46 | 6 | 70 | 31 |
| 202 | 37 | 62 | 12 | 210 | 120 |
| 176 | 37 | 54 | 4 | 60 | 25 |
| 157 | 32 | 52 | 11 | 230 | 80 |
| 156 | 33 | 54 | 15 | 225 | 73 |
| 138 | 33 | 68 | 2 | 110 | 43 |

**Source**: http://v8doc.sas.com/sashtml/

E2) The variance covariance matrix between 5 yield attributing parameters (X1, X2, X3, X4 and X5) and four quality attributes (Y1, Y2, Y3, Y4) based on a sample of size 200 is given as (after standardizing the variables)

| | X1 | X2 | X3 | X4 | X5 | Y1 | Y2 | Y3 | Y4 |
|---|---|---|---|---|---|---|---|---|---|
| X1 | 1.000 | 0.754 | -0.690 | -0.440 | 0.702 | -0.605 | -0.480 | 0.780 | -0.152 |
| X2 | 0.754 | 1.000 | -0.710 | -0.515 | 0.412 | -0.722 | -0.419 | 0.542 | -0.100 |
| X3 | -0.690 | -0.710 | 1.000 | 0.323 | -0.444 | 0.737 | 0.361 | -0.546 | 0.172 |
| X4 | -0.440 | -0.515 | 0.323 | 1.000 | -0.334 | 0.527 | 0.461 | -0.393 | -0.019 |
| X5 | 0.702 | 0.412 | -0.444 | -0.334 | 1.000 | -0.383 | -0.505 | 0.737 | -0.148 |
| Y1 | -0.605 | -0.722 | 0.737 | 0.527 | -0.383 | 1.000 | 0.251 | -0.490 | 0.250 |
| Y2 | -0.480 | -0.419 | 0.361 | 0.461 | -0.505 | 0.251 | 1.000 | -0.434 | -0.079 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Y3 | 0.780 | 0.542 | -0.546 | -0.393 | 0.737 | -0.490 | -0.434 | 1.000 | -0.163 |
| Y4 | -0.152 | -0.100 | 0.172 | -0.019 | -0.148 | 0.250 | -0.079 | -0.163 | 1.000 |

Perform canonical correlation analysis and interpret your results.

E3) Consider the following variance-covariance matrix

| | X1 | X2 | Y1 | Y2 |
|---|---|---|---|---|
| X1 | 100 | 0 | 0 | 0 |
| X2 | 0 | 1 | 0.95 | 0 |
| Y1 | 0 | 0.95 | 1 | 0 |
| Y2 | 0 | 0 | 0 | 100 |

Obtain the correlation coefficient between first pair of canonical variables.

Now, let us summarize the unit.

## 26.4 SUMMARY

In this unit, we have covered the following points.

1) **Canonical correlation** is a technique to identify and quantify the association between two sets of variables.

2) Canonical correlation analysis actually 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.

3) Simple and multiple correlations are special cases of canonical correlation in which one or both sets contain a single variable.

4) The canonical correlation also reduces the dimensionality.

5) Nonlinear canonical correlation analysis corresponds to categorical canonical correlation analysis with optimal scaling.

## 26.5 SOLUTIONS/ANSWERS

**E1)** Step 1: Obtain the variance covariance matrix of weight, waist, pulse rate, chins, situps and jumps.

| **Covariance Matrix** | | | | | | |
|---|---|---|---|---|---|---|
| | **Weight** | **Waist** | **Pulse** | **Chins** | **Situps** | **Jumps** |
| **Weight** | 609.621053 | 68.800000 | -65.115789 | -50.863158 | -761.715789 | -286.505263 |
| **Waist** | 68.800000 | 10.252632 | -8.147368 | -9.347368 | -129.336842 | -31.442105 |
| **Pulse** | -65.115789 | -8.147368 | 51.989474 | 5.742105 | 101.521053 | 12.915789 |
| **Chins** | -50.863158 | -9.347368 | 5.742105 | 27.944737 | 230.107895 | 134.384211 |
| **Situps** | -761.715789 | -129.336842 | 101.521053 | 230.107895 | 3914.576316 | 2146.984211 |
| **Jumps** | -286.505263 | -31.442105 | 12.915789 | 134.384211 | 2146.984211 | 2629.378947 |

Now variance-covariance matrix of physiological variables weight, waist and pulse is

$$\boldsymbol{\Sigma}_{11} = \begin{bmatrix} 609.621053 & 68.800000 & -65.115789 \\ 68.000000 & 10.252632 & -8.147368 \\ -65.115789 & -8.147368 & 51.989474 \end{bmatrix}.$$

Variance-covariance matrix of exercise variables chins, situps and jmps is

$$\boldsymbol{\Sigma}_{22} = \begin{bmatrix} 27.944737 & 230.107895 & 134.384211 \\ 230.107895 & 3914.576316 & 2146.984211 \\ 134.384211 & 2146.984211 & 2629.378947 \end{bmatrix}.$$

Covariance matrix between physiological variables and exercise variables is

$$\boldsymbol{\Sigma}_{12} = \begin{bmatrix} -50.863158 & -9.347368 & 5.742105 \\ -761.715789 & -129.336842 & 101.521053 \\ -286.505263 & 101.521053 & 12.915789 \end{bmatrix}.$$

Step 2: To obtain $\boldsymbol{\Sigma}_{11}^{-1/2}$, first obtain eigenvalues and eigenvectors of $\boldsymbol{\Sigma}_{11}$. Let $\lambda_i$ and $\gamma_i$ denote $i^{th}$ eigenvalue and $i^{th}$ eigenvector respectively; i = 1,2,3. Now $\Sigma_{11}^{-1/2} = \sum_{i=1}^{3} \lambda_i^{-1/2} \gamma_i \gamma_i'$. The eigenvalues $\boldsymbol{\Sigma}_{11}$ are 624.93238, 44.487838 and 2.4429359. The corresponding eigenvectors are: (0.987172, 0.1120006, -0.113786)'; (0.1150796, -0.005131, 0.993343)'; (-0.110671, 0.9936949, 0.017954)'. Now

$$\boldsymbol{\Sigma}_{11}^{-1/2} = \begin{bmatrix} 0.0488043 & -0.066027 & 0.0113741 \\ -0.066027 & 0.6322628 & 0.0101406 \\ 0.0113741 & 0.0101406 & 0.1486615 \end{bmatrix}.$$

Step 3: Compute $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2} = \mathbf{F}$ (say). In this case it is

$$\mathbf{F} = \begin{bmatrix} 0.2033438 & 0.2611853 & -0.044758 \\ 0.2611853 & 0.4540921 & -0.083341 \\ -0.044758 & -0.083341 & 0.0210455 \end{bmatrix}.$$

Step 4: Obtain eigenvalues and eigenvectors of **F**. The eigenvalues are 0.6332992, 0.040223, 0.005266 and the corresponding eigenvectors are
$\mathbf{a}_1' = (0.524930, 0.837384, -0.152437)'$ ;
$\mathbf{a}_2' = (0.847489, 0.497640, 0.184708)'$ ;
$\mathbf{a}_3' = (0.078813, 0.226147, 0.970900)'$ respectively.

Step 5: On the similar lines of Steps 2, 3 and 4 obtain $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2} = \mathbf{G}$, its eigenvalues and eigenvectors.

$\mathbf{G} = \begin{bmatrix} 0.2634144 & -0.013700 & -0.001758 \\ -0.013700 & 0.0205081 & -0.008377 \\ -0.001758 & -0.0083771 & 0.0248539 \end{bmatrix}$. The eigenvalues are 0.6332992, 0.040223, 0.005266 and the corresponding eigenvectors are
$\mathbf{b}_1' = (0.297668, 0.926878, -0.228672)'$ ;
$\mathbf{b}_2' = (-0.247629, 0.3062953, 0.9191642)'$ ;
$\mathbf{b}_3' = (0.9219943, -0.216980, 0.3206965)'$ respectively

Step 6: The first pair of canonical variables is now given by $\mathbf{a}_1'\Sigma_{11}^{-1/2}\mathbf{X}$ and $\mathbf{b}_1'\Sigma_{22}^{-1/2}\mathbf{Y}$. Here $\mathbf{X}$ denotes the matrix of physiological variables and $\mathbf{Y}$ denote the matrix of exercise variables. Similarly other two pairs of canonical variables can be obtained.

$\mathbf{a}_1'\Sigma_{11}^{-1/2}$ = [-0.031405 0.4932416 -0.008199];
$\mathbf{a}_2'\Sigma_{11}^{-1/2}$ = [0.0763195 -0.368723 0.032052];
$\mathbf{a}_3'\Sigma_{11}^{-1/2}$ = [-0.007735 0.1580337 0.1457322].

$\mathbf{b}_1'\Sigma_{22}^{-1/2}$ = [0.066114 0.0168462 -0.013972];
$\mathbf{b}_2'\Sigma_{22}^{-1/2}$ = [-0.071041 0.0019737 0.0207141];
$\mathbf{b}_3'\Sigma_{22}^{-1/2}$ =[0.2452754 -0.019768 0.0081675].

The canonical correlations between three pairs of canonical variables can be obtained by taking the square root of the eigenvalues of $\mathbf{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}}$ or $\mathbf{\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}}$. Here the canonical correlations are 0.795608, 0.200556 and 0.072570.

**E2)** Step 1: Variance-covariance matrix of five yield attributing parameters X1, X2, X3, X4 and X5) is $\mathbf{\Sigma}_{11}$ and is given by $5\times5$ cells in left hand upper corner of the given variance covariance matrix. The variance-covariance matrix of four quality parameters Y1, Y2, Y3 and Y4 is $\mathbf{\Sigma}_{22}$ and is given by $4\times4$ cells in right hand bottom corner of the given variance covariance matrix. The covariance matrix between yield attributing parameters and quality attributes is $\mathbf{\Sigma}_{12}$ and is the $5\times4$ cells in right hand upper corner of the given variance covariance matrix.

Step 2: To obtain $\mathbf{\Sigma}_{11}^{-1/2}$, first obtain eigenvalues and eigenvectors of $\mathbf{\Sigma}_{11}$. Let $\lambda_i$ and $\gamma_i$ denote $i^{th}$ eigenvalue and $i^{th}$ eigenvector respectively; i = 1,2,3. Now $\Sigma_{11}^{-1/2} = \sum_{i=1}^{3}\lambda_i^{-1/2}\gamma_i\gamma_i'$. The eigenvalues $\mathbf{\Sigma}_{11}$ are 3.1758116, 0.7417691, 0.6572676, 0.275192 and 0.1499598.
The corresponding eigenvectors are:
(0.5157006 0.2095696 0.0655101 0.3701104 -0.740851)';
(0.4876162 -0.15341 -0.3956110 0.5454541 0.5335426)';
(-0.457400 -0.206258 0.5068357 0.7007362 0.0181492)';
(-0.349891 0.8468537 -0.3073330 0.2408609 0.0891511)' and
(0.4057652 0.4157429 0.6984732 -0.128269 0.39773710)'.
Now

$$\mathbf{\Sigma}_{11}^{-1/2} = \begin{bmatrix} 1.8839857 & -0.564091 & 0.3180696 & 0.0793549 & -0.576395 \\ -0.564091 & 1.6560546 & 0.4178776 & 0.2766612 & 0.1107616 \\ 0.3180696 & 0.4178776 & 1.4205354 & 0.0207793 & 0.0802538 \\ 0.0793549 & 0.2766612 & 0.0207793 & 1.1490046 & 0.0970125 \\ -0.576395 & 0.1107616 & 0.0802538 & 0.0970125 & 1.334717 \end{bmatrix}.$$

Step 3: Compute $\mathbf{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}=F}$ (say). In this case it is

$$\mathbf{F} = \begin{bmatrix} 0.2863075 & 0.0949999 & -0.119406 & -0.084734 & 0.2775501 \\ 0.0949999 & 0.2120403 & -0.238331 & -0.18475 & 0.0624013 \\ -0.119406 & -0.238331 & 0.2852454 & 0.1846082 & -0.069907 \\ -0.084734 & -0.18475 & 0.1846082 & 0.1997659 & -0.089549 \\ 0.2775501 & 0.0624013 & -0.069907 & -0.089549 & 0.3175887 \end{bmatrix}.$$

Step 4: Obtain first four eigenvalues and eigenvectors of **F**. The eigenvalues are 0.8265343, 0.4002511, 0.0651783, 0.0089843 and the corresponding eigenvectors are

$\mathbf{a}_1' = (0.4743674, 0.4263205, -0.485436, -0.396704, 0.4474416)'$;

$\mathbf{a}_2' = (0.4682176, -0.392238, 0.428314, 0.2902513, 0.599352)'$;

$\mathbf{a}_3' = (0.3723098, 0.0200515, -0.485523, 0.736616, -0.287484)'$;

$\mathbf{a}_4' = (0.6273877, 0.1540741, 0.4326632, -0.259644, -0.572742)'$ respectively.

Step 5: On the similar lines of Steps 2, 3 and 4 obtain

$\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2} = \mathbf{G}$, its eigenvalues and eigenvectors.

$$\mathbf{G} = \begin{bmatrix} 1.1354777 & -0.073763 & 0.2631705 & -0.124552 \\ -0.073763 & 1.0984645 & 0.2394113 & 0.0889018 \\ 0.2631705 & 0.2394113 & 1.1877782 & 0.0714007 \\ -0.124552 & 0.0889018 & 0.0714007 & 1.0378877 \end{bmatrix}.$$

The eigenvalues are 0.8265343, 0.4002511, 0.0651783, 0.0089843 and the corresponding eigenvectors are

$\mathbf{b}_1' = (0.6527588, 0.4359061, -0.616865, 0.0580444)'$;

$\mathbf{b}_2' = (0.7101565, -0.072348, 0.6880977, -0.13025)'$;

$\mathbf{b}_3' = (-0.258398, 0.8350398, 0.279248, -0.397442)'$;

$\mathbf{b}_4' = (-0.05305, 0.3278111, 0.2608056, 0.90648)'$ respectively.

Step 6: The first pair of canonical variables is now given by $\mathbf{a}_1'\Sigma_{11}^{-1/2}\mathbf{X}$ and $\mathbf{b}_1'\Sigma_{22}^{-1/2}\mathbf{Y}$. Here **X** denotes the matrix of yield attributing parameters and **Y** denote the matrix of quality attributes. Similarly other three pairs of canonical variables can be obtained.

The canonical correlations between four pairs of canonical variables can be obtained by taking the square root of the eigenvalues of $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ or $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Here the canonical correlations are 0.9091, 0.6327, 0.2553 and 0.0948.

**E3)** Proceeding on the same steps as in E2), one can easily see that the canonical correlation between first pair of canonical variables is 0.95. Which is also obvious.

# UNIT 27 CONJOINT ANALYSIS

**Structure** **Page No.**



## 27.1 INTRODUCTION

Conjoint Analysis is a popular marketing research technique that helps the marketers in understanding how people make choices between products or services or a combination of product and service. This helps in designing new products or services that better meet customers' underlying needs. Conjoint analysis has been found to be an extremely powerful of way of capturing what really drives customers to buy one product over another and what customers really value. We shall start our unit by discussing the developments in conjoint analysis in Section 27.2.

**Conjoint analysis** is a statistical technique used in market research to determine how people value different features that make up an individual product or service. The objective of conjoint analysis is to determine what combination of a limited number of attributes is most influential on respondent choice or decision making. A controlled set of potential products or services is shown to respondents and by analyzing how they make preferences between these products or services, the implicit valuation of the individual elements making up the product or service can be determined. These implicit valuations (utilities or part-worths) can be used to create market models that estimate market share, revenue and even profitability of new designs, which is discussed in Section 27.3. In Section 27.4, we shall focus on procedures of conjoint analysis and in Section 27.5, we shall describe the conjoint analysis models. In Section 27.6, use of fractional factorial plans in conjoint analysis is discussed with an example. At the end of this unit, we shall explain the application and disadvantages of conjoint analysis in Section 27.7.

### Objectives

After studying this unit, one should be able to

- understand the concept and need of conjoint analysis
- explain the steps involved in conjoint analysis'
- apply conjoint analysis in different marketing needs

## 27.2 DEVELOPMENTS IN CONJOINT ANALYSIS

Conjoint originated in mathematical psychology and was developed by marketing professor Paul Green at the Wharton School of the University of Pennsylvania and Data Chan. Other prominent conjoint analysis pioneers include Richard Johnson (founder of Sawtooth Software) who developed the Adaptive Conjoint Analysis technique in the 1980s and Jordan Louviere who invented and developed Choice-based approaches to conjoint analysis and related techniques such as MaxDiff.

**Early Years - Full Profile and Part-Worths**
Conjoint analysis has its earliest roots in psychology and the testing of the theory that when people take decisions the result is the 'sum' of all the bits of value for each part in that decision. So when we buy a computer, there is a difference in value between computers with a small screen compared to a large screen, for instance. The difference in value to the customer can be linked just to this one feature and by adding up different values in a kind of 'configurator' style you can come to a conclusion about the overall value of the product or service.

This seemed sensible in theory, but to test it required a method for breaking a product down into constituent parts, building profiles from these parts, then gathering preference data, then finally testing untried combinations to see if the customer preference was as per expectation. This is then the root of conjoint design. A demonstration can be seen here.

The first such tests were called full-profile designs (since the respondent saw profiles with one of each type of part) and were designed on cards using some form of ranking exercise typically. A major hurdle was in reducing the number of profiles down to a manageable number. This drew on work from statistics looking at experimental design - how do you minimize the number of experiments required to test a number of combinations of properties, for instance in drug development? The result was the use of 'orthogonal fractional factorial plans' in building the profiles for testing. Since the objective was to validate the decision making process, tests were carried out and analyzed and then compared to 'hold-out' profiles. These were profiles that were not counted for statistical accuracy but were required to validate the model.

The method of analysis initially was analysis of variance (ANOVA) to produce a statistical model to predict the preference drivers. Early studies evaluated attributes as continuous (eg. price) or discrete (eg. colour), but it soon became apparent that it was more effective to treat all variables as if they were discrete variables as even 'linear' attributes such as price were often non-linear in nature. The result of the analysis was the calculation of 'part-worths'. That is the model betas which describe how much each variable contributes to the final model. A higher beta is being more important. Since these part-worths have no units and because they are predicting an abstract entity such as preference, it is possible to scale and adjust the parameters without affecting the underlying model outcomes. This meant many of the consultancies that started to use conjoint turned these part-worths into the more user-friendly 'utilities'.

**Developments of computer learning techniques**
Research into conjoint type techniques continues to develop. The advent of online research means that computer-based techniques such as adaptive are much lower cost than when computer interviewing was carried out face-to-face. Consequently there is an ongoing development of computer learning techniques using elements such as genetic algorithms and evolutionary learning to 'search' the decision making space with the consumer and so make judgments about people would make decisions. In addition the development of product configurations, such as those used by Dell selling computers, means that there are other ways of getting at choice data.

There is still also controversy at the academic level since there are still some very fundamental assumptions being made about the types of models that people use - we blithely treat everything as linear functions and often disregard non-linear factors such as diminishing returns.

## 27.3 CONJOINT ANALYSIS DESIGN

Conjoint analysis and trade-off studies are amongst the most sophisticated forms of market research because they provide estimates of demand based on product or service design. But like any form of research, the quality of the output depends heavily on the quality of the design. For conjoint analysis, in particular this means choosing the right

flavour or type of conjoint to use and ensuring that the design of the attributes and levels meets the research, analysis and business requirements.

The basis of conjoint design splits into three interrelated parts viz.

1. Choice of the flavour or type of conjoint analysis that is to be used. Different types of conjoint analysis have different strengths and weaknesses. Some forms can be carried out on paper, some require the use of computer or internet-based interviews, some allow large complex designs, others are more robust in the face of issues like pricing;
2. The attributes and levels that make up the product.
3. The sample itself and the contact methods available where the conjoint interview can be carried out face-to-face, or over the internet, or by post and phone, and how long will the questionnaire be.

**Attributes and Levels in Conjoint Analysis**

Attributes and levels form the fundamental basis of conjoint analysis. The idea is that a product or service can be broken down into its constituent parts - so for instance a mobile phone has a size, weight, battery life, size of address book, type of ring. Each of these elements making up a generic mobile phone is known as an attribute.

When we compare between mobile phones each will have a different specification on each of these attributes. We might have choices in terms of battery life between 12, 24, 36, 48 hours of battery life. Each of these options is known as a level of the battery life attribute. In another example, a car might have an attribute colour which is made up of the levels red, blue, green, white.

This breaking down of products and services into attributes and levels is an extremely powerful tool for examining what a business offers and what it should be offering. For new product development, combining this product breakdown with an understanding of what the customer values most means that the business can focus its efforts on the areas that find favour with customers.

In addition, each level can also be thought of as a performance target. If the customer, on the attribute 'delivery', wants the level "next day delivery" and we are perceived to be offering "48hr delivery" then there is a gap to make up. What is more with conjoint analysis, it is possible to calculate what the value of making up that gap is to the business. Compare this to the cost of doing it and the business can decide whether it is worth it.

In conjoint analysis, the attributes and levels have to behave in certain ways so as to make the results of the conjoint analysis valid and conjoint useful. The first and foremost point is each attribute should be independent. In other words, it should not overlap with other attributes. So, colour and fuel economy are clearly not related, so they can appear together. However, some things like "car shape" and "number of passengers" aren't independent and as such cannot be used. Certain attributes have halo-effect on others around them. For example, if one level is "gold-plated handle", many people would infer that the rest of the product is also of better quality when there is no other information to support this. The main difficulty this causes is that price and brand need to be treated extremely carefully in conjoint studies to produce valid results.

Each level also needs to be capable of being read and understood on its own. Although attributes are used to help break a product down and in analysis, when presented to respondents what respondent sees is the levels.

Independent and readable levels are important from an analysis point of view, but for the conjoint to be useful it also needs to ensure that the range of attributes cover all the areas that are important to the customer, and that the range of levels cover all possibilities. If we do want to focus on a particular element of choice, it is possible to use an "all other things being equal" scenario, but if we are looking at the type of car-radio this limitation would not tell us about the importance of the car-radio in the overall choice of car.

For many products, particularly in business markets, service can be more important than the actual product. By using both product and service attributes in the same conjoint it is possible to see how customers trade-off service against features. The care, however, has to be taken to balance the attributes to prevent biasing the outcome one way or another.

It is also vital that the levels used describe real and where possible measurable and actionable performance. We could run a conjoint with levels "very good fuel economy" "good fuel economy" "not so good fuel economy". But, as with the scales above, these do not mean anything. If we come out with "good fuel economy" and the customer wants "very good fuel economy" what do you do?

One good test is if we showed it to the shop-floor (or the call centre) would they know how to deliver it. A second test is whether it is a statement you could use on a packet or in a brochure.

Now, in the following section, we shall discuss procedures of conjoint analysis

## 27.4 PROCEDURES OF CONJOINT ANALYSIS

The various procedures of conjoint analysis are Adaptive conjoint analysis, Choice Based Conjoint Analysis, Discrete Choice Analysis, Full Profile Conjoint Analysis, etc. A brief description of each one of them is given in the sequel.

1) **Hybrid Designs and Adaptive Conjoint Analysis (ACA)**
One of the most fundamental problems in conjoint is reducing the number of profiles that need to be evaluated by respondents. It was developed by Richard Johnson as a a computer based approach to conduct conjoint analysis which relied on initial self-explicated exercises where respondents pre-rank and pre-value items before undertaking the preference task.

Adaptive Conjoint Analysis (ACA) is one of two most common methods for carrying out conjoint analysis. The benefits of ACA are that it allows for a large number of attributes (up to 30) and levels (up to 7 per attribute) to be used. However, ACA does require a computer-based interview and the large number of attributes means that it is common for an ACA interview to last 45 minutes or more. In addition, some of the methods it uses to simplify the task of working out utilities mean that some care is needed in choosing and designing the attributes in order to get reliable results.

Technically ACA is known as a hybrid technique as it contains elements of 'self-explication' followed by the trade-off tasks themselves. ACA itself is produced by Sawtooth Software and can be conducted face-to-face or on-line. Telephone use of ACA is difficult and paper-base questionnaires are not possible.

2) **Choice Based Conjoint Analysis - CBC**
Choice-based conjoint analysis (CBC) is the most common alternative ACA. Design, implementation and calculation in CBC are completely different from ACA.

ACA has respondents selecting from products described with two or three attributes, CBC shows full descriptions using all the attributes available. In addition, CBC can show more than just two "products" at the same time, together with a none-of-these option enabling more realistic choice decisions to be evaluated.

A respondent takes a long time in CBC. As a consequence, choice-based conjoint is limited to 5-7 attributes in contrast to 25-30 attributes for ACA.

An additional twist is that utilities and importance in CBC are calculated across a sample as a whole, whereas for ACA we get utilities and importance for each individual in the sample. Combined with the lower number of attributes, this means that choice-based studies require far shorter questionnaires (15-20 minutes) and can be designed to be purely paper-based.

The advantages choice-based conjoint gives us are greater robustness of results - particularly for pricing work (although there are ways of getting around ACA's pricing limitations), combined with shorter and therefore less costly fieldwork. It is also favoured for it's rigour academically.

The disadvantages are the lower number of attributes that are possible, and the lack of individual level utilities.

3) **Discrete Choice Analysis**

A more advanced form of choice-based conjoint is Discrete Choice Analysis. It is also known as "stated preference research". DCA studies are particularly popular for transportation studies looking at modal choice - the preference between a train, car and airline for instance. The main difference from CBC is the inclusion of continuous variables such as price and time. This allows the ability to examine the varying costs of the ticket with varying times taken to travel and so to establish the value of time for the journey. This enables transport economists to make statements like "2cm extra leg room is worth 10 minutes longer journey time or £40 extra fare" or "an extra train every 15 minutes would encourage x% of car drivers to switch to the train".

4) **Full Profile Conjoint Analysis**

An additional option that dates back a long time but that is still used is full profile conjoint analysis. Full-profile is the original form of conjoint and is still in use, though predominantly in the US it would appear. Full profile conjoint anlaysis makes use of a more limited number of attributes to describe the product or service, but sufficient treatments are shown to one respondent to enable individual level utilities to be calculated. A fractional factorial design is used to specify a fixed set of profiles that need to be shown for analysis. The difficulty is that this does limit the number of attributes quite severely. However these old school studies are still popular for simple, non-computer-based conjoint projects and are most common for students learning about conjoint for the first time.

Most conjoint analysts fit what is known as part-worth model to respondents' evaluating judgments, whether obtained by trade-off tables or full profile, self explicated, or hybrid approaches.

Now, try an exercise.

---

E1) List any two procedures of conjoint analysis. Also, give one example of each.

---

Data for conjoint analysis is most commonly gathered through a market research survey. Conjoint analysis can also be applied to a carefully designed configurator or data from an appropriately design test market experiment. Market research rules of thumb apply with regard to statistical sample size and accuracy when designing conjoint analysis interviews.

The length of the research questionnaire depends on the number of attributes to be assessed and the method of conjoint analysis in use. A typical Adaptive Conjoint questionnaire with 20-25 attributes may take more than 30 minutes to complete. Choice based conjoint, by using a smaller profile set distributed across the sample as a whole may be completed in less than 15 minutes. Choice exercises may be displayed as a store front type layout or in some other simulated shopping environment.

## 27.5 CONJOINT ANALYSIS MODELS

Let p = 1,2,…, t denote the set of *t* attributes that are used in the study design. Let $y_{ip}$ denote the level of $p^{th}$ attribute for the $i^{th}$ stimulus. First we assume that $y_{ip}$ is inherently continuous. The vector model assumes that the preferences $s_j$ for the $j^{th}$ stimulus is given by

$$s_j = \sum_{p=1}^{t} w_p y_{jp}$$

where $w_p$ denotes a respondent's weight for each of the $p^{th}$ attributes.

The ideal point model shows that preference $s_j$ is negatively related to the weighted squared distance $d_j^2$ of the location $y_{ip}$ of the $j^{th}$ stimulus from the individual's ideal point $x_p$ where $d_j^2$ is defined as

$$d_j^2 = \sum_{p=1}^{t} w_p (y_{jp} - x_p)^2$$

The part-worth model assumes that

$$s_j = \sum_{p=1}^{t} f_p (y_{jp})$$

where $f_p$ is a function denoting the part-worth of difference levels of $y_{ip}$ for the $p^{th}$ attribute. In practice $f_p(y_{jp})$ is estimated for a selected set of discrete levels of $y_{ip}$.

Any number of algorithms may be used to estimate utility functions. These utility functions indicate the perceived value of the feature and how sensitive consumer perceptions and preferences are to changes in product features. The actual mode of analysis will depend on the design of the task and profiles for respondents. For full profile tasks, linear regression may be appropriate, for choice based tasks, maximum likelihood estimation, usually with logistic regression are typically used. The original methods were monotonic analysis of variance or linear programming techniques, but these are largely obsolete in contemporary marketing research practice. Conjoint analysis makes extensive use of Orthogonal Arrays (Addelman, 1962) to reduce the number of stimulus descriptions to a small fraction of the total number of descriptions.

## 27.6 USE OF FRACTIONAL FACTORIAL PLANS IN CONJOINT ANALYSIS: AN EXAMPLE

In Conjoint Analysis the profile of different products are presented to the consumers for their responses. These profiles are generated by varying the levels of its attribute. For example, suppose we are conducting a Conjoint Analysis based study of dish

washers. Let us assume that the most important attributes considered by its customers are Brand, Price, Washing Capacity, Colour and Shape. Let us further assume that the following levels of attributes are considered relevant and interesting by the marketer for the study.

| S.No. | **Attribute** | **Levels** |
|---|---|---|
| 1. | Brand | Kitchen Master; Elegant; Torrent Evermaid |
| 2. | Price | Rs. 15,000/-; Rs. 20,000/-; Rs. 25,000/- |
| 3 | Washing Capacity | High; Medium; Low |
| 4. | Colour | Steel grey; White; Light Blue; Pink |
| 5. | Shape | Cylindrical; Box Type |

Since the 5 attributes can take 4, 3, 3, 4 and 2 levels, the total number of possible product concepts that can be generated by configuring these attributes is $4\times3\times3\times4\times2=288$. In order to determine the part worth utilities of each of the levels of all these attributes, we shall have to take 288 different product concepts for getting his/her responses. This number is certainly too large for any consumer. Therefore, we resort to the method of Fractional Factorial Design of Experiment to make it manageable.

The statistical technique of Fractional Factorial Design of Experiment finds out the minimum number of product designs which are necessary to use in the study and yet provide us all the information that we originally sought. These designs are also mutually independent (orthogonal) to avoid any redundancy in the data and allow the representation of each of the attributes and their respective levels in an unbiased manner.

In the example of dish washer considered here, this technique has given us only 16 designs out of the 288 possible dish washers. However, it should be noted that such reduction in number of product designs is possible only after making certain assumptions. For example, we had assumed that none of the attributes interact among themselves. Or in other words, the attributes are considered to be independent of each other. Only under this assumption we got the number of product concepts as 16. At the other end, if we would have allowed all the attributes to interact with each other the required number of product concepts would have remained as 288. With different types of assumptions the number of concepts required would be in between these extremes.

The 16 product concepts found through this method are not unique. Many other sets of 16 cards would have also been equally good. However, all of these sets would have to be independent and represent all the attributes and their respective levels in an unbiased manner. We are illustrating below one such set of 16 cards representing the product concepts of dish washers using Fractional Factorial Design of Experiment.

**Table 1: Description of Product Concept Cards**

| **Card#** | **Brand** | **Price** | **Cap.** | **Colour** | **Shape** |
|---|---|---|---|---|---|
| Card 1 | Kit Master | Rs. 15,000 | High | St. Grey | Cylinder |
| Card 2 | Elegant | Rs. 25,000 | Medium | St. Grey | Box Type |
| Card 3 | Torrent | Rs. 20,000 | Medium | St. Grey | Box Type |
| Card 4 | Ever Maid | Rs. 20,000 | Low | St. Grey | Cylinder |
| Card 5 | Kit Master | Rs. 20,000 | Medium | White | Box Type |
| Card 6 | Elegant | Rs. 20,000 | Low | White | Cylinder |
| Card 7 | Torrent | Rs. 25,000 | High | White | Cylinder |
| Card 8 | Ever Maid | Rs. 15,000 | Medium | White | Box Type |

| | | | | | |
|---|---|---|---|---|---|
| Card 9 | Kit Master | Rs. 25,000 | Low | Light B1 | Box Type |
| Card 10 | Elegant | Rs. 15,000 | Medium | Light B1 | Cylinder |
| Card 11 | Torrent | Rs. 20,000 | Medium | Light B1 | Cylinder |
| Card 12 | Ever Maid | Rs. 20,000 | High | Light B1 | Box Type |
| Card 13 | Kit Master | Rs. 20,000 | Medium | Pink | Cylinder |
| Card 14 | Elegant | Rs. 20,000 | High | Pink | Box Type |
| Card 15 | Torrent | Rs. 15,000 | Low | Pink | Box Type |
| Card 16 | Ever Maid | Rs. 25,000 | Medium | Pink | Cylinder |

- **Physical Design of Stimuli**

After selecting the product concepts required for Conjoint Analysis study, they need to be exposed to the consumers as stimuli. This may be done in a variety of ways mainly depending on the demands of the situation and the convenience of the researcher. Of course, it would be most desirable to present real life prototypes of the products according to the product concepts specified. These, may be given to the consumers for their usage or trials. But, such extreme ways of presenting the products may not always be possible or even necessary. In such cases, product models, diagrams or even verbal descriptions may be adopted. In our example of dish washers, it may not be possible to produce the 16 prototypes and take them to the consumers. Just their models or pictures may be sufficient.

- **Data Collection**

Ease of data collection is a key feature of Conjoint Analysis. The consumers are asked only to assign rating scores to each of the product stimuli or even rank the different concepts presented to them. This is quite a realistic task and is close to the shopping experiences where the customer merely makes choices. He does not have to respond to each of the attributes separately.

This feature of conjoint analysis is possible due to the use of Fractional Factorial Design of Experiment before collection of data and the use of Conjoint Analysis after collecting the data. In other words, the use of the technique eases the burden of the respondents.

- **Determination of part worth utilities**

The rating or ranking data obtained from the consumers are analyzed next. Two methods are more popular for this purpose. In one method, the part worth utility for each of the levels of each attributes are arbitrarily assigned. Based on these assumed values, consumers overall rating or ranking (as the case may be) are estimated. These estimated responses may understandably, be quite different from the actual data. After a few iterations convergence is achieved so that the part worth utilities found approximate the estimate responses to the actual data best.

In the alternative method, the part worth utilities are derived in one step. Here, an error function describing the difference between the estimated and actual data is defined. This function is then minimized.

After using any of the available method, the output is obtained for each of the respondent separately. This is quite significant as the disaggregate data can be combined in any of the desired way. But, if the output was only at the aggregate level then disaggregation might not have been possible.
In our example of dish washer, the part worth utilities may be found for each of the attributes and their levels as following:

**Table 2: Attributes with their levels**

| | **Attribute** | **Levels** | **Part Worth Utility** |
|---|---|---|---|
| 1. | Brand (25.4%) | Kitchen Master | 2.2 |
| | | Elegant | -2.3 |

| | | | |
|---|---|---|---|
| | | Torrent | -0.8 |
| | | Ever Maid | 0.9 |
| 2. | Price (66.2%) | Rs. 15,000 | 5.0 |
| | | Rs.20,000 | -0.1 |
| | | Rs. 25,000 | -5.8 |
| 3. | Washing Capacity (4.9%) | High | 0.0 |
| | | Medium | 0.1 |
| | | Low | -0.1 |
| 4. | Colour (2.8%) | Steel Grey | -0.3 |
| | | White | -0.1 |
| | | Light Blue | 0.1 |
| | | Pink | 0.2 |
| 5. | Shape (0.7%) | Cylindrical | -0.1 |
| | | Box Type | 0.1 |

From the above table, we find that price plays most important role (66.2%) in the minds of customers. This is followed by Brand (25.4%), Washing Capacity (4.9%), Colour (2.8%) and Shape (0.7%). These relative importance values for the maximum and minimum values of the part, worth utilities of the respective attributes. Customers respond quite predictably towards different price levels. They prefer lesser price to the higher ones. Among brands, we find that Kitchen Master is the most liked brand and Elegant in the least liked one. However, they do not prefer the "High" capacity most. They prefer "Medium" size most. In terms of the colour, they prefer Pink colour most. Between the two shapes, box type dish washers are preferred more. But, the importance scores for the wash capacity, colour and shape themselves are quite low.

The part worth utilities can now be added to determine the total utility for each of the possible product concepts. This allows us to scan the consumer preference pattern for all of the 288 product concepts although he has been exposed to only 16 of them. We can now also rank all of these 288 product concepts.

## 27.7 APPLICATIONS OF CONJOINT ANALYSIS

Calculation of the part worth utilities becomes just the starting point for many interesting applications of conjoint analysis. The important ones among them are described next:

1) **Optimum Product Design:** Since all possible product concepts can be compared after adding their respective attribute levels part worth utilities, it is possible to determine the demand for different products out of any given set of available products in the market place. The demand levels can be converted into profit figures as cost of producing and marketing can also be calculated. These cost calculations are possible as the volume of operations and the features of the products are now known. Thus, the optimum product can be chosen from the profits point of view (or any of the other given management's objective). Customers differential rates of purchase of products are also duly considered at this stage.

   Quite often, a manager may like to know the effects of slight change in any of the attribute by his own company or the competitor's. Conjoint Analysis allows this kind of "What if" analysis very easily with the data base of part worth utilities. In fact, different kinds of scenarios can be simulated and the manager can optimize not only the product but other aspects of his marketing strategy. Similarly, whenever there is any change in competitor’s actions or in the environment a fresh scenario can be drawn for simulation. Of course, the simulation shall be limited to the attributes considered in the analysis. This feature does also help in increasing the shelf life of the conjoint analysis output.

2) **Market segmentation:** Since the Conjoint Analysis is done at the individual customer level, the individual customer's identity can be retained through out the analysis. Thus, customers can be segmented according to their sensitivities to different product attributes.

   It is also possible to identify the customers segments, which would be attracted most for the proposed product position. This helps in having a focused matching between the chosen product position and the target customer segment. It can also help in identifying that part f competitor's market which needs to be poached for snatching market share from them. Similarly, the same type of analysis can be done to identify the most vulnerable section of one's own market segment.

   Sometimes, an additional product offer appears to be quite attractive. But, this may be at the cost of cannibalisation. Conjoint Analysis can help in estimating the effects of cannibalisation as well. Thus, it helps in maximizing net profits of the organization.

3) **SWOT Analysis:** First of all, the part worth utility of the brand itself can tell about the relative brand strength. Similarly by looking at the other features of one's own and competitor's offers Conjoint Analysis enables the marketers to conduct his detailed SWOT Analysis.

4) **Estimating Customer Level Brand Equity:** Conjoint Analysis is a good bridge between the consumer level perceptions and the financial worth of the offers. This can be used for estimating the important parameter of brand equity at the consumers level. There is scope of differentiating the “Loyal”, “Acceptors” and “Switchers” for more accurate calculations of brand equity.

Now let us discuss the advantages and disadvantages of conjoint analysis.

**Advantages**

- It estimates psychological tradeoffs that consumers make when evaluating several attributes together.
- It measures preferences at the individual level.
- It uncovers real or hidden drivers which may not be apparent to the respondent themselves.
- It is realistic choice or shopping task.
- It is able to use physical objects.
- If it is appropriately designed, the ability to model interactions between attributes can be used to develop needs based segmentation.

**Disadvantages**

- Designing the conjoint studies can be complex with too many options, respondents resort to simplification strategies.
- It is difficult to use for product positioning research because there is no procedure for converting perceptions about actual features to perceptions about a reduced set of underlying features.
- The respondents are unable to articulate attitudes toward new categories, or may feel forced to think about issues they would otherwise not give much thought to poorly designed studies may over-value emotional/preference variables and undervalue concrete variables.
- It does not take into account the number items per purchase so it can give a poor reading of market share

Inspite of lot of developments in the field of conjoint anlaysis and its widespread use, one can still find market research professionals who have no experience of conjoint analysis, or who still rely on self-explication to try for the prediction of the consumer behaviour.

Now, try the following exercises.

E2) State conjoint analysis and state its potential applications.

E3) What are the steps involved in a conjoint analysis? Explain with the help of examples.

E4) Write three advantages and three disadvantages of conjoint analysis.

Now, let us summarize the unit.

## 27.8 SUMMARY

In this unit, we have covered the following points.

1) **Conjoint analysis** is a statistical technique used in market research to determine how people value different features that make up an individual product or service. The objective of conjoint analysis is to determine what combination of a limited number of attributes is most influential on respondent choice or decision making.

2) The main steps involved in using conjoint analysis include determinationof salient attributes for the given product from the points of view of consumer, assigning a set of discrete levels or a range of continuous values to each of the attributes, utilizing fractional factorial plan for desgning the stimuli of the experiment, physically designing the stimuli, data collection and determination of part worth utilities.

3) Some of the procedures of conjoint analysis are Adaptive conjoint analysis, Choice Based Conjoint Analysis, Discrete Choice Analysis, Full Profile Conjoint Analysis, etc.

4) Data for conjoint analysis is most commonly gathered through a market research survey. Conjoint analysis can also be applied to a carefully designed configurator or data from an appropriately design test market experiment.

5) The possible application of conjoint analysis include product design, market segmentation, SWOT analysis, etc.