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 Σ 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 λ, 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 λ, 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 λ must be chosen that

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

Thus a non-zero solution for Equation (4) exists if and only if λ 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 λ to be largest eigenvalue, namely λ_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 λ and Γ. 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 Γ 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 Σ 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 Σ

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 λ_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 λ_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} = $ Constant into the ellipsoid $\mathbf{Z}'\boldsymbol{\Lambda}^{-1} =$ 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.

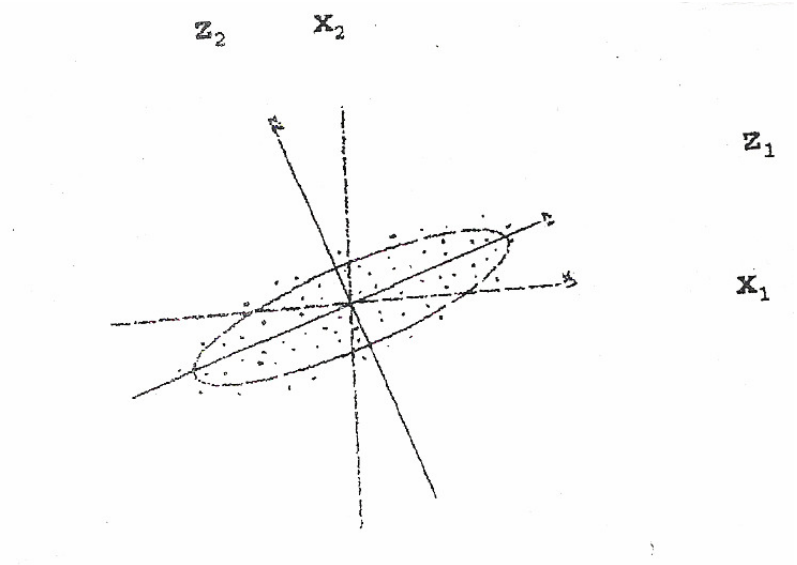

Fig. 1: Geometrical Interpretation of Principal Components

If the eigenvalues of Σ 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 Σ. 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.