UNIT 16 INTRODUCTION TO MULTIVARIATE ANALYSIS

Structure

16.0 OBJECTIVES

After going through this unit, you will be able to:

- explain the concept of multivariate analysis;
- apply the specific techniques used under multivariate analysis;
- decide on the technique to be used in a research problem; and
- describe the statistical issues involved in multivariate data analysis.

16.1 INTRODUCTION

Multivariate analysis involves a set of techniques to analyse data sets on more than one variable. Many of these techniques are modern and often involve quite sophisticated use of computing tools. Such analyses refer to all statistical methods that simultaneously analyse multiple measurements on each individual or object under investigation. Hence, any analysis simultaneously involving analysis of more than or equal to two variables can loosely be considered multivariate analysis. This unit will provide a list of such analyses in order to help decide when to use a given statistical technique for a given type of data or statistical question. It also gives a brief description of each technique. It is organized according to the number of data sets to analyze: one or two (or more). With two data sets we consider two cases: in the first case, one set of data plays the role of predictors or independent variables and the second set of data corresponds to measurements or dependent variables; in the second case, the different sets of data correspond to different sets of dependent variables.

Let us begin with analysis of situations involving a single dataset.

16.2 DEALING WITH ONE DATA SET

In case of one data set, the data tables to be analyzed are made of several measurements collected on a set of units (e.g., subjects). This implies that the

investigator is not really interested in the causal relationships between the dependent variable and the predictors or independent variables, but is interested in either creating groups or clusters of related variables or observations related to a single variable.

Interval or Ratio Level of Measurement: Principal Component Analysis

When faced with a large number of variables, principal component analysis (PCA) is a helpful measure to reduce the number of variables. PCA decomposes the entire data with correlated variables into a number of uncorrelated variables. These variables are called, depending upon the context, principal components, factors, eigenvectors, singular vectors, or loadings. Each unit is also assigned a set of scores, which correspond to its projection on the components. The results of the analysis are often presented with graphs plotting the projections of the units onto the components, and the loadings of the variables.

The importance of each component is expressed by the variance (i.e., eigenvalue) of its projections or by the proportion of the variance explained. Hence, PCA is also interpreted as an orthogonal decomposition of the variance (also called inertia) of a data table.

Nominal or Ordinal Level of Measurement: Correspondence Analysis, Multiple Correspondence Analysis

Correspondence Analysis (CA) is a generalization of PCA to contingency tables. The factors of correspondence analysis give an orthogonal decomposition of the Chi-square associated to the table. In correspondence analysis, rows and columns of the table play a symmetric role and can be represented in the same plot. When several nominal variables are analyzed, correspondence analysis is generalized as Multiple Correspondence Analysis (MCA). The correspondence analysis is also known as dual or optimal scaling or reciprocal averaging.

Similarity or Distance: Multidimensional Scaling, Additive Tree, Cluster Analysis

These techniques are applied when the rows and the columns of the data table represent the same units and when the measure is a distance or a similarity. The goal of the analysis is to represent graphically these distances or similarities. Multidimensional Scaling (MDS) is used to represent the units as points on a map such that their Euclidean distances on the map approximate the original similarities (classic MDS, which is equivalent to PCA, is used for distances, nonmetric MDS for similarities). Additive tree analysis and cluster analysis are used to represent the units as "leaves" of a tree with the distance \on the tree" approximating the original distance or similarity.

16.3 DEALING WITH TWO DATA SETS: ONE DEPENDENT AND ONE INDEPENDENT

When we are dealing with two data sets there could be two situations: one, we are dealing with one set of data taken as independent variables and the other data set as dependent variables; and two, both data sets are dependent variables. While dealing with one set of independent variables and one dependent variable we have several alternative scenarios. We can proceed with any of the following techniques.

Multiple Linear Regression Analysis

Multiple linear regression (MLR) is one of the strongest statistical tools to establish causal relationships between variables. In MLR, there are more than one independent

or explanatory variable (which are supposed to be regressed without error) that are used to predict a dependent variable. If the explanatory variables are orthogonal, the problem reduces to a set of univariate regressions. MLR can no longer be performed when explanatory variables become linearly dependent on each other. This leads to a problem that is popularly known as multicollinearity.

In MLR, if Y is the dependent variable, and X_1, X_2, ..., X_n are explanatory or independent variables, then the framework is expressed as:

$$Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3 + B_4 X_4 .... + B_n X_n + u \qquad ...(16.1)$$

where u reflects the random disturbance term with mean zero and constant variance.

There could be situations where we have to deal with regression models with too many predictors and/or several dependent variables. In such situations the problem of multicollinearity is likely to come up.

Partial Least Squares Regression

Partial least squares regression (PLSR) is used when the goal is to predict or explain more than one dependent variable. This technique is extremely versatile in use. One of the modes of addressing the multicollinearity problem is by partial least squares regression (PLSR). PLSR addresses the multicollinearity problem by computing latent vectors (similar to the components of PCA) that explains both the explanatory variables and the dependent variables. Hence, in that sense, it combines the characteristics of PCA and Multiple Linear Regression. The score of the units as well as the loadings of the variables can be plotted as in PCA, and the dependent variables can be estimated (with a confidence interval) as in MLR.

Principal Component Regression

This is an interesting method often used when data suffers from multicollinearity, or when there are too many variables to be dealt with in an analysis that makes the analysis complicated, we often expose the explanatory variables to a process of Principal Component Analysis. Eventually, after obtaining the scores of the units, the units are being treated as explanatory variables, and the dependent variable is regressed over the explanatory variables.

Ridge Regression

Ridge Regression accommodates the multicollinearity problem by adding a small constant (the ridge) to the diagonal of the correlation matrix. This makes the computation of the regression estimates possible.

Reduced Rank Regression or Redundancy Analysis

In reduced rank regression (RRR), the dependent variables are first submitted to a PCA and the scores of the units are then used as dependent variables in a series of standard MLR's where the original independent variables are used as predictors (a procedure akin to an inverse principal component regression).

Multivariate Analysis of Variance

Multivariate analysis of variance (MANOVA) is a technique to assess group differences across multiple metric dependent variables simultaneously, based on a set of categorical (non-metric) variables acting as independent variables. It provides information on the nature and predictiv power of the independent measures, as well as the relationships and differences seen in the dependent measures. MANOVA involves a *structured* method to specify the comparisons of group differences for the dependent variables and still maintain statistical efficiency. In MANOVA the explanatory variables have structure similar to that of a standard ANOVA, and are

used to predict a set of dependent variables, MANOVA computes a series of ordered orthogonal linear combinations of the dependent variables (i.e., factors) with the constraint that the first factor generates the largest F if used in an ANOVA. The sampling distribution of this F is adjusted to take into account its construction.

16.4 PREDICTING A NOMINAL VARIABLE: DISCRIMINANT ANALYSIS

Discriminant analysis (DA) helps to determine which variables discriminate between two or more naturally occurring groups. Mathematically equivalent to MANOVA, it is extensively used when a set of explanatory variables are used to predict the group to which a given unit belongs (which is a nominal dependent variables). It combines the explanatory variables in order to create the largest F when the groups are used as a fixed factor in an ANOVA.

The model is constructed with a set of observations for which the classes are known. The set of observations are sometimes referred to as the training set. Based on the training set, the technique constructs a set of linear functions of the predictors, known as discriminant functions, such that

$$L = b_1X_1 + b_2X_2 + \ldots + b_nX_n + c \qquad \ldots(16.2)$$

where the b's are discriminant coefficients, the x's are the input variables or predictors and c is a constant.

For example, an educational researcher may want to investigate which variables discriminate between high school graduates who decide (a) to go to college, (b) to attend a trade or professional school, or (c) to seek no further training or education. For that purpose the researcher could collect data on numerous variables prior to students' graduation. After graduation, most students will naturally fall into one of the three categories. Discriminant analysis could then be used to determine which variable(s) are the best predictors of students' subsequent educational choice.

16.5 FITTING A MODEL: CONFIRMATORY FACTOR ANALYSIS

Confirmatory factor analysis (CFA) seeks to determine whether the number of factors and the loadings of measured (indicator) variables on them conform to what is expected on the basis of pre-established theory. Indicator variables are selected on the basis of prior theory and factor analysis is used to see if they load as predicted on the expected number of factors. The researcher first generates one (or a few) model(s) of an underlying explanatory structure (i.e., a construct) which is often expressed as a graph. The researcher's *à priori* assumption is that each factor (the number and labels of which may be specified *à priori*) is associated with a specified subset of indicator variables. A minimum requirement of confirmatory factor analysis is that one hypothesize beforehand the number of factors in the model, but usually also the researcher will posit expectations about which variables will load on which factors (Kim and Mueller, 1978b: 55). The researcher seeks to determine, for instance, if measures created to represent a latent variable really belong together. The correlations between the dependent variables are fitted to this structure. Models are evaluated by comparing how well they fit the data. Variations over CFA are called structural equation modelling (SEM), LISREL, or EQS.

16.6 DEALING WITH TWO DATA SETS: TWO DEPENDENT VARIABLES SETS

Canonical Correlation Analysis

Canonical correlation analysis (CC) allows the investigation of the relationship between *two sets* of variables. For example, a sociologist may want to investigate the relationship between two predictors of social mobility based on interviews, with actual subsequent social mobility as measured by four different indicators. A medical researcher may want to study the relationship of various risk factors to the development of a group of symptoms. In all of these cases, the researcher is interested in the relationship between two sets of variables, and Canonical Correlation would be the appropriate method of analysis.

Canonical Correlation combines the dependent variables to find pairs, of new variables called canonical variables, CV, one for each data table having the highest correlation. However, the CV's, even when highly correlated, do not necessarily explain a large portion of the variance of the original tables. This makes the interpretation of the CV sometimes difficult, but CC is nonetheless an important theoretical tool because most multivariate techniques can be interpreted as a special case of CC.

Multiple Factor Analysis

Multiple factor analysis (MFA) combines several data tables into one single analysis. The first step is to perform a PCA of each table. Then each data table is normalized by dividing all the entries of the table by the first eigenvalue of its PCA. This transformation – akin to the univariate z-score of the normal distribution – equalizes the weight of each table in the final solution and therefore makes possible the simultaneous analysis of several heterogeneous data tables.

Multiple Correspondence Analysis

Correspondence analysis is an exploratory technique used to analyze simple two-way and multi-way tables containing measures of correspondence between the rows and columns of any given data. The results provide information almost similar to those produced by Factor Analysis techniques, and they allow us to explore the structure of categorical variables included in the table. Multiple correspondence analysis (MCA) is an extension of simple correspondence analysis to more than two variables. MCA can be used to analyze several contingency tables by generalizing CA.

PARAFAC and TUCKER3

Both these techniques are used for three-way data analysis. PARAFAC model is the simplest three-way model. These techniques handle three-way data matrices by generalizing the PCA decomposition into scores and loadings in order to generate three matrices of loading (one for each dimension of the data). They differ by the constraints they impose on the decomposition (TUCKER3 generates orthogonal loadings, PARAFAC does not).

Indscal

Indscal is used when each of several subjects generates a data matrix with the same units and the same variables for all the subjects. Indscal generates a common Euclidean solution (with dimensions) and expresses the differences between subjects as differences in the importance given to the common dimensions.

Statis

Statis is used when at least one dimension of the three-way table is common to all tables (e.g., same units measured on several occasions with different variables). The first step of the method performs a PCA of each table and generates a similarity table (i.e., cross-product) between the units for each table.

The similarity tables are then combined by computing a cross-product matrix and performing its PCA (without centering). The loadings on the first component of this analysis are then used as weights to compute the compromise data table which is the weighted average of all the tables. The original table (and their units) are projected into the compromise space in order to explore their communalities and differences.

Procustean Analysis

Procustean analysis (PA) is used to compare distance tables obtained on the same objects. The first step is to represent the tables by MDS maps. Then procustean analysis finds a set of transformations that will make the position of the objects in both maps as close as possible (in the least squares sense).

Check Your Progress 1

1) Explain he purpose of carrying out a discriminant analysis.

.....

.....

.....

.....

.....

2) Explain the following concepts:

a) Canonical Correlation Analysis

b) Multiple Factor Analysis

c) MANOVA

.....

.....

.....

.....

.....

16.7 LET US SUM UP

In this Unit we explained some of the techniques that can be used in analysis of multivariate data. There could be two situations where multivariate analysis is undertaken depending upon whether we have one data set or more than one data sets. There are several techniques available to researchers in each category. We have discussed the underlying ideas in each of these techniques in brief. This will serve as a prelude to the following two Units in the Block.

16.8 KEY WORDS

Ridge Regression : Ridge Regression accommodates the multicollinearity problem by adding a small constant (the ridge) to the diagonal of the correlation matrix. This makes the computation of the regression estimates possible.

Confirmatory factor analysis : It seeks to determine whether the number of factors and the loadings of measured (indicator) variables on them conform to what is expected on the basis of pre-established theory. Indicator variables are selected on the basis of prior theory and factor analysis is used to see if they load as predicted on the expected number of factors.

Multiple factor analysis : It combines several data tables into one single analysis. The first step is to perform a PCA of each table. Then each data table is normalized by dividing all the entries of the table by the first eigenvalue of its PCA.

16.9 SOME USEFUL BOOKS/ REFERENCES*

Borg I., and Groenen P., 1997, *Modern Multidimensional Scaling*, Springer-Verlag, New York.

Johnson R.A., & Wichern D.W., 2002, *Applied Multivariate Statistical Analysis*. Prentice-Hall, Upper Saddle River (NJ).

Kim, Jae-On and Charles W. Mueller, 1978, *Introduction to Factor Analysis: What it is and how to do* it, Quantitative Applications in the Social Sciences Series, No. 13. Sage Publications, Thousand Oaks, CA.

Naes T., and Risvik E. (Eds.), 1996, *Multivariate Analysis of Data in Sensory Science*. Elsevier, New York.

Weller S.C., and Romney A.K., 1990, *Metric Scaling: Correspondence Analysis*. Thousand Oaks, Sage Publications, CA.

16.10 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 16. 2 and Section 16.4 and answer.

2) See Section 16.5 and answer.

UNIT 17 PRINCIPAL COMPONENTS ANALYSIS

Structure

17.0 OBJECTIVES

After going through this unit, you will be able to:

- explain the basic principles of principal component analysis;
- reduce the dimensionality of a dataset; and
- identify new meaningful variables.

17.1 INTRODUCTION

In the context of multivariate data analysis, one might be faced with a large number of variables that are correlated with each other, eventually acting as a proxy of each other. This makes the coexistence of the variables in the framework redundant, thereby complicating the analyses. Under such circumstances, the investigator might be interested in reducing the dimensionality of the data set by identifying and classifying the commonality in the patterns of the related variables. Principal component analysis (PCA) is a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*.

17.2 PRINCIPAL COMPONENTS ANALYSIS

As one of the oldest multivariate statistical methods of data reduction, Principal Component Analysis (PCA)simplifies a dataset by producing a small number of derived variables that are uncorrelated and that account for most of the variation in the original data set. Eventually, the derived variables are combinations of the original variables. For example, it might be that students take 10 examinations and some students do well in one exam whilst other students do better in another. It is difficult to compare one student with another when we have marks from 10 examinations to consider. One obvious way of comparing students is to calculate the mean score. This is a constructed combination of the existing variables. However, we

may get a more useful comparison of overall performances by considering other constructed combinations of the 10 exam marks. The PCA is one way of constructing such combinations, doing so in such a way as to account for as much as possible of the variation in the original data. One can then compare students' performance by considering this much smaller number of variables.

The PCA is amongst the oldest of the multivariate statistical methods of data reduction. It is a technique for simplifying a dataset, by reducing multidimensional datasets to lower dimensions for analysis. It produces a small number of derived variables that are uncorrelated and that account for most of the variation in the original data set. By reducing the number of variables in this way, we can understand the underlying structure of the data. The derived variables are combinations of the original variables. For example, it might be that students take 10 examinations and some students do well in one examination while other students do better in another. It is difficult to compare one student with another when we have 10 marks to consider. One obvious way of comparing students is to calculate the mean score. This is a constructed combination of the existing variables. However, one might get a more useful comparison of overall performances by considering other constructed combinations of the 10 exam marks. The PCA is one way of constructing such combinations, doing so in such a way as to account for the maximum possible variation in the original data. We can then compare students' performance by considering this much smaller number of variables.

PCA states and then solves a well-defined statistical problem, and except for special cases always gives a unique solution with some very nice mathematical properties. We can even describe some very artificial practical problems for which PCA provides the exact solution. The difficulty comes in trying to relate PCA to real-life scientific problems; the match is simply not very good. Actually PCA often provides a good approximation to common factor analysis, but that feature is now unimportant since both methods are now easy enough.

17.2.1 Transformation of Data

PCA is a linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. The PCA can be used for dimensionality reduction in a dataset while retaining those characteristics of the dataset that contribute most to its variance, by keeping lower-order principal components and ignoring higher-order ones. Such low-order components often contain the "most important" aspects of the data. But this is not necessarily the case, depending on the application.

Let p and m denote respectively the original and reduced number of variables. The original variables are denoted X. In the simplest case our measure of accuracy of reconstruction is the sum of p squared multiple correlations between X-variables and the predictions of X made from the factors. In the more general case we can weight each squared multiple correlation by the variance of the corresponding X-variable. Since we can set those variances ourselves by multiplying scores on each variable, by any constant we choose, this amounts to the ability to assign any weights we choose to the different variables.

We now have a problem which is well-defined in the mathematical sense: reduce p variables to a set of m linear functions of those variables which best summarize the original p in the sense just described. It turns out, however, that infinitely many linear functions provide equally good summaries. To narrow the problem to one unique solution, we introduce three conditions. First, the m derived linear functions must be mutually uncorrelated. Second, any set of m linear functions must include the functions for a smaller set. For instance, the best 4 linear functions must include the best 3, which include the best 2, which include the best one. Third, the squared weights defining each linear function must sum to 1. These three conditions provide, for most data sets, one unique solution. Typically there are p linear functions (called

principal components) declining in importance; by using all p one gets perfect reconstruction of the original X-scores, and by using the first m (where m ranges from 1 to p) one gets the best reconstruction possible for that value of m.

Define each component's *eigenvector* or *characteristic vector* or *latent vector* as the column of weights used to form it from the X-variables. If the original matrix R is a correlation matrix, define each component's *eigenvalue* or *characteristic value* or *latent value* as its sum of squared correlations with the X-variables. If R is a covariance matrix, define the eigenvalue as a weighted sum of squared correlations, with each correlation weighted by the variance of the corresponding X-variable. The sum of the eigenvalues always equals the sum of the diagonal entries in R.

Non-unique solutions arise only when two or more eigenvalues are exactly equal; it then turns out that the corresponding eigenvectors are not uniquely defined. This case rarely arises in practice.

Each component's eigenvalue is called the "amount of variance" the component explains. The major reason for this is the eigenvalue's definition as a weighted sum of squared correlations. However, it also turns out that the actual variance of the component scores equals the eigenvalue. Thus in PCA the "factor variance" and "amount of variance the factor explains" are always equal. Therefore the two phrases are often used interchangeably, even though conceptually they stand for very different quantities.

17.2.2 The Number of Principal Components

While there are p original variables the number of principal components is m such that $m < p$. It may happen that m principal components will explain all the variance in a set of X-variables--that is, allow perfect reconstruction of X--even though $m < p$. However, in the absence of this event, there is no significance test on the number of principal components. To see why, consider first a simpler problem: testing the null hypothesis that a correlation between two variables is 1.0. This hypothesis implies that all points in the population fall in a straight line. From that it follows that if the correlation is 1.0 in the population, it must also be 1.0 in every single sample from that population. Any deviation from 1.0, no matter how small, contradicts the null hypothesis. A similar argument applies to the hypothesis that a multiple correlation is 1.0. But the hypothesis that m components account for all the variance in p variables is essentially the hypothesis that when the variables are predicted from the components by multiple regression, the multiple correlations are all 1.0. Thus even the slightest failure to observe this in a sample contradicts the hypothesis concerning the population.

17.2.3 Eigenvalue-based Rules for Selecting the Number of Components

Henry Kaiser suggested a rule for selecting a number of components m less than the number needed for perfect reconstruction: set m equal to the number of eigenvalues greater than 1. This rule is often used in common factor analysis as well as in PCA. Several lines of thought lead to Kaiser's rule, but the simplest is that since an eigenvalue is the amount of variance explained by one more component, it doesn't make sense to add a component that explains less variance than is contained in one variable. Since a component analysis is supposed to summarize a set of data, to use a component that explains less than a variance of 1 is something like writing a summary of a book in which one section of the summary is longer than the book section it summarizes--which makes no sense. However, Kaiser's major justification for the rule was that it matched pretty well the ultimate rule of doing several component analyses with different numbers of components, and seeing which analysis made sense. That ultimate rule is much easier today than it was a generation ago, so Kaiser's rule seems obsolete.

Raymond B. Cattell suggested an alternative method called the scree test. In this method the successive eigenvalues are plotted, and look for a spot in the plot where

the plot abruptly levels out. Cattell named this test after the tapering "scree" or rockpile at the bottom of a landslide. One difficulty with the scree test is that it can lead to very different conclusions if you plot the square roots or the logarithms of the eigenvalues instead of the eigenvalues themselves, and it is not clear why the eigenvalues themselves are a better measure than these other values.

Another approach is very similar to the *scree test*, but relies more on calculation and less on graphs. For each eigenvalue L, define S as the sum of all later eigenvalues plus L itself. Then L/S is the proportion of previously unexplained variance explained by L. For instance, suppose that in a problem with 7 variables the last 4 eigenvalues were 0.8, 0.2, 0.15, and 0.1. These sum to 1.25, so 1.25 is the amount of variance unexplained by a 3-component model. But 0.8/1.25 = 0.64, so adding one more component to the 3-component model would explain 64% of previously unexplained variance. A similar calculation for the fifth eigenvalue yields 0.2/(0.2+0.15+0.1) = 0.44, so the fifth principal component explains only 44% of previously unexplained variance.

17.2.4 PCA by Covariance Method

Following is a detailed description of PCA using the covariance method. The goal is to transform a given data set X of dimension M to an alternative data set Y of smaller dimension L.

Derivation of PCA using the covariance method

Let X be a d-dimensional random vector expressed as column vector. Without loss of generality, assume X has zero empirical mean. We want to find a $d \times d$ orthonormal projection matrix P such that

$$Y = P^T X \qquad \ldots(17.1)$$

where P^T is the traspose of matrx P,

with the constraint that

$$\text{cov}(Y) \text{ is a diagonal matrix and } P^{-1} = P^T \qquad \ldots(17.2)$$

By substitution, and matrix algebra, we obtain:

$$\begin{aligned} \text{cov}(Y) &= E\left[YY^T\right] \\ &= E\left[\left(P^T X\right)\left(P^T X\right)^T\right] \\ &= E\left[\left(P^T X\right)\left(X^T P\right)\right] \\ &= P^T E\left[XX^T\right]P \\ &= P^T \text{cov}(X)P \end{aligned} \qquad \ldots(17.3)$$

We now have:

$$\begin{aligned} P\,\text{cov}(Y) &= PP^T \text{cov}(X)PE \\ &= \text{cov}(X)P \end{aligned} \qquad \ldots(17.4)$$

Rewrite **P** as d $d \times 1$ column vectors, so

$$P = \left[P_1, P_2, \ldots, P_d\right] \qquad .(17.5)$$

and $\text{cov}(Y)$as

$$\begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_d \end{bmatrix} \qquad \ldots(17.6)$$

Substituting into equation above, we obtain:

$$[\lambda_1 P_1, \lambda_2 P_2, ..., \lambda_d P_d] = [\text{cov}(X)P_1, \text{cov}(X)P_2, ..., \text{cov}(X)P_d] \qquad \ldots(17.7)$$

Notice that in $\lambda_i P_i = \text{cov}(X)P_i$, P_i is an eigenvector of X's covariance matrix. Therefore, by finding the eigenvectors of X's covariance matrix, we find a projection matrix P that satisfies the original constraints. The following steps are involved in computing the principal components.

Organize the data set

Suppose you have data comprising a set of observations of M variables, and you want to reduce the data so that each observation can be described with only L variables, $L < M$. Suppose further, that the data are arranged as a set of N data vectors $X_1 ... X_N$ with each X_n representing a single grouped observation of the M variables.

- Write $X_1 ... X_N$ as column vectors, each of which has M rows.
- Place the column vectors into a single matrix **X** of dimensions $M \times N$.

Calculate the empirical mean

- Find the empirical mean along each dimension $m = 1 ... M$.
- Place the calculated mean values into an empirical mean vector vector **u** of dimensions $M \times 1$.

$$X_1 ... X_N$$

$$u[m] = \frac{1}{N} \sum_{n=1}^{N} X[m, n] \qquad \ldots(17.8)$$

Calculate the deviations from the mean

Subtract the empirical mean vector **u** from each column of the data matrix **X** Sto mean-subtracted data in the $M \times N$ matrix **B**

$$\mathbf{B} = \mathbf{X} - \mathbf{u} \cdot \mathbf{h} \qquad \ldots(17.9)$$

where **h** is a 1 x N row vector of all 1's

$$h[n] = 1 \qquad \text{for } n = 1 \ldots N \qquad \ldots(17.10)$$

Find the covariance matrix

- Find the $M \times M$ empirical covariance matrix **C** from the outer product of matrix **B** with itself:

$$\mathbf{C} = \text{E}[\mathbf{B} \otimes \mathbf{B}] = E[\mathbf{B} \cdot \mathbf{B}^*] = \frac{1}{N-1} \mathbf{B} \cdot \mathbf{B}^* \qquad \ldots(17.11)$$

where

E is the expected value operator,

$\otimes$ is the outer product operator, and

* is the conjugate transpose operator

Find the eigenvectors and eigenvalues of the covariance matrix

- Compute the eigenvalue matrix **D** and the eigenvector matrix **V** of the covariance matrix **C**:

$$\mathbf{C} \cdot \mathbf{V} = \mathbf{V} \cdot \mathbf{D} \qquad \ldots(17.12)$$

- Matrix **D** will take the form of an $M \times M$ diagonal matrix, where

$$D[p,q] = \lambda_m \qquad \text{for } p = q = m \qquad \ldots(17.13)$$

is the mth eigenvalue of the covariance matrix **C**, and

$$D[p,q] = 0 \qquad \text{for } p \neq q = m \qquad \ldots(17.14)$$

- Matrix **V**, also of dimension $M \times M$, contains M column vectors, each of length M, which represent the M eigenvectors of the covariance matrix **C**.

- The eigenvalues and eigenvectors are ordered and paired. The mth eigenvalue corresponds to the mth eigenvector.

Rearrange the eigenvectors and eigenvalues

- Sort the columns of the eigenvector matrix **V** and eigenvalue matrix **D** in order of *decreasing* eigenvalue.

- Make sure to maintain the correct pairings b etween the columns in each matrix.

Compute the cumulative energy content for each eigenvector

- The eigenvalues represent the distribution of the source data's energy among each of the eigenvectors, where the eigenvectors form a basis for the data. The cumulative energy content g for the mth eigenvector is the sum of the energy content across all of the eigenvectors from 1 through m:

$$g[m] = \sum_{q=1}^{m} D[p,q] \qquad \text{for} \quad p = q \quad \text{and } m = 1 \ldots M \qquad \ldots(17.15)$$

Select a subset of the eigenvectors as basis vectors

- Save the first L columns of **V** as the $M \times L$ matrix **W**:

$$W[p,q] = V[p,q] \qquad \text{for } p = 1 \ldots M \quad q = 1 \ldots L \qquad \ldots(17.16)$$

where

$$1 \leq L \leq M$$

- Use the vector **g** as a guide in choosing an appropriate value for L. The goal is to choose as small a value of L as possible while achieving a reasonably high value of g on a percentage basis. For example, you may want to choose L so that the cumulative energy g is above a certain threshold, like 90 percent. In this case, choose the smallest value of L such that

$$g[m = L] \geq 90\% \qquad \ldots(17.17)$$

Convert the source data to z-scores

- Create an $M \times 1$ empirical standard deviation vector **s** from the square root of each element along the main diagonal of the covariance matrix **C**:

$$s = \{s[m]\} = \sqrt{C[p,q]} \quad \text{for} \quad p = q \quad \text{and } m = 1...M \qquad ...(17.18)$$

score matrix:

$$Z = \frac{B}{\mathbf{s} \cdot h} \text{ (divide element-by-element)} \qquad ...(17.19)$$

Geometric Interpretation

If we consider a covariance matrix (computed from a bivariate data set, x_1 and x_2) as below:

$$C_X = \begin{pmatrix} 20.28 & 15.58 \\ 15.58 & 24.06 \end{pmatrix} \qquad ...(17.20)$$

we see that the covariance is symmetrical around the diagonal (the variances of x_1 and x_2 respectively). If we extract the eigenvectors and eigenvalues from this covariance matrix we have a new set of basis functions that are more efficient in representing the data from which the covariance matrix was derived.

$$U = \begin{pmatrix} 0.66 & -0.75 \\ 0.75 & 0.66 \end{pmatrix} \qquad ...(17.21)$$

Where U are the eigenvectors and

$$\lambda' = \begin{pmatrix} 37.87 \\ 6.47 \end{pmatrix} \qquad ...(17.22)$$

the eigenvalues. We can consider the columns or rows of the matrix in Eqn (17.20) as vectors and plot them from the origin out to their end points in Fig. 1. If we now plot the first eigenvector (column one in Eqn (17.10)) with a length of 37.87, remember the eigenvectors are orthonormal, and the second eigenvector (second column in Eqn (17.21)) with a length of 6.47, we have now plotted the semimajor and semiminor axes of an ellipse that en-circles both the eigenvalue-scaled-eigenvectors and the covariance vectors. This ellipse is oriented along the eigenvectors and has the magnitudes of the eigenvalues. We can now plot an alternate coordinate, by using the major and minor axes of this ellipse, in Fig.1.

If, in Figure 1, we were to project vector 1 (the first vector formed from the covariance matrix, whose original "coordinates" were 20.3, 15.6) back onto the major and minor axes of the ellipse (the first and second eigenvectors), we would get the "more efficient" representation coordinates of 25.01, 4.88. Most of the information is loaded onto the first principal component and this would be true of each individual sample as well. We call these more efficient coordinates the *principal component factors*.

We say, "more efficient", because these factors redistribute the total variance in a preferential way. The total variance is given by the sum of the diagonal of the covariance matrix (the sum of the diagonal of a matrix is called the *trace*), in this case 44.34. A very useful feature of eigen-analysis is this: the sum of the eigenvalues always equal the total variance. We can now evaluate how much of the total variance is included in the first component of the original data X. From the covariance matrix we get the variance of x_1: 20.28/44.34 or about 46%. Similarly for the second component x_2: 24.06/44.34 or about 54%. In the new coordinate system given by the eigenvectors the amount of variance contained in the principal components (eigenvectors) are given by the eigenvalues, or in percentages, 37.87/44.34 or about 85% and 6.47/44.34 or about 15%. This is what we mean by more *efficient*, the first factor scores account for 85% of the total variance; if one had to compress their data down to one vector, the principal component scores offer an obvious choice.

So, in principal component analysis we have a more efficient coordinate system to describe our data. To put it another way, *if* we have to reduce the number of numbers describing our data to just one number converting to the principal component scores first minimizes the information lost.

There are other traits of PCA. The total variance of the data set is equal to the *trace* of the covariance matrix, which also equal to the sum of the eigenvalues (the trace of the diagonal matrix containing the eigenvalues). Using this we can apportion how much of the original, total variance is accounted for by the individual principal component (eigenvectors). If the matrix U above looked like:

$$U = \begin{pmatrix} 0.66 & -0.75 \\ 0.75 & 0.66 \end{pmatrix} \qquad \ldots(17.23)$$

Each column would represent the eigenvectors (unit length) and the amount of variance accounted for by the first principal component would be given by:

$$\frac{\lambda_1}{\text{var}(X)} \qquad \ldots(17.24)$$

where var(X) represents the total variance of the data set X and λ is the diagonal matrix containing the eigenvalues. Another way of writing this is:

$$\sum_{i=1} \lambda_i$$

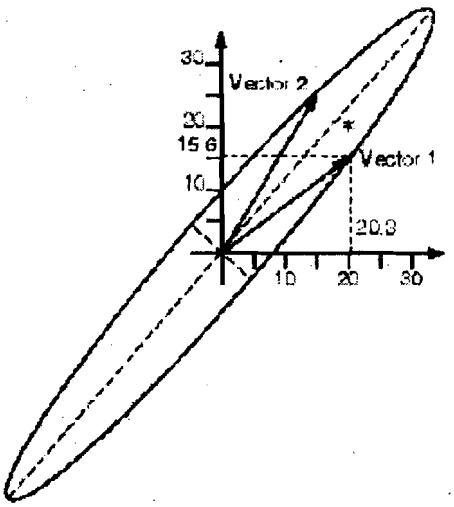

Fig. 17.1: Covariance Vectors Plotted as Vectors 1 and 2.

In Fig. 17.1, the ellipse major axis is the first eigenvector (of unit length) multiplied by the corresponding eigenvalue. The minor axis represents the second eigenvector/value. Vectors 1 and 2 represent a basis for this data set, but they are not totally independent and are not necessarily efficient. Because the eigenvectors are always orthonormal they are always independent and more efficient in their representation of the original data.

17.2.5 Difference in Goals between PCA and FA

In PCA the eigenvalues must ultimately account for all of the variance. There is no probability, no hypothesis, no test because strictly speaking PCA is not a statistical procedure. The PCA is merely a mathematical manipulation to recast *m* variables as *m* factors. Factor analysis (FA), however, brings *a priori* knowledge to the problem solving exercise.

There is a short list of primary assumptions behind factor analysis. But basically, factor analysis assumes that there are correlations/covariances between the m variables in the data set that are a result of p underlying, mutually uncorrelated factors.

Check Your Progress 1

1) Consider a random vector $\boldsymbol{Z}$ with mean and covariance matrix given below.

$$\mu_{\boldsymbol{Z}} = \begin{pmatrix} 2.1 \\ 0.4 \\ 1.6 \end{pmatrix} \text{ and } \Sigma_{\boldsymbol{Z}} = \begin{pmatrix} 13.82 & 10.73 & 12.21 \\ 10.73 & 16.82 & 1.74 \\ 12.21 & 1.74 & 18.18 \end{pmatrix}$$

a) Calculate the determinant of the corresponding correlation matrix.

b) Is $\boldsymbol{Z}$ singular, multicollinear, or neither of these?

c) Calculate the eigenvalues and eigenvectors of $\sum Z$.

d) Represent $\boldsymbol{Z}$ in terms of its principal components as in.

e) What is the covariance matrix $\sum D$ of the vector of principal components $\boldsymbol{D}$?

f) Construct an approximation $\tilde{Z}$ for $\boldsymbol{Z}$ based on the first two principal components of $\boldsymbol{Z}$.

g) Construct the covariance matrix of $\tilde{Z}$. Compare your result with the covariance matrix of $\boldsymbol{Z}$.

.....

.....

.....

.....

.....

17.3 LET US SUM UP

The principal component analysis is amongst the oldest of the multivariate statistical methods of data reduction. There are the methods for producing a small number of constructed variables, derived from the larger number of variables originally collected. The idea is to produce a small number of derived variables that are uncorrelated and that account for most of the variation in the original data set. The main reason that we might want to reduce the number of variables in this way is that it helps us to understand the underlying structure of the data.

Principal component analysis does not have an underlying statistical model. It is just a mathematical technique and, as such, is used in other statistical analyses that are driven by models, for example, factor analysis. The emphasis in factor analysis is to identify underlying factors that might explain the variability in a large and complex data set. Factor analysis is a two-stage process and PCA is the most commonly used method for the first stage, the extraction of an initial solution. Thus, the mathematical technique of PCA underlies other multivariate statistical methods.

17.4 KEY WORDS

Diagonal matrix : It is a square matrix in which the entries outside the main diagonal are all zero.

Dimensionality reduction : It can be divided into two categories: feature selection and feature extraction. Dimensionality reduction is also a phenomenon discussed widely in physics, whereby a physical system exists in three dimensions, but its properties behave like those of a lower-dimensional system. It has been experimentally realised at the quantum critical point in an insulating magnet called 'Han purple'.

Eigenvalue (or characteristic root) : It is a mathematical property of a matrix; used in relation to the decomposition of a covariance matrix, both as a criterion of determining the number of factors to extract and a measure of variance accounted for by a given dimension.

Eigenvector : It is a vector associated with its respective eigenvalue; obtained in the process of initial factoring; when these vectors are appropriately standardized, they become factor loadings.

Expected value (or mathematical expectation) : For a random variable it is the sum of the probability of each possible outcome of the experiment multiplied by its payoff ("value"). Thus, it represents the average amount one "expects" as the outcome of the random trial when identical odds are repeated many times. Note that the value itself may not be expected in the general sense; it may be unlikely or even impossible.

Linear transformation (also called linear map or linear operator) : is a function between two vector spaces that preserves the operations of vector addition and scalar multiplication. In the language of abstract algebra, a linear transformation is a homomorphism of vector spaces, or a morphism in the category of vector spaces over a given field.

Linear transformation (also called linear map or linear operator) Outer product : It typically refers to the tensor product or to operations with similar cardinality such as exterior product. The cardinality of these operations is that of cartesian products.

Principal Components : These are linear combinations of observed variables, possessing properties such as being orthogonal to each other, and the first principal component representing the largest amount of variance in the data, the second representing the second largest and so on; often considered variants of common factors, but more accurately they are contrasted with common factors which are hypothetical.

Scree Test : It is a rule of thumb criterion for determining the number of significant factors to retain; it is based on the graph of roots (eigenvalues); claimed to be appropriate in handling disturbances due to minor (unarticulated) factors.

Varimax : It is a method of orthogonal rotation which simplifies the factor structure by maximizing the variance of a column of the pattern matrix.

17.5 SOME USEFUL BOOKS/REFERENCES

Darlington, Richard B., Principal Component Analysis, at the website: http://comp9.psych.cornell. edu/Darlington/factor.htm

Dunteman, George H., 1989, *Principal Components Analysis*. Thousand Oaks, CA: Sage Publications, Quantitative Applications in the Social Sciences Series, No. 69.

Garson, G. David, 2006, Principal Component Analysis, Wikipedia, the free encyclopedia.

Glover, Jenkins and Doney, 2004, "Principal Component and Factor Analysis", Chapter 4 in *Modeling, Data Analysis and Numerical Techniques for Geochemistry*.

Website "http://en.wikipedia.org/wiki/Principal_components_analysis".

17.6 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1)

a) The correlation matrix has determinant $|\rho| = .009314$.

b) Based upon item (a), $\boldsymbol{Z}$ is multicollinear.

c) The eigenvalues λ_1, λ_2 and λ_3 are 33.0548, 15.6856 and 0.0759 respectively. The corresponding eigenvectors form the columns of

$$\boldsymbol{v} = \begin{pmatrix} .6453 & -.0270 & .7635 \\ .4892 & -.7529 & -.4401 \\ .5868 & .6575 & -.4726 \end{pmatrix}$$

d) The principal components of $\boldsymbol{Z}$ are the normalized eigenvectors of the covariance matrix $\sum Z$. The eigenvectors we presented for item (c) are normalized, so they are the principal components of $\boldsymbol{Z}$.

e)

$$Z = D_1 v_1 + D_2 v_2 + D_3 v_3 + \mu_Z$$

$$= vD + \mu_Z$$

$$= \begin{pmatrix} .6453 & -.0270 & .7635 \\ .4892 & -.7529 & -.4401 \\ .5868 & .6575 & -.4726 \end{pmatrix} \begin{pmatrix} D_1 \\ D_2 \\ D_3 \end{pmatrix} + \begin{pmatrix} 2.1 \\ 0.4 \\ 1.6 \end{pmatrix}$$

f)

$$\Sigma_D = \begin{pmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{pmatrix} = \begin{pmatrix} 33.0584 & 0 & 0 \\ 0 & 15.6856 & 0 \\ 0 & 0 & 0.0759 \end{pmatrix}$$

g)

$$Z = D_1 v_1 + D_2 v_2 + \mu_Z$$

$$= \begin{pmatrix} .6453 & -.0270 \\ .4892 & -.7529 \\ .5868 & .6575 \end{pmatrix} \begin{pmatrix} D_1 \\ D_2 \end{pmatrix} + \begin{pmatrix} 2.1 \\ 0.4 \\ 1.6 \end{pmatrix}$$

UNIT 18 FACTOR ANALYSIS

Structure

18.0 OBJECTIVES

After going through this Unit you will be able to:

- explain the underlying patterns in a the interrelationship between variables;
- estimate factors or latent variables from a data set; and
- reduce the dimensionality of a large number of variables to a fewer number of factors.

18.1 INTRODUCTION

Factor analysis (FA) explains variability among observed random variables in terms of fewer unobserved random variables called factors. The observed variables are expressed in terms of linear combinations of the factors, plus "error" terms. Factor analysis originated in psychometrics, and is used in social sciences, marketing, product management, operations research, and other applied sciences that deal with large quantities of data.

Factor analysis is applied to a set of variables to discover coherent subsets that are relatively independent of one another. Variables, correlated with each other and independent of other subsets of variables are combined into factors. Factors, which are generated, are thought to be representative of the underlying processes that have created the correlations among variables.

FA can be exploratory in nature; FA is used as a tool in attempts to reduce a large set of variables to a more meaningful, smaller set of variables. As FA is sensitive to the magnitude of correlations robust comparisons must be made to ensure the quality of the analysis.

Accordingly, FA is sensitive to outliers, missing data, and poor correlations between variables due to poorly distributed variables. As a result data transformations have a large impact upon FA.

18.2 USES AND CLASSIFICATIONS

Factor analysis is integrated in structural equation modeling (SEM), helping create the latent variables modeled by SEM. However, factor analysis can be and is often used on a stand-alone basis for similar purposes.

- To select a subset of variables from a larger set based on which original variables have the highest correlations with the principal component factors.
- To create a set of factors to be treated as uncorrelated variables as one approach to handling multicollinearity in such procedures as multiple regression
- To validate a scale or index by demonstrating that its constituent items load on the same factor, and to drop proposed scale items which cross-load on more than one factor.
- To establish that multiple tests measure the same factor, thereby giving justification for administering fewer tests.
- To identify clusters of cases and/or outliers.
- To determine network groups by determining which sets of people cluster together (using Q-mode factor analysis)

Let us consider this example. Suppose, a mother finds various bumps and shapes under a blanket at the bottom of a bed. With one shape moving toward the top of the bed, all the other bumps and shapes also move toward the top, making the mother infer that what is under the blanket is a single thing, most likely her child. Factor analysis takes as input a number of measures and tests, analogous to the bumps and shapes. Those that move together are considered to be a single factor. That is, in factor analysis the researcher is assuming that there is a "child" out there in the form of an underlying factor, and he or she takes simultaneous movement (correlation) as evidence of its existence. If correlation is spurious for some reason, this inference will be mistaken, of course, so it is important when conducting factor analysis that possible variables, which might introduce spuriousness, such as anteceding causes, be included in the analysis and taken into account.[1]

Factor analysis makes many of the same assumptions as multiple regression: linear relationships, interval or near-interval data, untruncated variables, proper specification (relevant variables included, extraneous ones excluded), lack of high multicollinearity, and multivariate normality for purposes of significance testing. Eventually, it generates a table in which the rows are the observed raw indicator variables and the columns are the factors or latent variables, which explain as much of the variance in these variables as possible. The cells in this table are factor loadings, and the meaning of the factors must be induced from seeing which variables are most heavily loaded on which factors. This inferential labeling process can be fraught with subjectivity as diverse researchers impute different labels. [2]

Of the various types of factor analyses, principal components analysis (PCA) is the most common. However, principal axis factoring (PAF), also called common factor analysis, is preferred for purposes of confirmatory factory analysis in structural equation modeling.

18.2.1 Understanding the Causes

Many statistical methods are used to study the relation between independent and dependent variables. Factor analysis is different; it is used to study the patterns of relationship among many dependent variables, with the goal of discovering

[1] http://www2.chass.ncsu.edu/garson/pa765/factor.htm
[2] http://www2.chass.ncsu.edu/garson/pa765/factor.htm

something about the nature of the independent variables that affect them, even though those independent variables were not measured directly. Thus answers obtained by factor analysis are necessarily more hypothetical and tentative than is true when independent variables are observed directly. The inferred independent variables are called *factors*. A typical factor analysis suggests answers to four major questions:

1) How many different factors are needed to explain the pattern of relationships among these variables?
2) What is the nature of those factors?
3) How well do the hypothesized factors explain the observed data?
4) How much purely random or unique variance does each observed variable include?

18.2.2 Assumptions

There is a simple list of fundamental assumptions that underlie factor analysis and distinguish it from principal component analysis (even if they share a lot of common mathematical machinery).

1) The correlations and covariances that exist between m variables are a result of p underlying, mutually uncorrelated factors. Usually p is less than m.
2) Usually p is known in advance. The number of factors, hidden in the data set, is one of the pieces of *a priori* knowledge that is brought to the table to solve the factor analysis problem.
3) The rank of a matrix and the number of eigenvectors are interrelated, the eigenvalues are the square of the non-zero singular values. The eigenvalues are ordered by the amount of variance accounted for.

Factor analysis starts with the basic principal component approach, but differs in two important ways. First of all, factor analysis is *always* done with *standardized* data. This implies that we want the individual variables to have equal weight in their influence on the underlying variance-covariance structure. In addition, this requirement is necessary for us to be able to convert the principal component vectors into factors. Secondly, the eigenvectors must be computed in such a way that they are *normalized*, *i.e.* of unit length or orthonormal.

18.2.3 Factor Loadings Matrix

As we stated above, we start factor analysis with principal component analysis, but we quickly diverge as we apply the *a priori knowledge* we brought to the problem. This knowledge may be of the form that we "know" how many factors there should be or it may be more of the nature that allows our experience and intuition about the data guide us as to how many factors there should be. As before, with PCA, we can take the eigenvectors (of unit length) and weight them with the square root of the corresponding eigenvalue:

$$A^R = UI\sqrt{\lambda'} \qquad \ldots(18.1)$$

Here A^R represents a matrix such as:

	I	II	III	...	p
$\mathbf{X_1}$	L_{11}	L_{12}	L_{13}	...	L_{1p}
$\mathbf{X_2}$	L_{21}	L_{22}	L_{23}	...	
$\mathbf{X_3}$	L_{31}				
$\mathbf{X_m}$					

where the m variables X run down the side and the *p factors* go across the top and the L represent the *Loadings* of each variable on individual factors. When $p=m$ you have the same thing as PCA.

18.2.4 Communalities

The communalities h_j^2 represent the fraction of the total variance accounted for of variable j. By calculating the communalities we can keep track of how much of the original variance that was contained in variable j is still being accounted for by the number of factors we have retained. As a consequence, when $p = m, h_j^2 = 1$, always, so long as the data was standardized first. The communalities are calculated in the following fashion:

$$h_j^2 = \sum_{r=1}^{p} (A_{jr}^R)^2 \qquad \ldots(18.2)$$

One can read/think of this as: summing the squares of the factor loadings horizontally across the factor loadings matrix.

18.2.5 Number of Factors

Above we make several references to the fact that if $p=m$ *(i.e.* the number of factors equal the number of variables) then factor analysis is no different than PCA with standardized variables. But of course, in factor analysis, you want $p<<m$ and so the question remains: how do you decide which factors to keep? When doing factor analysis it helps to keep the results in the following organization:

	I	II	III	...	p	h_j^2
$\mathbf{X_1}$	A_{11}^R	A_{12}^R	A_{13}^R		A_{1p}^R	h_1^2
$\mathbf{X_2}$	A_{21}^R	A_{22}^R	A_{23}^R			h_2^2
$\mathbf{X_3}$	A_{31}^R					h_3^2
$\mathbf{X_m}$						h_m^2

Factor	Eigenvalue	% Total Var	Cummulative % Var
I	λ_1	.....	.....
II	λ_2	.....	.....
III	λ_3	.....	.....
p	λ_p	.....	100%

At first the A_{jr}^R could be the entries from U (the raw eigenvectors), but later you can use the entries from A^R (the factor loadings) as the analysis continues and the number of factors decreases. In this manner, you can keep track of the number of factors you are dealing with, how much of the original total variance is being accounted for, where the variables are loading on the individual factors, and the communalities on each individual variable. This will help you see how your choices in the number of factors kept have effected these measures of performance.

As to the question of how you decide, unfortunately there is no hard and fast rule as to how many factors to keep. One *rule of thumb* is to keep all of the factors whose eigenvalue is greater than one, provided you started with standardized data. If you get a lot of factors with eigenvalues greater than one, then you might have to face the likelihood that maybe the factor theory approach isn't applicable to your problem, at least in the way you have presented it. Typically the more successful factor analyses have been those where a "few" factors account for most of the variance. See the discussion of simple structure concepts in the next section.

18.2.6 Varimax Rotation and Simple Structure Concepts

After you have chosen the few factors you wish to keep in your analysis, you can "improve" the fit of this reduced dimensionality coordinate system to your data by a technique known as *factor rotation*. Even though the number of factors may have reduced the dimensionality of your problem, the factors may not be easy to interpret. Factor rotation allows you to reorganize the *loadings* onto *rotated* factors. This is accomplished by maximizing the variance of the loadings on the factors. For each (K^{th}) factor we can compute:

$$s_k^2 = \frac{p\sum_j^m \left(\frac{a_{jp}^2}{h_j^2}\right)^2 - \left(\sum_j^m \frac{a_{jp}^2}{h_j^2}\right)^2}{p^2}$$

where p is the number of retained factors, m is the number of original variables, a_{jp} is the loading of variable j on factor p, and h_j^2 is the communality of the j^{th} variable. Using this expression of the variance of the loading on the k^{th} factor, one maximizes the following:

$$V = \sum_k^p s_k^2$$

This is an iterative process where you rotate two factors at a time, holding the others constant, until the increase in the overall variance V drops below a preset value. This is the heart of the Kaiser Varimax orthogonal rotation.

The various factor rotation methods have, as a guiding principle, the *simple structure concepts*. That is to say, the results, after rotation, should have become simple in their appearance. To put it another way, these simple structure concepts should be considered when trying to determine whether or not a given factor rotation has clarified the underlying structure of the data. Five simple structure precepts have been put forth by Thurstone:

1) There should be at least one zero in each row of the factor loadings matrix.

2) There should be at least k zeros in each column of the factor matrix, where k is the number of factors extracted.

3) For every pair of factors, some variables (in the R-mode) should have high loadings on one and near-zero loadings on the other.

4) For every pair of factors, several variables should have small loadings on both factors.

5) For every pair of factors, only a few variables should have non-vanishing loadings on both.

But in real world to get fulfilled all these five rules is very rare with orthogonal factor rotation.

Check Your Progress 1

1) What is meant by a factor in the context of factor analysis?

.....

.....

.....

.....

2) Sketch the three basic matrices involved in the factor analysis procedure: Input data matrix, Correlation matrix, and Factor matrix.

.....

.....

.....

.....

3) Discuss the meaning of a factor loading. What is its maximum and minimum value? Explain.

.....

.....

.....

.....

4) What is accomplished by rotating a factor-loading matrix?

.....

.....

.....

.....

18.3 LET US SUM UP

The factor analysis (FA) and principal component analysis (PCA) are amongst the oldest of the multivariate statistical methods of data reduction. These are the methods for producing a small number of constructed variables, derived from the larger number of variables originally collected. The idea is to produce a small number of derived variables that are uncorrelated and that account for most of the variation in the original data set. The main reason that we might want to reduce the number of variables in this way is that it helps us understand the underlying structure of the data.

Principal component analysis does not have an underlying statistical model. It is just a mathematical technique and, as such, is used in other statistical analyses that are driven by models, for example, factor analysis. The emphasis in factor analysis is on identification of underlying factors that might explain the variability in a large and complex data set. Factor analysis is a two-stage process while PCA is the most

commonly used method for the first stage, the extraction of an initial solution. Thus, the mathematical technique of PCA underlies other multivariate statistical methods.

18.4 KEY WORDS

Communality : It shows the variance of an observed variable accounted for by the common factors; in an orthogonal factor model. It is equivalent to the sum of the squared factor loadings.

Common Factor : It implies the unmeasured (or hypothetical) underlying variable which is the source of variation in at least two observed variables under consideration.

Eigenvalue (or characteristic root) : It is a mathematical property of a matrix; used in relation to the decomposition of a covariance matrix, both as a criterion of determining the number of factors to extract and a measure of variance accounted for by a given dimension.

Eigenvector : It is a vector associated with its respective eigenvalue; obtained in the process of initial factoring; when these vectors are appropriately standardized, they become factor loadings.

Factor Extraction : It is the initial stage of factor analysis in which the covariance matrix is resolved into a smaller number of underlying factors or components.

Factors : It implies hypothesized, unmeasured, and underlying variables which are presumed to be the sources of the observed variables; often divided into unique and common factors.

Factor Loading : It is a general term referring to a coefficient in a factor pattern or structure matrix.

Factor Pattern Matrix : It refers to a matrix of coefficients where the columns usually refer to common factors and the rows to the observed variables; elements of the matrix represent regression weights for the common factors where an observed variable is assumed to be a linear combination of the factors; for an orthogonal solution, the pattern matrix is equivalent to correlations between factors and variables.

Orthogonal Factors : It indicates the factors that are not correlated with each other; factors obtained through orthogonal rotation.

Orthogonal Rotation : It refers to the operation through which a simple structure is sought under the restriction that factors be orthogonal (or uncorrelated); factors obtained through this rotation are by definition uncorrelated.

Principal Axis Factoring : It is a method of initial factoring in which the adjusted correlation matrix is decomposed hierarchically; a principal axis factor analysis

with iterated commonalities leads to a least-squares solutions of initial factoring.

Principal Components : It reflects linear combinations of observed variables, possessing properties such as being orthogonal to each other, and the first principal component representing the largest amount of variance in the data, the second representing the second largest and so on; often considered variants of common factors, but more accurately they are contrasted with common factors which are hypothetical.

Scree Test : It is a rule of thumb criterion for determining the number of significant factors to retain; it is based on the graph of roots (eigenvalues); claimed to be appropriate in handling disturbances due to minor (unarticulated) factors.

Varimax : It refers to a method of orthogonal rotation which simplifies the factor structure by maximizing the variance of a column of the pattern matrix.

18.5 SOME USEFUL BOOKS/REFERENCES

Kachigan, Sam K., 1991, *Multivariate Statistical Analysis: A Conceptual Introduction*

Kin, Jae-On and Charles W. Mueller, 1978, *Factor Analysis: Statistical Methods and Practical Issues*, Sage Publications

18.6 ANSWERS/HINTS TO CHECK YOUR PROGRESS EXERCISES

Check Your Progress 1

1) See Section 18.2 and answer.

2) See Section 18.2 and answer.

3) See Sub-section 18.2.3 and answer.

4) See Sub-section 18.2.6 and answer.

18.7 EXERCISE

The following Table provides data on the crime rates on 35 demographic groups. Using an appropriate statistical package (SPSS/ Stata, etc.) run a factor analysis and interpret the factors.

HRS = Average hours worked during the year
RATE = Average hourly wage (dollars)
ERSP = Average yearly earnings of a spouse (dollars)
ERNO = Average yearly earnings of other family members (dollars)
NEIN = Average yearly non-earned income (dollars)

ASSET = Average family asset holdings (bank account, etc.) (dollars)
AGE = Average age of respondent
DEP = Average number of dependents
SCHOOL = Average highest grade of school completed.

HRS	RATE	ERSP	ERNO	NEIN	ASSET	AGE	DEP	SCHOOL
2157	2.905	1121	291	380	7250	38.5	2.34	10.5
2174	2.97	1128	301	398	7744	39.3	2.335	10.5
2062	2.35	1214	326	185	3068	40.1	2.851	8.9
2111	2.511	1203	49	117	1632	22.4	1.159	11.5
2134	2.791	1013	594	730	12710	57.7	1.229	8.8
2185	3.04	1135	287	382	7706	38.6	2.602	10.7
2210	3.222	1100	295	474	9338	39	2.187	11.2
2105	2.493	1180	310	255	4730	39.9	2.616	9.3
2267	2.838	1298	252	431	8317	38.9	2.024	11.1
2205	2.356	885	264	373	6789	38.8	2.662	9.5
2121	2.922	1251	328	312	5907	39.8	2.287	10.3
2109	2.499	1207	347	271	5069	39.7	3.193	8.9
2108	2.796	1036	300	259	4614	38.2	2.04	9.2
2047	2.453	1213	297	139	1987	40.3	2.545	9.1
2174	3.582	1141	414	498	10239	40	2.064	11.7
2067	2.909	1805	290	239	4439	39.1	2.301	10.5
2159	2.511	1075	289	308	5621	39.3	2.486	9.5
2257	2.516	1093	176	392	7293	37.9	2.042	10.1
1985	1.423	553	381	146	1866	40.6	3.833	6.6
2184	3.636	1091	291	560	11240	39.1	2.328	11.6
2084	2.983	1327	331	296	5653	39.8	2.208	10.2
2051	2.573	1194	279	172	2806	40	2.362	9.1
2127	3.262	1226	314	408	8042	39.5	2.259	10.8
2102	3.234	1188	414	352	7557	39.8	2.019	10.7
2098	2.28	973	364	272	4400	40.6	2.661	8.4
2042	2.304	1085	328	140	1739	41.8	2.444	8.2
2181	2.912	1072	304	383	7340	39	2.337	10.2
2186	3.015	1122	30	352	7292	37.2	2.046	10.9
2188	3.01	990	366	374	7325	38.4	2.847	10.6
2077	1.901	350	209	95	1370	37.4	4.158	8.2
2196	3.009	947	294	342	6888	37.5	3.047	10.6
2093	1.899	342	311	120	1425	37.5	4.512	8.1
2173	2.959	1116	296	387	7625	39.2	2.342	10.5
2179	2.971	1128	312	397	7779	39.4	2.341	10.5
2200	2.98	1126	204	393	7885	39.2	2.341	10.6

Source: Greenberg, D.H. and M. Kosters, 1970, *Income Guarantees and Working Poor,* The Rand Corporation, R-579-OEO, December.

Hints:

How to Run Factor Analysis in SPSS?

After opening the SPSS (whatever be the version) window,

Step 1 : Go to "Analyaze" in the menu given on the top
Step 2 : Click on "Data Reduction"
Step 3 . Click on "Factor"
Step 4 : Select all the variables, and place in the blank space
Step 5 : Click on "Extraction" and select "principal axis factoring" and press "continue"
Step 6 : Click on "OK".