# 13

# Multivariate Analysis Techniques

All statistical techniques which simultaneously analyse more than two variables on a sample of observations can be categorized as multivariate techniques. We may as well use the term 'multivariate analysis' which is a collection of methods for analyzing data in which a number of observations are available for each object. In the analysis of many problems, it is helpful to have a number of scores for each object. For instance, in the field of intelligence testing if we start with the theory that general intelligence is reflected in a variety of specific performance measures, then to study intelligence in the context of this theory one must administer many tests of mental skills, such as vocabulary, speed of recall, mental arithmetic, verbal analogies and so on. The score on each test is one variable, $X_i$, and there are several, $k$, of such scores for each object, represented as $X_1, X_2 \ldots X_k$. Most of the research studies involve more than two variables in which situation analysis is desired of the association between one (at times many) criterion variable and several independent variables, or we may be required to study the association between variables having no dependency relationships. All such analyses are termed as multivariate analyses or multivariate techniques. In brief, techniques that take account of the various relationships among variables are termed multivariate analyses or multivariate techniques.

## GROWTH OF MULTIVARIATE TECHNIQUES

Of late, multivariate techniques have emerged as a powerful tool to analyse data represented in terms of many variables. The main reason being that a series of univariate analysis carried out separately for each variable may, at times, lead to incorrect interpretation of the result. This is so because univariate analysis does not consider the correlation or inter-dependence among the variables. As a result, during the last fifty years, a number of statisticians have contributed to the development of several multivariate techniques. Today, these techniques are being applied in many fields such as economics, sociology, psychology, agriculture, anthropology, biology and medicine. These techniques are used in analyzing social, psychological, medical and economic data, specially when the variables concerning research studies of these fields are supposed to be correlated with each other and when rigorous probabilistic models cannot be appropriately used. Applications of multivariate techniques in practice have been accelerated in modern times because of the advent of high speed electronic computers.

## CHARACTERISTICS AND APPLICATIONS

Multivariate techniques are largely empirical and deal with the reality; they possess the ability to analyse complex data. Accordingly in most of the applied and behavioural researches, we generally resort to multivariate analysis techniques for realistic results. Besides being a tool for analyzing the data, multivariate techniques also help in various types of decision-making. For example, take the case of college entrance examination wherein a number of tests are administered to candidates, and the candidates scoring high total marks based on many subjects are admitted. This system, though apparently fair, may at times be biased in favour of some subjects with the larger standard deviations. Multivariate techniques may be appropriately used in such situations for developing norms as to who should be admitted in college. We may also cite an example from medical field. Many medical examinations such as blood pressure and cholesterol tests are administered to patients. Each of the results of such examinations has significance of its own, but it is also important to consider relationships between different test results or results of the same tests at different occasions in order to draw proper diagnostic conclusions and to determine an appropriate therapy. Multivariate techniques can assist us in such a situation. In view of all this, we can state that "if the researcher is interested in making probability statements on the basis of sampled multiple measurements, then the best strategy of data analysis is to use some suitable multivariate statistical technique."[1]

The basic objective underlying multivariate techniques is to represent a collection of massive data in a simplified way. In other words, multivariate techniques transform a mass of observations into a smaller number of composite scores in such a way that they may reflect as much information as possible contained in the raw data obtained concerning a research study. Thus, the main contribution of these techniques is in arranging a large amount of complex information involved in the real data into a simplified visible form. Mathematically, multivariate techniques consist in "forming a linear composite vector in a vector subspace, which can be represented in terms of projection of a vector onto certain specified subspaces."[2]

For better appreciation and understanding of multivariate techniques, one must be familiar with fundamental concepts of linear algebra, vector spaces, orthogonal and oblique projections and univariate analysis. Even then before applying multivariate techniques for meaningful results, one must consider the nature and structure of the data and the real aim of the analysis. We should also not forget that multivariate techniques do involve several complex mathematical computations and as such can be utilized largely with the availability of computer facility.

## CLASSIFICATION OF MULTIVARIATE TECHNIQUES

Today, there exist a great variety of multivariate techniques which can be conveniently classified into two broad categories viz., dependence methods and interdependence methods. This sort of classification depends upon the question: Are some of the involved variables dependent upon others? If the answer is 'yes', we have dependence methods; but in case the answer is 'no', we have interdependence methods. Two more questions are relevant for understanding the nature of multivariate techniques. Firstly, in case some variables are dependent, the question is how many variables are dependent? The other question is, whether the data are metric or non-metric? This means whether

[1]K. Takeuchi, H. Yanai and B.N. Mukherji, *The Foundations of Multivariate Analysis*, p. 54.

[2] Ibid., p. iii.

the data are quantitative, collected on interval or ratio scale, or whether the data are qualitative, collected on nominal or ordinal scale. The technique to be used for a given situation depends upon the answers to all these very questions. Jadish N. Sheth in his article on "The multivariate revolution in marketing research"[3] has given the flow chart that clearly exhibits the nature of some important multivariate techniques as shown in Fig. 13.1.

Thus, we have two types of multivariate techniques: one type for data containing both dependent and independent variables, and the other type for data containing several variables without dependency relationship. In the former category are included techniques like multiple regression analysis, multiple discriminant analysis, multivariate analysis of variance and canonical analysis, whereas in the latter category we put techniques like factor analysis, cluster analysis, multidimensional scaling or MDS (both metric and non-metric) and the latent structure analysis.

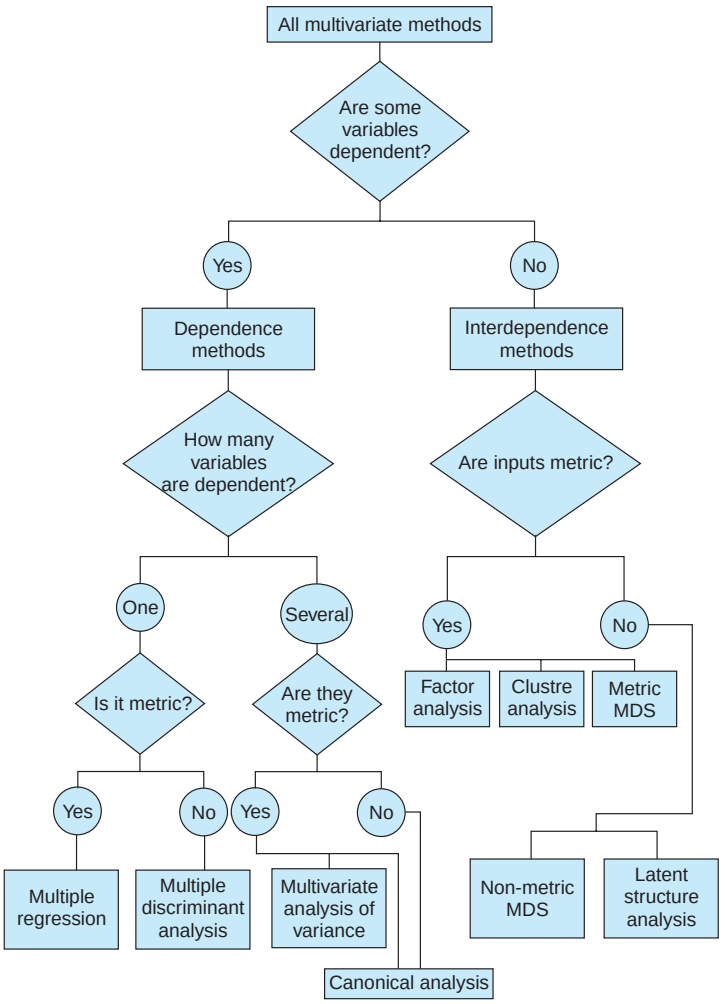


**Fig. 13.1**

[3]*Journal of Marketing, American Marketing Association,* Vol. 35, No. 1 (Jan. 1971), pp. 13–19.

## VARIABLES IN MULTIVARIATE ANALYSIS

Before we describe the various multivariate techniques, it seems appropriate to have a clear idea about the term, 'variables' used in the context of multivariate analysis. Many variables used in multivariate analysis can be classified into different categories from several points of view. Important ones are as under:

(i) *Explanatory variable and criterion variable:* If $X$ may be considered to be the cause of $Y$, then $X$ is described as explanatory variable (also termed as causal or independent variable) and $Y$ is described as criterion variable (also termed as resultant or dependent variable). In some cases both explanatory variable and criterion variable may consist of a set of many variables in which case set ($X_1$, $X_2$, $X_3$, …., $X_p$) may be called a set of explanatory variables and the set ($Y_1$, $Y_2$, $Y_3$, …., $Y_q$) may be called a set of criterion variables if the variation of the former may be supposed to cause the variation of the latter as a whole. In economics, the explanatory variables are called external or exogenous variables and the criterion variables are called endogenous variables. Some people use the term external criterion for explanatory variable and the term internal criterion for criterion variable.

(ii) *Observable variables and latent variables:* Explanatory variables described above are supposed to be observable directly in some situations, and if this is so, the same are termed as observable variables. However, there are some unobservable variables which may influence the criterion variables. We call such unobservable variables as latent variables.

(iii) *Discrete variable and continuous variable:* Discrete variable is that variable which when measured may take only the integer value whereas continuous variable is one which, when measured, can assume any real value (even in decimal points).

(iv) *Dummy variable* (*or Pseudo variable*)*:* This term is being used in a technical sense and is useful in algebraic manipulations in context of multivariate analysis. We call $X_i$ ( $i$ = 1, …., $m$) a dummy variable, if only one of $X_i$ is 1 and the others are all zero.

## IMPORTANT MULTIVARIATE TECHNIQUES

A brief description of the various multivariate techniques named above (with special emphasis on factor analysis) is as under:

(i) *Multiple regression**: In multiple regression we form a linear composite of explanatory variables in such way that it has maximum correlation with a criterion variable. This technique is appropriate when the researcher has a single, metric criterion variable. Which is supposed to be a function of other explanatory variables. The main objective in using this technique is to predict the variability the dependent variable based on its covariance with all the independent variables. One can predict the level of the dependent phenomenon through multiple regression analysis model, given the levels of independent variables. Given a dependent variable, the linear-multiple regression problem is to estimate constants $B_1$, $B_2$, ... $B_k$ and $A$ such that the expression $Y = B_1X_1 + B_2X_2 + ... + B_kX_k + A$ pare rovides a good estimate of an individual's $Y$ score based on his $X$ scores.

In practice, $Y$ and the several $X$ variables are converted to standard scores; $z_y$, $z_1$, $z_2$, ... $z_k$; each $z$ has a mean of 0 and standard deviation of 1. Then the problem is to estimate constants, $\beta_i$ , such that

$$z'_y = \beta_1 z_1 + \beta_2 z_2 + ... + \beta_k z_k$$

* See Chapter 7 also for other relevant information about multiple regression.

where $z'_y$ stands for the predicted value of the standardized $Y$ score, $z_y$. The expression on the right side of the above equation is the linear combination of explanatory variables. The constant $A$ is eliminated in the process of converting $X$'s to $z$'s. The least-squares-method is used, to estimate the beta weights in such a way that the sum of the squared prediction errors is kept as small as possible i.e., the expression $\sum\left(z_y - z'_y\right)^2$ is minimized. The predictive adequacy of a set of beta weights is indicated by the size of the correlation coefficient $r_{zy \cdot z'y}$ between the predicted $z'_y$ scores and the actual $z_y$ scores. This special correlation coefficient from Karl Pearson is termed the multiple correlation coefficient ($R$). The squared multiple correlation, $R^2$, represents the proportion of criterion ($z_y$) variance accounted for by the explanatory variables, i.e., the proportion of total variance that is 'Common Variance'.

Sometimes the researcher may use step-wise regression techniques to have a better idea of the independent contribution of each explanatory variable. Under these techniques, the investigator adds the independent contribution of each explanatory variable into the prediction equation one by one, computing betas and $R^2$ at each step. Formal computerized techniques are available for the purpose and the same can be used in the context of a particular problem being studied by the researcher.

(ii) *Multiple discriminant analysis:* Through discriminant analysis technique, researcher may classify individuals or objects into one of two or more mutually exclusive and exhaustive groups on the basis of a set of independent variables. Discriminant analysis requires interval independent variables and a nominal dependent variable. For example, suppose that brand preference (say brand *x* or *y)* is the dependent variable of interest and its relationship to an individual's income, age, education, etc. is being investigated, then we should use the technique of discriminant analysis. Regression analysis in such a situation is not suitable because the dependent variable is, not intervally scaled. Thus discriminant analysis is considered an appropriate technique when the single dependent variable happens to be non-metric and is to be classified into two or more groups, depending upon its relationship with several independent variables which all happen to be metric. The objective in discriminant analysis happens to be to predict an object's likelihood of belonging to a particular group based on several independent variables. In case we classify the dependent variable in more than two groups, then we use the name multiple discriminant analysis; but in case only two groups are to be formed, we simply use the term discriminant analysis.

We may briefly refer to the technical aspects* relating to discriminant analysis.

(i) There happens to be a simple scoring system that assigns a score to each individual or object. This score is a weighted average of the individual's numerical values of his independent variables. On the basis of this score, the individual is assigned to the 'most likely' category. For example, an individual is 20 years old, has an annual income of Rs 12,000,and has 10 years of formal education. Let $b_1$, $b_2$, and $b_3$ be the weights attached to the independent variables of age, income and education respectively. The individual's score ($z$), assuming linear score, would be:

$$z = b_1 (20) + b_2 (12000) + b_3 (10)$$

* Based on Robert Ferber, ed., *Handbook of Marketing Research.*

This numerical value of *z* can then be transformed into the probability that the individual is an early user, a late user or a non-user of the newly marketed consumer product (here we are making three categories viz. early user, late user or a non-user).

(ii) The numerical values and signs of the *b*'s indicate the importance of the independent variables in their ability to discriminate among the different classes of individuals. Thus, through the discriminant analysis, the researcher can as well determine which independent variables are most useful in predicting whether the respondent is to be put into one group or the other. In other words, discriminant analysis reveals which specific variables in the profile account for the largest proportion of inter-group differences.

(iii) In case only two groups of the individuals are to be formed on the basis of several independent variables, we can then have a model like this

$$z_i = b_0 + b_1X_{1i} + b_2X_{2i} + \ldots + b_nX_{ni}$$

where $X_{ji}$ = the *i*th individual's value of the *j*th independent variable;

$b_j$ = the discriminant coefficient of the *j*th variable;

$z_i$ = the *i*th individual's discriminant score;

$z_{\text{crit.}}$ = the critical value for the discriminant score.

The classification procedure in such a case would be

If $z_i > z_{\text{crit.}}$, classify individual *i* as belonging to Group I

If $z_i < z_{\text{crit}}$, classify individual *i* as belonging to Group II.

When *n* (the number of independent variables) is equal to 2, we have a straight line classification boundary. Every individual on one side of the line is classified as Group I and on the other side, every one is classified as belonging to Group II. When $n = 3$, the classification boundary is a two-dimensional plane in 3 space and in general the classification boundary is an $n - 1$ dimensional hyper-plane in *n* space.

(iv) In *n*-group discriminant analysis, a discriminant function is formed for each pair of groups. If there are 6 groups to be formed, we would have $6(6 - 1)/2 = 15$ pairs of groups, and hence 15 discriminant functions. The *b* values for each function tell which variables are important for discriminating between particular pairs of groups. The *z* score for each discriminant function tells in which of these two groups the individual is more likely to belong. Then use is made of the transitivity of the relation "more likely than". For example, if group II is more likely than group I and group III is more likely than group II, then group III is also more likely than group I. This way all necessary comparisons are made and the individual is assigned to the most likely of all the groups. Thus, the multiple-group discriminant analysis is just like the two-group discriminant analysis for the multiple groups are simply examined two at a time.

(v) For judging the statistical significance between two groups, we work out the Mahalanobis statistic, $D^2$, which happens to be a generalized distance between two groups, where each group is characterized by the same set of *n* variables and where it is assumed that variance-covariance structure is identical for both groups. It is worked out thus:

$$D^2 = (U_1 - U_2)v^{-1}(U_1 - U_2)'$$

where $U_1$ = the mean vector for group I

$U_2$ = the mean vector for group II
$v$ = the common variance matrix

By transformation procedure, this $D^2$ statistic becomes an $F$ statistic which can be used to see if the two groups are statistically different from each other.

From all this, we can conclude that the discriminant analysis provides a predictive equation, measures the relative importance of each variable and is also a measure of the ability of the equation to predict actual class-groups (two or more) concerning the dependent variable.

(iii) *Multivariate analysis of variance:* Multivariate analysis of variance is an extension of bivariate analysis of variance in which the ratio of among-groups variance to within-groups variance is calculated on a set of variables instead of a single variable. This technique is considered appropriate when several metric dependent variables are involved in a research study along with many non-metric explanatory variables. (But if the study has only one metric dependent variable and several non-metric explanatory variables, then we use the ANOVA technique as explained earlier in the book.) In other words, multivariate analysis of variance is specially applied whenever the researcher wants to test hypotheses concerning multivariate differences in group responses to experimental manipulations. For instance, the market researcher may be interested in using one test market and one control market to examine the effect of an advertising campaign on sales as well as awareness, knowledge and attitudes. In that case he should use the technique of multivariate analysis of variance for meeting his objective.

(iv) *Canonical correlation analysis:* This technique was first developed by Hotelling wherein an effort is made to simultaneously predict a set of criterion variables from their joint co-variance with a set of explanatory variables. Both metric and non-metric data can be used in the context of this multivariate technique. The procedure followed is to obtain a set of weights for the dependent and independent variables in such a way that linear composite of the criterion variables has a maximum correlation with the linear composite of the explanatory variables. For example, if we want to relate grade school adjustment to health and physical maturity of the child, we can then use canonical correlation analysis, provided we have for each child a number of adjustment scores (such as tests, teacher's ratings, parent's ratings and so on) and also we have for each child a number of health and physical maturity scores (such as heart rate, height, weight, index of intensity of illness and so on). The main objective of canonical correlation analysis is to discover factors separately in the two sets of variables such that the multiple correlation between sets of factors will be the maximum possible. Mathematically, in canonical correlation analysis, the weights of the two sets viz., $a_1, a_2, \ldots a_k$ and $y_1, y_2, y_3, \ldots y_j$ are so determined that the variables $X = a_1X_1 + a_2X_2 + \ldots + a_kX_k + a$ and $Y = y_1Y_1 + y_2Y_2 + \ldots y_jY_j + y$ have a maximum common variance. The process of finding the weights requires factor analyses with two matrices.* The resulting canonical correlation solution then gives an over all description of the presence or absence of a relationship between the two sets of variables.

(v) *Factor analysis:* Factor analysis is by far the most often used multivariate technique of research studies, specially pertaining to social and behavioural sciences. It is a technique applicable when there is a systematic interdependence among a set of observed or manifest variables and the researcher is interested in finding out something more fundamental or latent which creates this commonality. For instance, we might have data, say, about an individual's income, education, occupation and dwelling

* See, Eleanor W. Willemsen, *Understanding Statistical Reasoning*, p. 167–168.

area and want to infer from these some factor (such as social class) which summarises the commonality of all the said four variables. The technique used for such purpose is generally described as factor analysis. Factor analysis, thus, seeks to resolve a large set of measured variables in terms of relatively few categories, known as factors. This technique allows the researcher to group variables into factors (based on correlation between variables) and the factors so derived may be treated as new variables (often termed as latent variables) and their value derived by summing the values of the original variables which have been grouped into the factor. The meaning and name of such new variable is subjectively determined by the researcher. Since the factors happen to be linear combinations of data, the coordinates of each observation or variable is measured to obtain what are called factor loadings. Such factor loadings represent the correlation between the particular variable and the factor, and are usually place in a matrix of correlations between the variable and the factors.

*The mathematical basis of factor analysis* concerns a data matrix* (also termed as score matrix), symbolized as *S*. The matrix contains the scores of *N* persons of *k* measures. Thus $a_1$ is the score of person 1 on measure *a*, $a_2$ is the score of person 2 on measure *a*, and $k_N$ is the score of person *N* on measure *k*. The score matrix then take the form as shown following:

SCORE MATRIX (or Matrix *S*)

Measures (variables)

| Persons (objects) | *a* | *b* | *c* | *k* |
|---|---|---|---|---|
| 1 | $a_1$ | $b_1$ | $c_1$ | $k_1$ |
| 2 | $a_2$ | $b_2$ | $c_2$ | $k_2$ |
| 3 | $a_3$ | $b_3$ | $c_3$ | $k_3$ |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| *N* | $a_N$ | $b_N$ | $c_N$ | $k_N$ |

It is assumed that scores on each measure are standardized [i.e., $x_i = (X - \bar{X}_i)^2/\sigma_i$]. This being so, the sum of scores in any column of the matrix, *S*, is zero and the variance of scores in any column is 1.0. Then factors (a factor is any linear combination of the variables in a data matrix and can be stated in a general way like: $A = W_a a + W_b b + \ldots + W_k k$) are obtained (by any method of factoring). After this, we work out factor loadings (i.e., factor-variable correlations). Then communality, symbolized as $h^2$, the eigen value and the total sum of squares are obtained and the results interpreted. For realistic results, we resort to the technique of rotation, because such rotations reveal different structures in the data. Finally, factor scores are obtained which help in explaining what the factors mean. They also facilitate comparison among groups of items as groups. With factor scores, one can also perform several other multivariate analyses such as multiple regression, cluster analysis, multiple discriminant analysis, etc.

*Alternatively the technique can be applied through the matrix of correlations, *R* as stated later on.

## IMPORTANT METHODS OF FACTOR ANALYSIS

There are several methods of factor analysis, but they do not necessarily give same results. As such factor analysis is not a single unique method but a set of techniques. Important methods of factor analysis are:

(i) the centroid method;
(ii) the principal components method;
(ii) the maximum likelihood method.

Before we describe these different methods of factor analysis, it seems appropriate that some basic terms relating to factor analysis be well understood.

(i) *Factor:* A factor is an underlying dimension that account for several observed variables. There can be one or more factors, depending upon the nature of the study and the number of variables involved in it.

(ii) *Factor-loadings:* Factor-loadings are those values which explain how closely the variables are related to each one of the factors discovered. They are also known as factor-variable correlations. In fact, factor-loadings work as key to understanding what the factors mean. It is the absolute size (rather than the signs, plus or minus) of the loadings that is important in the interpretation of a factor.

(iii) *Communality* ($h^2$)*:* Communality, symbolized as $h^2$, shows how much of each variable is accounted for by the underlying factor taken together. A high value of communality means that not much of the variable is left over after whatever the factors represent is taken into consideration. It is worked out in respect of each variable as under:

$$h^2 \text{ of the } i\text{th variable} = (i\text{th factor loading of factor } A)^2 + (i\text{th factor loading of factor } B)^2 + \ldots$$

(iv) *Eigen value (or latent root):* When we take the sum of squared values of factor loadings relating to a factor, then such sum is referred to as Eigen Value or latent root. Eigen value indicates the relative importance of each factor in accounting for the particular set of variables being analysed.

(v) *Total sum of squares:* When eigen values of all factors are totalled, the resulting value is termed as the total sum of squares. This value, when divided by the number of variables (involved in a study), results in an index that shows how the particular solution accounts for what all the variables taken together represent. If the variables are all very different from each other, this index will be low. If they fall into one or more highly redundant groups, and if the extracted factors account for all the groups, the index will then approach unity.

(vi) *Rotation:* Rotation, in the context of factor analysis, is something like staining a microscope slide. Just as different stains on it reveal different structures in the tissue, different rotations reveal different structures in the data. Though different rotations give results that appear to be entirely different, but from a statistical point of view, all results are taken as equal, none superior or inferior to others. However, from the standpoint of making sense of the results of factor analysis, one must select the right rotation. If the factors are independent orthogonal rotation is done and if the factors are correlated, an oblique rotation is made. Communality for each variables will remain undisturbed regardless of rotation but the eigen values will change as result of rotation.

(vii) *Factor scores:* Factor score represents the degree to which each respondent gets high scores on the group of items that load high on each factor. Factor scores can help explain what the factors mean. With such scores, several other multivariate analyses can be performed.

We can now take up the important methods of factor analysis.

## (A) Centroid Method of Factor Analysis

This method of factor analysis, developed by L.L. Thurstone, was quite frequently used until about 1950 before the advent of large capacity high speed computers.* The centroid method tends to maximize the sum of loadings, disregarding signs; it is the method which extracts the largest sum of absolute loadings for each factor in turn. It is defined by linear combinations in which all weights are either + 1.0 or – 1.0. The main merit of this method is that it is relatively simple, can be easily understood and involves simpler computations. If one understands this method, it becomes easy to understand the mechanics involved in other methods of factor analysis.

Various steps** involved in this method are as follows:

(i) This method starts with the computation of a matrix of correlations, *R*, wherein unities are place in the diagonal spaces. The product moment formula is used for working out the correlation coefficients.

(ii) If the correlation matrix so obtained happens to be positive manifold (i.e., disregarding the diagonal elements each variable has a large sum of positive correlations than of negative correlations), the centroid method requires that the weights for all variables be +1.0. In other words, the variables are not weighted; they are simply summed. But in case the correlation matrix is not a positive manifold, then reflections must be made before the first centroid factor is obtained.

(iii) The first centroid factor is determined as under:

  (a) The sum of the coefficients (including the diagonal unity) in each column of the correlation matrix is worked out.

  (b) Then the sum of these column sums (*T*) is obtained.

  (c) The sum of each column obtained as per (a) above is divided by the square root of *T* obtained in (b) above, resulting in what are called centroid loadings. This way each centroid loading (one loading for one variable) is computed. The full set of loadings so obtained constitute the first centroid factor (say *A*).

(iv) To obtain second centroid factor (say *B*), one must first obtain a matrix of residual coefficients. For this purpose, the loadings for the two variables on the first centroid factor are multiplied. This is done for all possible pairs of variables (in each diagonal space is the square of the particular factor loading). The resulting matrix of factor cross products may be named as $Q_1$. Then $Q_1$ is subtracted clement by element from the original matrix of

*But since 1950, Principal components method, to be discussed a little later, is being popularly used.

**See, Jum C. Nunnally, *Psychometric Theory*, 2nd ed., p. 349–357, for details.

correlation, $R$, and the result is the first matrix of residual coefficients, $R_1$.* After obtaining $R_1$, one must *reflect* some of the variables in it, meaning thereby that some of the variables are given negative signs in the sum [This is usually done by inspection. The aim in doing this should be to obtain a reflected matrix, $R'_1$, which will have the highest possible sum of coefficients ($T$)]. For any variable which is so reflected, the signs of all coefficients in that column and row of the residual matrix are changed. When this is done, the matrix is named as 'reflected matrix' form which the loadings are obtained in the usual way (already explained in the context of first centroid factor), but the loadings of the variables which were reflected must be given negative signs. The full set of loadings so obtained constitutes the second centroid factor (say $B$). Thus loadings on the second centroid factor are obtained from $R'_1$.

(v) For subsequent factors ($C$, $D$, etc.) the same process outlined above is repeated. After the second centroid factor is obtained, cross products are computed forming, matrix, $Q_2$. This is then subtracted from $R_1$ (and not from $R'_1$) resulting in $R_2$. To obtain a third factor ($C$), one should operate on $R_2$ in the same way as on $R_1$. First, some of the variables would have to be reflected to maximize the sum of loadings, which would produce $R'_2$. Loadings would be computed from $R'_2$ as they were from $R'_1$. Again, it would be necessary to give negative signs to the loadings of variables which were reflected which would result in third centroid factor ($C$).

We may now illustrate this method by an example.

### *Illustration 1*

Given is the following correlation matrix, $R$, relating to eight variables with unities in the diagonal spaces:

| | | *Variables* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| | 1 | 1.000 | .709 | .204 | .081 | .626 | .113 | .155 | .774 |
| | 2 | .709 | 1.000 | .051 | .089 | .581 | .098 | .083 | .652 |
| | 3 | .204 | .051 | 1.000 | .671 | .123 | .689 | .582 | .072 |
| *Variables* | 4 | .081 | .089 | .671 | 1.000 | .022 | .798 | .613 | .111 |
| | 5 | .626 | .581 | .123 | .022 | 1.000 | .047 | .201 | .724 |
| | 6 | .113 | .098 | .689 | .798 | .047 | 1.000 | .801 | .120 |
| | 7 | .155 | .083 | .582 | .613 | .201 | .801 | 1.000 | .152 |
| | 8 | .774 | .652 | .072 | .111 | .724 | .120 | .152 | 1.000 |

Using the centroid method of factor analysis, work out the first and second centroid factors from the above information.

* One should understand the nature of the elements in $R_1$ matrix. Each diagonal element is a partial variance i.e., the variance that remains after the influence of the first factor is partialed. Each off-diagonal element is a partial co-variance i.e., the covariance between two variables after the influence of the first factor is removed. This can be verified by looking at the partial correlation coefficient between any two variables say 1 and 2 when factor $A$ is held constant

$$r_{12 \cdot A} = \frac{r_{12} - r_{1A} \cdot r_{2A}}{\sqrt{1 - r_{1A}^2}\sqrt{1 - r_{2A}^2}}$$

(The numerator in the above formula is what is found in $R_1$ corresponding to the entry for variables 1 and 2. In the denominator, the square of the term on the left is exactly what is found in the diagonal element for variable 1 in $R_1$. Likewise the partial variance for 2 is found in the diagonal space for that variable in the residual matrix.) *contd.*

***Solution:*** Given correlation matrix, *R*, is a positive manifold and as such the weights for all variables be +1.0. Accordingly, we calculate the first centroid factor (*A*) as under:

**Table 13.1(a)**

| | | *Variables* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| *Variables* | 1 | 1.000 | .709 | .204 | .081 | .626 | .113 | .155 | .774 |
| | 2 | .709 | 1.000 | .051 | .089 | .581 | .098 | .083 | .652 |
| | 3 | .204 | .051 | 1.000 | .671 | .123 | .689 | .582 | .072 |
| | 4 | .081 | .089 | .671 | 1.000 | .022 | .798 | .613 | .111 |
| | 5 | .626 | .581 | .123 | .022 | 1.000 | .047 | .201 | .724 |
| | 6 | .113 | .098 | .689 | .798 | .047 | 1.000 | .801 | .120 |
| | 7 | .155 | .083 | .582 | .613 | .201 | .801 | 1.000 | .152 |
| | 8 | .774 | .652 | .072 | .111 | .724 | .120 | .152 | 1.000 |
| Column sums | | 3.662 | 3.263 | 3.392 | 3.385 | 3.324 | 3.666 | 3.587 | 3.605 |

Sum of the column sums ($T$) = 27.884 $\quad \therefore \sqrt{T} = 5.281$

First centroid factor $A = \frac{3.662}{5.281}, \frac{3.263}{5.281}, \frac{3.392}{5.281}, \frac{3.385}{5.281}, \frac{3.324}{5.281}, \frac{3.666}{5.281}, \frac{3.587}{5.281}, \frac{3.605}{5.281}$

$= .693, .618, .642, .641, .629, .694, .679, .683$

We can also state this information as under:

**Table 13.1 (b)**

| *Variables* | *Factor loadings concerning first Centroid factor A* |
|---|---|
| 1 | .693 |
| 2 | .618 |
| 3 | .642 |
| 4 | .641 |
| 5 | .629 |
| 6 | .694 |
| 7 | .679 |
| 8 | .683 |

To obtain the second centroid factor *B*, we first of all develop (as shown on the next page) the first matrix of factor cross product, $Q_1$:

Since in $R_1$ the diagonal terms are partial variances and the off-diagonal terms are partial covariances, it is easy to convert the entire table to a matrix of partial correlations. For this purpose one has to divide the elements in each row by the square-root of the diagonal element for that row and then dividing the elements in each column by the square-root of the diagonal element for that column.

First Matrix of Factor Cross Product ($Q_1$)

| First centroid factor $A$ | .693 | .618 | .642 | .641 | .629 | .694 | .679 | .683 |
|---|---|---|---|---|---|---|---|---|
| .693 | .480 | .428 | .445 | .444 | .436 | .481 | .471 | .473 |
| .618 | .428 | .382 | .397 | .396 | .389 | .429 | .420 | .422 |
| .642 | .445 | .397 | .412 | .412 | .404 | .446 | .436 | .438 |
| .641 | .444 | .396 | .412 | .411 | .403 | .445 | .435 | .438 |
| .629 | .436 | .389 | .404 | .403 | .396 | .437 | .427 | .430 |
| .694 | .481 | .429 | .446 | .445 | .437 | .482 | .471 | .474 |
| .679 | .471 | .420 | .436 | .435 | .427 | .471 | .461 | .464 |
| .683 | .473 | .422 | .438 | .438 | .430 | .474 | .464 | .466 |

Now we obtain first matrix of residual coefficient ($R_1$) by subtracting $Q_1$ from $R$ as shown below:

First Matrix of Residual Coefficient ($R_1$)

| | *Variables* | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| *Variables* | 1 | .520 | .281 | –.241 | –.363 | .190 | –.368 | –.316 | .301 |
| | 2 | .281 | .618 | –.346 | –.307 | .192 | –.331 | –.337 | .230 |
| | 3 | –.241 | –.346 | .588 | .259 | –.281 | .243 | .146 | –.366 |
| | 4 | –.363 | –.307 | .259 | .589 | –.381 | .353 | .178 | –.327 |
| | 5 | .190 | .192 | –.281 | –.381 | .604 | –.390 | –.217 | .294 |
| | 6 | –.368 | –.331 | .243 | .353 | –.390 | .518 | .330 | –.354 |
| | 7 | –.316 | –.337 | .146 | .178 | –.226 | .330 | .539 | –.312 |
| | 8 | .301 | .230 | –.366 | –.327 | .294 | –.354 | –.312 | .534 |

Reflecting the variables 3, 4, 6 and 7, we obtain reflected matrix of residual coefficient ($R'_1$) as under and then we can extract the second centroid factor ($B$) from it as shown on the next page.

Reflected Matrix of Residual Coefficients ($R'_1$)
and Extraction of 2nd Centroid Factor ($B$)

| | *Variables* | 1 | 2 | 3* | 4* | 5 | 6* | 7* | 8 |
|---|---|---|---|---|---|---|---|---|---|
| *Variables* | 1 | .520 | .281 | .241 | .363 | .190 | .368 | .316 | .301 |
| | 2 | .281 | .618 | .346 | .307 | .192 | .331 | .337 | .230 |
| | 3* | .241 | .346 | .588 | .259 | .281 | .243 | .146 | .366 |
| | 4* | .363 | .307 | .259 | .589 | .381 | .353 | .178 | .327 |
| | 5 | .190 | .192 | .281 | .381 | .604 | .390 | .217 | .294 |
| | 6* | .368 | .331 | .243 | .353 | .390 | .518 | .330 | .354 |
| | 7* | .316 | .337 | .146 | .178 | .226 | .330 | .539 | .312 |

*Contd.*

*Variables*

| | 1 | 2 | 3* | 4* | 5 | 6* | 7* | 8 |
|---|---|---|---|---|---|---|---|---|
| 8 | .301 | .230 | .366 | .327 | .294 | .354 | .312 | .534 |
| Column sums: | 2.580 | 2.642 | 2.470 | 2.757 | 2.558 | 2.887 | 2.375 | 2.718 |
| Sum of column sums ($T$) = 20.987 $\therefore$ $\sqrt{T}$ = 4.581 | | | | | | | | |
| Second centroid factor $B$ = .563 .577 –.539 –.602 .558 –.630 –.518 .593 | | | | | | | | |

*These variables were reflected.

Now we can write the matrix of factor loadings as under:

| *Variables* | *Factor loadings* | |
|---|---|---|
| | *Centroid Factor A* | *Centroid Factor B* |
| 1 | .693 | .563 |
| 2 | .618 | .577 |
| 3 | .642 | –.539 |
| 4 | .641 | –.602 |
| 5 | .629 | .558 |
| 6 | .694 | –.630 |
| 7 | .679 | –.518 |
| 8 | .683 | .593 |

***Illustration 2***

Work out the communality and eigen values from the final results obtained in illustration No. 1 of this chapter. Also explain what they (along with the said two factors) indicate.

***Solution:*** We work out the communality and eigen values for the given problem as under:

**Table 13.2**

| *Variables* | *Factor loadings* | | *Communality ($h^2$)* |
|---|---|---|---|
| | *Centroid Factor A* | *Centroid Factor B* | |
| 1 | .693 | .563 | $(.693)^2 + (.563)^2 = .797$ |
| 2 | .618 | .577 | $(.618)^2 + (.577)^2 = .715$ |
| 3 | .642 | –.539 | $(.642)^2 + (–.539)^2 = .703$ |
| 4 | .641 | –.602 | $(.641)^2 + (–.602)^2 = .773$ |
| 5 | .629 | .558 | $(.629)^2 + (.558)^2 = .707$ |
| 6 | .694 | –.630 | $(.694)^2 + (–.630)^2 = .879$ |
| 7 | .679 | –.518 | $(.679)^2 + (–.518)^2 = .729$ |
| 8 | .683 | .593 | $(.683)^2 + (.593)^2 = .818$ |

*Contd.*

| *Variables* | *Factor loadings* | | *Communality* ($h^2$) |
|---|---|---|---|
| | *Centroid Factor A* | *Centroid Factor B* | |
| Eigen value (Variance accounted for i.e., common variance) | 3.490 | 2.631 | 6.121 |
| Proportion of total variance | .44 (44%) | .33 (33%) | .77 (77%) |
| Proportion of common variance | .57 (57%) | .43 (43%) | 1.00 (100%) |

Each communality in the above table represents the proportion of variance in the corresponding (row) variable and is accounted for by the two factors (*A* and *B*). For instance, 79.7% of the variance in variable one is accounted for by the centroid factor *A* and *B* and the remaining 20.3% of the total variance in variable one scores is thought of as being made up of two parts: a factor specific to the attribute represented by variable one, and a portion due to errors of measurement involved in the assessment of variable one (but there is no mention of these portions in the above table because we usually concentrate on common variance in factor analysis).

It has become customary in factor analysis literature for a loading of 0.33 to be the minimum absolute value to be interpreted. The portion of a variable's variance accounted for by this minimum loading is approximately 10%. This criterion, though arbitrary, is being used more or less by way of convention, and as such must be kept in view when one reads and interprets the multivariate research results. In our example, factor *A* has loading in excess of 0.33 on all variables; such a factor is usually called "*the general factor*" and is taken to represent whatever it is that all of the variables have in common. We might consider all the eight variables to be product of some unobserved variable (which can be named subjectively by the researcher considering the nature of his study). The factor name is chosen in such a way that it conveys what it is that all variables that correlate with it (that "load on it") have in common. Factor *B* in our example has all loadings in excess of 0.33, but half of them are with negative signs. Such a factor is called a "*bipolar factor*" and is taken to represent a single dimension with two poles. Each of these poles is defined by a cluster of variables—one pole by those with positive loadings and the other pole with negative loadings.

We can give different names to the said two groups to help us interpret and name factor *B*. The rows at the bottom of the above table give us further information about the usefulness of the two factors in explaining the relations among the eight variables. The total variance (*V*) in the analysis is taken as equal to the number of variables involved (on the presumption that variables are standardized). In this present example, then $V = 8.0$. The row labeled "Eigen value" or "Common variance" gives the numerical value of that portion of the variance attributed to the factor in the concerning column above it. These are found by summing up the squared values of the corresponding factor loadings. Thus the total value, 8.0, is partitioned into 3.490 as eigen value for factor *A* and 2.631 as eigen value for factor *B* and the total 6.121 as the sum of eigen values for these two factors. The corresponding proportion of the total variance, 8.0, are shown in the next row; there we can notice that 77% of the

total variance is related to these two factors, i.e., approximately 77% of the total variance is common variance whereas remaining 23% of it is made up of portions unique to individual variables and the techniques used to measure them. The last row shows that of the common variance approximately 57% is accounted for by factor *A* and the other 43% by factor *B*. Thus it can be concluded that the two factors together "explain" the common variance.

## (B) Principal-components Method of Factor Analysis

Principal-components method (or simply P.C. method) of factor analysis, developed by H. Hotelling, seeks to maximize the sum of squared loadings of each factor extracted in turn. Accordingly PC factor explains more variance than would the loadings obtained from any other method of factoring.

The aim of the principal components method is the construction out of a given set of variables $X_j$'s ($j = 1, 2, \ldots, k$) of new variables ($p_i$), called principal components which are linear combinations of the $X_s$

$$p_1 = a_{11} X_1 + a_{12} X_2 + \ldots + a_{1k} X_k$$
$$p_2 = a_{21} X_1 + a_{22} X_2 + \ldots + a_{2k} X_k$$
$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots \qquad \vdots$$
$$p_k = a_{k1} X_1 + a_{k2} X_2 + \ldots + a_{kk} X_k$$

The method is being applied mostly by using standardized variables, i.e., $z_j = \left(X_j - \bar{X}_j\right)^2 / \sigma_j$.

The $a_{ij}$'s are called loadings and are worked out in such a way that the extracted principal components satisfy two conditions: (i) principal components are uncorrelated (orthogonal) and (ii) the first principal component ($p_1$) has the maximum variance, the second principal component ($p_2$) has the next maximum variance and so on.

*Following steps are usually involved in principal components method*

(i) Estimates of $a_{ij}$'s are obtained with which *X*'s are transformed into orthogonal variables i.e., the principal components. A decision is also taken with regard to the question: how many of the components to retain into the analysis?

(ii) We then proceed with the regression of *Y* on these principal components i.e.,

$$Y = \hat{y}_1 p_1 + \hat{y}_2 p_2 + \ldots + \hat{y}_m p_m \quad (m < k)$$

(iii) From the $\hat{a}_{ij}$ and $\hat{y}_{ij}$, we may find $b_{ij}$ of the original model, transferring back from the *p*'s into the standardized *X*'s.

*Alternative method for finding the factor loadings is as under:*

(i) Correlation coefficients (by the product moment method) between the pairs of *k* variables are worked out and may be arranged in the form of a correlation matrix, *R*, as under:

**Correlation Matrix, *R***

*Variables*

| | | $X_1$ | $X_2$ | $X_3$ | .... | $X_k$ |
|---|---|---|---|---|---|---|
| *Variables* | $X_1$ | $r_{11}$ | $r_{12}$ | $r_{13}$ | .... | $r_{1k}$ |
| | $X_2$ | $r_{21}$ | $r_{22}$ | $r_{23}$ | .... | $r_{3k}$ |
| | $X_3$ | $r_{31}$ | $r_{32}$ | $r_{33}$ | .... | $r_{3k}$ |
| | . | . | . | . | | |
| | . | . | . | . | | |
| | $X_k$ | $r_{k1}$ | $r_{k2}$ | $r_{k3}$ | .... | $r_{kk}$ |

The main diagonal spaces include unities since such elements are self-correlations. The correlation matrix happens to be a symmetrical matrix.

(ii) Presuming the correlation matrix to be positive manifold (if this is not so, then reflections as mentioned in case of centroid method must be made), the first step is to obtain the sum of coefficients in each column, including the diagonal element. The vector of column sums is referred to as $U_{a1}$ and when $U_{a1}$ is normalized, we call it $V_{a1}$. This is done by squaring and summing the column sums in $U_{a1}$ and then dividing each element in $U_{a1}$ by the square root of the sum of squares (which may be termed as normalizing factor). Then elements in $V_{a1}$ are accumulatively multiplied by the first row of $R$ to obtain the first element in a new vector $U_{a2}$. For instance, in multiplying $V_{a1}$ by the first row of $R$, the first element in $V_{a1}$ would be multiplied by the $r_{11}$ value and this would be added to the product of the second element in $V_{a1}$ multiplied by the $r_{12}$ value, which would be added to the product of third element in $V_{a1}$ multiplied by the $r_{13}$ value, and so on for all the corresponding elements in $V_{a1}$ and the first row of $R$. To obtain the second element of $U_{a2}$, the same process would be repeated i.e., the elements in $V_{a1}$ are accumulatively multiplied by the 2nd row of $R$. The same process would be repeated for each row of $R$ and the result would be a new vector $U_{a2}$. Then $U_{a2}$ would be normalized to obtain $V_{a2}$. One would then compare $V_{a1}$ and $V_{a2}$. If they are nearly identical, then convergence is said to have occurred (If convergence does not occur, one should go on using these trial vectors again and again till convergence occurs). Suppose the convergence occurs when we work out $V_{a8}$ in which case $V_{a7}$ will be taken as $V_a$ (the characteristic vector) which can be converted into loadings on the first principal component when we multiply the said vector (i.e., each element of $V_a$) by the square root of the number we obtain for normalizing $U_{a8}$.

(iii) To obtain factor $B$, one seeks solutions for $V_b$, and the actual factor loadings for second component factor, $B$. The same procedures are used as we had adopted for finding the first factor, except that one operates off the first residual matrix, $R_1$ rather than the original correlation matrix $R$ (We operate on $R_1$ in just the same way as we did in case of centroid method stated earlier).

(iv) This very procedure is repeated over and over again to obtain the successive PC factors (viz. $C$, $D$, etc.).

*Other steps involved in factor analysis*

(a) Next the question is: How many principal components to retain in a particular study? Various criteria for this purpose have been suggested, but one often used is Kaiser's criterion. According to this criterion only the principal components, having latent root greater than one, are considered as essential and should be retained.

(b) The principal components so extracted and retained are then rotated from their beginning position to enhance the interpretability of the factors.

(c) Communality, symbolized, $h^2$, is then worked out which shows how much of each variable is accounted for by the underlying factors taken together. A high communality figure means that not much of the variable is left over after whatever the factors represent is taken into consideration. It is worked out in respect of each variable as under:

$$h^2 \text{ of the } i\text{th variable} = (i\text{th factor loading of factor } A)^2 + (i\text{th factor loading of factor } B)^2 + \ldots$$

Then follows the task of interpretation. The amount of variance explained (sum of squared loadings) by each PC factor is equal to the corresponding characteristic root. When these roots are divided by the number of variables, they show the characteristic roots as proportions of total variance explained.

(d) The variables are then regressed against each factor loading and the resulting regression coefficients are used to generate what are known as factor scores which are then used in further analysis and can also be used as inputs in several other multivariate analyses.

***Illustration 3***

Take the correlation matrix, *R*, for eight variables of illustration 1 of this chapter and then compute:

(i) the first two principal component factors;

(ii) the communality for each variable on the basis of said two component factors;

(iii) the proportion of total variance as well as the proportion of common variance explained by each of the two component factors.

***Solution:*** Since the given correlation matrix is a positive manifold, we work out the first principal component factor (using trial vectors) as under:

**Table 13.3**

| | | *Variables* | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| | 1 | 1.000 | .709 | .204 | .081 | .626 | .113 | .155 | .774 |
| | 2 | .709 | 1.000 | .051 | .089 | .581 | .098 | .083 | .652 |
| | 3 | .204 | .051 | 1.000 | .671 | .123 | .689 | .582 | .072 |
| | 4 | .081 | .089 | .671 | 1.000 | .022 | .798 | .613 | .111 |
| *Variables* | 5 | .626 | .581 | .123 | .022 | 1.000 | .047 | .201 | .724 |
| | 6 | .113 | .098 | .689 | .798 | .047 | 1.000 | .801 | .120 |

*Contd.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 7 | .155 | .083 | .582 | .613 | .201 | .801 | 1.000 | .152 |
| 8 | .774 | .652 | .072 | .111 | .724 | .120 | .152 | 1.000 |
| Column sums $U_{a1}$ | 3.662 | 3.263 | 3.392 | 3.385 | 3.324 | 3.666 | 3.587 | 3.605 |
| Normalizing $U_{a1}$ we obtain $V_{a1}$ i.e., $V_{a1}$ = $U_a$/Normalizing factor* | .371 | .331 | .344 | .343 | .337 | .372 | .363 | .365 |

*Normalizing factor = $\sqrt{(3.662)^2 + (3.263)^2 + (3.392)^2 + (3.385)^2 + (3.324)^2 + (3.666)^2 + (3.587)^2 + (3.605)^2}$

$= \sqrt{97.372} = 9.868$

Then we obtain $U_{a2}$ by accumulatively multiplying $V_{a1}$ row by row into $R$ and the result comes as under:

$U_{a2}$ : [1.296, 1.143, 1.201, 1.201, 1.165, 1.308, 1.280, 1.275]

Normalizing it we obtain (normalizing factor for $U_{a2}$ will be worked out as above and will be = 3.493):

$V_{a2}$ : [.371, .327, .344, .344, .334, .374, .366, .365]

Comparing $V_{a1}$ and $V_{a2}$, we find the two vectors are almost equal and this shows convergence has occurred. Hence $V_{a1}$ is taken as the characteristic vector, $V_a$. Finally, we compute the loadings on the first principal component by multiplying $V_a$ by the square root of the number that we obtain for normalizing $U_{a2}$. The result is as under:

| *Variables* | (*Characteristic vector* $V_a$) | × | $\sqrt{\text{normalizing factor of } U_{a2}}$ | = | *Principal Component I* |
|---|---|---|---|---|---|
| 1 | .371 | × | 1.868 | = | .69 |
| 2 | .331 | × | 1.868 | = | .62 |
| 3 | .344 | × | 1.868 | = | .64 |
| 4 | .343 | × | 1.868 | = | .64 |
| 5 | .337 | × | 1.868 | = | .63 |
| 6 | .372 | × | 1.868 | = | .70 |
| 7 | .363 | × | 1.868 | = | .68 |
| 8 | .365 | × | 1.868 | = | .68 |

For finding principal component II, we have to proceed on similar lines (as stated in the context of obtaining centroid factor *B* earlier in this chapter) to obtain the following result*:

| *Variables* | *Principal Component II* |
|---|---|
| 1 | +.57 |
| 2 | +.59 |
| 3 | –.52 |
| 4 | –.59 |
| 5 | +.57 |
| 6 | –.61 |
| 7 | –.49 |
| 8 | –.61 |

The other parts of the question can now be worked out (after first putting the above information in a matrix form) as given below:

| *Variables* | *Principal Components* | | *Communality, $h^2$* |
|---|---|---|---|
| | *I* | *II* | |
| 1 | .69 | +.57 | $(.69)^2 + (.57)^2 = .801$ |
| 2 | .62 | +.59 | $(.62)^2 + (.59)^2 = .733$ |
| 3 | .64 | –.52 | $(.64)^2 + (–.52)^2 = .680$ |
| 4 | .64 | –.59 | $(.64)^2 + (–.59)^2 = .758$ |
| 5 | .63 | +.57 | $(.63)^2 + (.57)^2 = .722$ |
| 6 | .70 | –.61 | $(.70)^2 + (–.61)^2 = .862$ |
| 7 | .68 | –.49 | $(.68)^2 + (–.49)^2 = .703$ |
| 8 | .68 | –.61 | $(.68)^2 + (–.61)^2 = .835$ |
| Eigen value i.e., common variance | 3.4914 | 2.6007 | 6.0921 |
| Proportion of total variance | .436 (43.6%) | .325 (32.5%) | .761 (76%) |
| Proportion of common variance | .573 (57%) | .427 (43%) | 1.000 (100%) |

All these values can be interpreted in the same manner as stated earlier.

*This can easily be worked out. Actual working has been left as an exercise for the students.

## (C) Maximum Likelihood (ML) Method of Factor Analysis

The ML method consists in obtaining sets of factor loadings successively in such a way that each, in turn, explains as much as possible of the population correlation matrix as estimated from the sample correlation matrix. If $R_s$ stands for the correlation matrix actually obtained from the data in a sample, $R_p$ stands for the correlation matrix that would be obtained if the entire population were tested, then the ML method seeks to extrapolate what is known from $R_s$ in the best possible way to estimate $R_p$ (but the PC method only maximizes the variance explained in $R_s$). Thus, the ML method is a statistical approach in which one maximizes some relationship between the sample of data and the population from which the sample was drawn.

The arithmetic underlying the ML method is relatively difficult in comparison to that involved in the PC method and as such is understandable when one has adequate grounding in calculus, higher algebra and matrix algebra in particular. Iterative approach is employed in ML method also to find each factor, but the iterative procedures have proved much more difficult than what we find in the case of PC method. Hence the ML method is generally not used for factor analysis in practice.[*]

The loadings obtained on the first factor are employed in the usual way to obtain a matrix of the residual coefficients. A significance test is then applied to indicate whether it would be reasonable to extract a second factor. This goes on repeatedly in search of one factor after another. One stops factoring after the significance test fails to reject the null hypothesis for the residual matrix. The final product is a matrix of factor loadings. The ML factor loadings can be interpreted in a similar fashion as we have explained in case of the centroid or the PC method.

## ROTATION IN FACTOR ANALYSIS

One often talks about the rotated solutions in the context of factor analysis. This is done (i.e., a factor matrix is subjected to rotation) to attain what is technically called "simple structure" in data. Simple structure according to L.L. Thurstone is obtained by rotating the axes[**] until:

(i) Each row of the factor matrix has one zero.

(ii) Each column of the factor matrix has $p$ zeros, where $p$ is the number of factors.

(iii) For each pair of factors, there are several variables for which the loading on one is virtually zero and the loading on the other is substantial.

(iv) If there are many factors, then for each pair of factors there are many variables for which both loadings are zero.

(v) For every pair of factors, the number of variables with non-vanishing loadings on both of them is small.

All these criteria simply imply that the factor analysis should reduce the complexity of all the variables.

[*] The basic mathematical derivations of the ML method are well explained in S.A. Mulaik's, *The Foundations of Factor Analysis.*

[**] Rotation constitutes the geometric aspects of factor analysis. Only the axes of the graph (wherein the points representing variables have been shown) are rotated keeping the location of these points relative to each other undisturbed.

There are several methods of rotating the initial factor matrix (obtained by any of the methods of factor analysis) to attain this simple structure. *Varimax rotation* is one such method that maximizes (simultaneously for all factors) the variance of the loadings within each factor. The variance of a factor is largest when its smallest loadings tend towards zero and its largest loadings tend towards unity. In essence, the solution obtained through varimax rotation produces factors that are characterized by large loadings on relatively few variables. The other method of rotation is known as *quartimax rotation* wherein the factor loadings are transformed until the variance of the squared factor loadings throughout the matrix is maximized. As a result, the solution obtained through this method permits a general factor to emerge, whereas in case of varimax solution such a thing is not possible. But both solutions produce orthogonal factors i.e., uncorrelated factors. It should, however, be emphasised that right rotation must be selected for making sense of the results of factor analysis.

## *R*-TYPE AND *Q*-TYPE FACTOR ANALYSES

Factor analysis may be R-type factor analysis or it may be Q-type factor analysis. In *R-type factor analysis*, high correlations occur when respondents who score high on variable 1 also score high on variable 2 and respondents who score low on variable 1 also score low on variable 2. Factors emerge when there are high correlations within groups of variables. In *Q-type factor analysis*, the correlations are computed between pairs of respondents instead of pairs of variables. High correlations occur when respondent 1's pattern of responses on all the variables is much like respondent 2's pattern of responses. Factors emerge when there are high correlations within groups of people. *Q*-type analysis is useful when the object is to sort out people into groups based on their simultaneous responses to all the variables.

Factor analysis has been mainly used in developing psychological tests (such as *IQ* tests, personality tests, and the like) in the realm of psychology. In marketing, this technique has been used to look at media readership profiles of people.

*Merits:* The main merits of factor analysis can be stated thus:

(i) The technique of factor analysis is quite useful when we want to condense and simplify the multivariate data.

(ii) The technique is helpful in pointing out important and interesting, relationships among observed data that were there all the time, but not easy to see from the data alone.

(iii) The technique can reveal the latent factors (i.e., underlying factors not directly observed) that determine relationships among several variables concerning a research study. For example, if people are asked to rate different cold drinks (say, Limca, Nova-cola, Gold Spot and so on) according to preference, a factor analysis may reveal some salient characteristics of cold drinks that underlie the relative preferences.

(iv) The technique may be used in the context of empirical clustering of products, media or people i.e., for providing a classification scheme when data scored on various rating scales have to be grouped together.

*Limitations:* One should also be aware of several limitations of factor analysis. Important ones are as follows:

(i) Factor analysis, like all multivariate techniques, involves laborious computations involving heavy cost burden. With computer facility available these days, there is no doubt that factor analysis has become relatively faster and easier, but the cost factor continues to be the same i.e., large factor analyses are still bound to be quite expensive.

(ii) The results of a single factor analysis are considered generally less reliable and dependable for very often a factor analysis starts with a set of imperfect data. "The factors are nothing but blurred averages, difficult to be identified."[4] To overcome this difficulty, it has been realised that analysis should at least be done twice. If we get more or less similar results from all rounds of analyses, our confidence concerning such results increases.

(iii) Factor-analysis is a complicated decision tool that can be used only when one has thorough knowledge and enough experience of handling this tool. Even then, at times it may not work well and may even disappoint the user.

To conclude, we can state that in spite of all the said limitations "when it works well, factor analysis helps the investigator make sense of large bodies of intertwined data. When it works unusually well, it also points out some interesting relationships that might not have been obvious from examination of the input data alone".[5]

### (vi) *Cluster Analysis*

Cluster analysis consists of methods of classifying variables into clusters. Technically, a cluster consists of variables that correlate highly with one another and have comparatively low correlations with variables in other clusters. The basic objective of cluster analysis is to determine how many mutually and exhaustive groups or clusters, based on the similarities of profiles among entities, really exist in the population and then to state the composition of such groups. Various groups to be determined in cluster analysis are not predefined as happens to be the case in discriminant analysis.

*Steps:* In general, cluster analysis contains the following steps to be performed:

(i) First of all, if some variables have a negative sum of correlations in the correlation matrix, one must reflect variables so as to obtain a maximum sum of positive correlations for the matrix as a whole.

(ii) The second step consists in finding out the highest correlation in the correlation matrix and the two variables involved (i.e., having the highest correlation in the matrix) form the nucleus of the first cluster.

(iii) Then one looks for those variables that correlate highly with the said two variables and includes them in the cluster. This is how the first cluster is formed.

(iv) To obtain the nucleus of the second cluster, we find two variables that correlate highly but have low correlations with members of the first cluster. Variables that correlate highly with the said two variables are then found. Such variables along the said two variables thus constitute the second cluster.

(v) One proceeds on similar lines to search for a third cluster and so on.

[4] Srinibas Bhattacharya, *Psychometrics and Behavioural Research,* p. 177.

[5] William D. Wells and Jagdish N. Sheth in their article on "Factor Analysis" forming chapter 9 in Robert Ferber, (ed.), *Handbook of Marketing Research,* p. 2–471.

From the above description we find that clustering methods in general are judgemental and are devoid of statistical inferences. For problems concerning large number of variables, various cut-and-try methods have been proposed for locating clusters. McQuitty has specially developed a number of rather elaborate computational routines* for that purpose.

In spite of the above stated limitation, cluster analysis has been found useful in context of market research studies. Through the use of this technique we can make segments of market of a product on the basis of several characteristics of the customers such as personality, socio-economic considerations, psychological factors, purchasing habits and like ones.

### (vii) *Multidimensional Scaling***

Multidimensional scaling (MDS) allows a researcher to measure an item in more than one dimension at a time. The basic assumption is that people perceive a set of objects as being more or less similar to one another on a number of dimensions (usually uncorrelated with one another) instead of only one.

There are several MDS techniques (also known as techniques for dimensional reduction) often used for the purpose of revealing patterns of one sort or another in interdependent data structures. If data happen to be non-metric, MDS involves rank ordering each pair of objects in terms of similarity. Then the judged similarities are transformed into distances through statistical manipulations and are consequently shown in *n*-dimensional space in a way that the interpoint distances best preserve the original interpoint proximities. After this sort of mapping is performed, the dimensions are usually interpreted and labeled by the researcher.

The significance of MDS lies in the fact that it enables the researcher to study "The perceptual structure of a set of stimuli and the cognitive processes underlying the development of this structure.... MDS provides a mechanism for determining the truly salient attributes without forcing the judge to appear irrational."[6] With MDS, one can scale objects, individuals or both with a minimum of information. The MDS analysis will reveal the most salient attributes which happen to be the primary determinants for making a specific decision.

### (viii) *Latent Structure Analysis*

This type of analysis shares both of the objectives of factor analysis viz., to extract latent factors and express relationship of observed (manifest) variables with these factors as their indicators and to classify a population of respondents into pure types. This type of analysis is appropriate when the variables involved in a study do not possess dependency relationship and happen to be non-metric.

In addition to the above stated multivariate techniques, we may also describe the salient features of what is known as "Path analysis", a technique useful for decomposing the total correlation between any two variables in a causal system.

* These are beyond the scope of this book and hence have been omitted. Readers interested in such methods are referred to "Cluster Analysis" by R. C. Tryon and D. E. Bailey.

** See, Chapter No. 5 of this book for other details about MDS.

[6] Robert Ferber, ed., *Handbook of Marketing Research,* p. 3–52.

## PATH ANALYSIS

The term 'path analysis' was first introduced by the biologist Sewall Wright in 1934 in connection with decomposing the total correlation between any two variables in a causal system. The technique of path analysis is based on a series of multiple regression analyses with the added assumption of causal relationship between independent and dependent variables. This technique lays relatively heavier emphasis on the heuristic use of visual diagram, technically described as a path diagram. An illustrative path diagram showing interrelationships between Fathers' education, Fathers' occupation, Sons' education, Sons' first and Sons' present occupation can be shown in the Fig. 13.2.

Path analysis makes use of standardized partial regression coefficients (known as beta weights) as effect coefficients. In linear additive effects are assumed, then through path analysis a simple set of equations can be built up showing how each variable depends on preceding variables. "The main principle of path analysis is that any correlation coefficient between two variables, or a gross or overall measure of empirical relationship can be decomposed into a series of parts: separate paths of influence leading through chronologically intermediate variable to which both the correlated variables have links."[7]

The merit of path analysis in comparison to correlational analysis is that it makes possible the assessment of the relative influence of each antecedent or explanatory variable on the consequent or criterion variables by first making explicit the assumptions underlying the causal connections and then by elucidating the indirect effect of the explanatory variables.

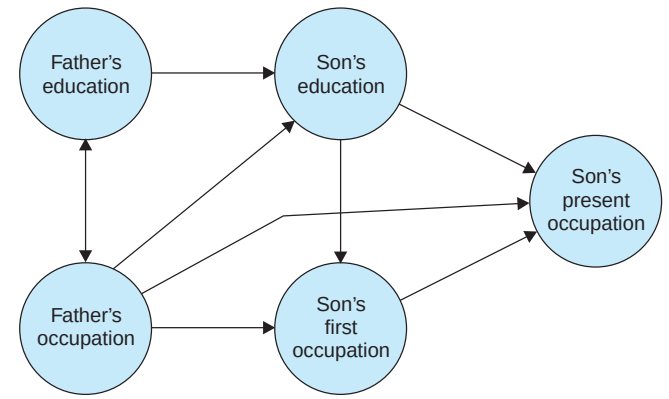


**Fig.13.2**

"The use of the path analysis technique requires the assumption that there are linear additive, a symmetric relationships among a set of variables which can be measured at least on a quasi-interval scale. Each dependent variable is regarded as determined by the variables preceding it in the path diagram, and a residual variable, defined as uncorrelated with the other variables, is postulated to account for the unexplained portion of the variance in the dependent variable. The determining variables are assumed for the analysis to be given (exogenous in the model)."[8]

[7] K. Takeuchi, *et al. op. cit., The Foundations of Multivariate Analysis*, p. 122.

[8] *Ibid.*, p. 121–122.

We may illustrate the path analysis technique in connection with a simple problem of testing a causal model with three explicit variables as shown in the following path diagram:

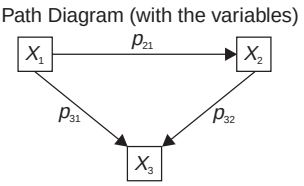


**Fig. 13.3**

The structural equation for the above can be written as:

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} e_1 \\ p_{21}X_1 + e_2 \\ p_{31}X_2 + p_{32}X_2 + e_3 \end{bmatrix} = pX + e$$

where the $X$ variables are measured as deviations from their respective means. $p_{21}$ may be estimated from the simple regression of $X_2$ on $X_1$ i.e., $X_2 = b_{21}X_1$ and $p_{31}$ and $p_{32}$ may be estimated from the regression of $X_3$ on $X_2$ and $X_1$ as under:

$$\hat{X}_3 = b_{31.2}\, X_1 + b_{2.1}\, X_2$$

where $b_{31.2}$ means the standardized partial regression coefficient for predicting variable 3 from variable 1 when the effect of variable 2 is held constant.

In path analysis the beta coefficient indicates the direct effect of $X_j$ ($j$ = 1, 2, 3, ..., $p$) on the dependent variable. Squaring the direct effect yields the proportion of the variance in the dependent variable $Y$ which is due to each of the $p$ number of independent variables $X_j$ ($i$ = 1, 2, 3, ..., $p$). After calculating the direct effect, one may then obtain a summary measure of the total indirect effect of $X_j$ on the dependent variable $Y$ by subtracting from the zero correlation coefficient $r_{yxj}$, the beta coefficient $b_j$ i.e.,

$$\text{Indirect effect of } X_j \text{ on } Y = c_{jy} = r_{yxj} - b_j$$
$$\text{for all } j = 1, 2, ..., p.$$

Such indirect effects include the unanalysed effects and spurious relationships due to antecedent variables.

In the end, it may again be emphasised that the main virtue of path analysis lies in making explicit the assumptions underlying the causal connections and in elucidating the indirect effects due to antecedent variables of the given system.

## CONCLUSION

From the brief account of multivariate techniques presented above, we may conclude that such techniques are important for they make it possible to encompass all the data from an investigation in one analysis. They in fact result in a clearer and better account of the research effort than do the piecemeal analyses of portions of data. These techniques yield more realistic probability statements

in hypothesis testing and interval estimation studies. Multivariate analysis (consequently the use of multivariate techniques) is specially important in behavioural sciences and applied researches for most of such studies involve problems in which several response variables are observed simultaneously. The common source of each individual observation generally results into dependence or correlation among the dimensions and it is this feature that distinguishes multivariate data and techniques from their univariate prototypes.

In spite of all this, multivariate techniques are expensive and involve laborious computations. As such their applications in the context of research studies have been accelerated only with the advent of high speed electronic computers since 1950's.

## Questions

1. What do you mean by multivariate techniques? Explain their significance in context of research studies.
2. Write a brief essay on "Factor analysis" particularly pointing out its merits and limitations.
3. Name the important multivariate techniques and explain the important characteristic of each one of such techniques.
4. Enumerate the steps involved in Thurstone's centroid method of factor analysis.
5. Write a short note on '*rotation*' in context of factor analysis.
6. Work out the first two centroid factors as well as first two principal components from the following correlation matrix, *R*, *relating* to six variables:

| | | *Variables* | | | | | |
|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 |
| | 1 | 1.00 | .55 | .43 | .32 | .28 | .36 |
| | 2 | | 1.00 | .50 | .25 | .31 | .32 |
| *Variables* | 3 | | | 1.00 | .39 | .25 | .33 |
| | 4 | | | | 1.00 | .43 | .49 |
| | 5 | | | | | 1.00 | .44 |
| | 6 | | | | | | 1.00 |

***Answers:***

| *Variables* | *Centroid factors* | | *Principal Components* | |
|---|---|---|---|---|
| | *I* | *II* | *I* | *II* |
| 1 | .71 | .40 | .71 | .39 |
| 2 | .70 | .46 | .71 | .48 |
| 3 | .70 | .37 | .70 | .32 |
| 4 | .69 | –.41 | .69 | –.42 |
| 5 | .65 | –.43 | .64 | –.45 |
| 6 | .71 | –.39 | .71 | –.38 |

7. Compute communality for each of the variable based on first two centroid factors in question six above and state what does it indicate.

8. Compute the proportion of total variance explained by the two factors worked out in question six above by the principal components method. Also point out the proportion of common variance explained by each of the two factors. ‘
9. What is the significance of using multiple discriminant analysis? Explain in brief the technical details involved in such a technique.
10. Write short notes on:
    (i) Cluster analysis;
    (ii) Multidimensional scaling;
    (iii) Reflections in context of factor analysis;
    (iv) Maximum likelihood method of factor analysis;
    (v) Path analysis.

# Appendix

# Summary Chart:
## Showing the Appropriateness of a Particular Multivariate Technique

| *Techniques of multivariate analysis* | *Number of* | | | |
|---|---|---|---|---|
| | *Explanatory variables* | | *Criterion variables* | |
| 1. Multiple regression analysis (along with path analysis) | | many | | one |
| 2. Multiple discriminant analysis | | many | one (to be classified into many groups) | |
| 3. Multivariate analysis of variance | many | | | many |
| 4. Canonical correlation analysis | | many | many[*1] | many[*2] |
| 5. Factor analysis | | many | | |
| 6. Cluster analysis | | many | | |
| 7. Multidimensional scaling (MDS) | many | many | | |
| 8. Latent structure analysis | many | | | |
| Nature of data | ↑ Non-metric | ↑ metric | ↑ Non-metric | ↑ metric |

[*1] Any one of the two.

[*2] Any one of the two.