

Multivariate Statistical Analyses and Applications

Prof. SK Mishra
Dept. of Economics
NEHU, Shillong

I. Introduction: Multivariate Statistical Analyses comprise a number of statistical methods/tools that may be applied to several fields of empirical investigations. Most of these methods have three characteristics in common: (i) they presume availability of data set, say $Z(n \times k)$ (in n observations and k variables; $k \geq 2$, $n > k$), collected from the field (observationally or experimentally); (ii) they analyze the dispersion matrix or correlation matrix obtained from Z , and (iii) they obtain a set of composite variables from Z by a weighted (often linear) combination. Among these methods six are the prominent ones, (i) Regression analysis, (ii) Principal Components analysis, (iii) Canonical Correlation Analysis, (iv) Discriminant Analysis, (v) Factor analysis, and (vi) Cluster analysis (see, *Kendall & Stuart, 1968*).

II. Regression Analysis: In regression analysis, Z is made up of two sets of variables, Y and X , that is, $Z = [Y \ X]$. The purpose of analysis is to explain variations in a particular Y , say $y_1 \in Y$, called as the dependent variable, in terms of other variables, $[Y_1 \ X_1]$, often called the explanatory variables, such that $Y_1 \in Y$; $y_1 \notin Y_1$; $X_1 \in X$.

Generally, a regression model is described as $YA + XB + U = 0$. In particular, when A is a 1×1 matrix normalized as $A = [-1]$, we call it a single equation model, described as $y = Xb + u$. It is often assumed that the disturbance term u obeys the Gauss-Markov assumptions - $E(u_i) = 0$; $E(uu') = \sigma^2 I$ and X are non-stochastic (*Intriligator, 1978*). Under these assumptions we estimate b such that $\hat{b} = [X'X]^{-1} X'y$. The quality of explanation is often judged by the value of $R^2 = r^2(y, \hat{y} : \hat{y} = X\hat{b})$. Here, R^2 is the coefficient of determination and r is the coefficient of multiple correlation.

However, the error term u often does not obey the Gauss-Markov assumptions. Whenever $E(uu') = \Omega$: $\Omega \neq \sigma^2 I$ we say that errors are non-spherical. To deal with such a situation we use Aitken's estimator (often call the Generalized Least squares). To use this method we have to estimate Ω from $\hat{u} = y - X\hat{b}$: $\hat{b} = [X'X]^{-1} X'y$ and incorporate the estimated Ω in the estimation method such that $\hat{b} = [X'\hat{\Omega}^{-1}X]^{-1} X'\hat{\Omega}^{-1}y$.

To deal with the problem of stochastic X , Instrumental variable method may be applied. The stochastic variables in X are replaced by their surrogates, called instrumental variables. Let W be the matrix corresponding to X in which the stochastic vectors of X have been replaced by the instrumental variables, then we estimate the coefficients b by $\beta = [W'X]^{-1} W'y$. An instrumental variable is chosen in such a manner that while it is

highly correlated with the variable it proxies, it is uncorrelated with the stochastic part of the variable that it proxies. As a result, β is a consistent estimator of b . However, applicability of this method is contingent on the availability of suitable instrumental variables.

Especially in the sciences that depend on observational data, errors often incorporate outliers. It has been found that in presence of outliers, the Least Squares based methods falter. Therefore, one may opt for least absolute deviation (LAD) estimators (for details see *Mishra & Dasgupta 2003*). These estimators minimize

$$S = \sum_{i=1}^n \left| y_i - \sum_{j=1}^m X_{ij} b_j \right|. \text{ It is possible to obtain this estimator by using the Fair-}$$

Schlossmacher method, which is very similar to the iterated weighted least squares method.

Multiple regression analysis often suffers from the problem of multi-collinearity, meaning highly correlated explanatory variables. Apparently, there is no straightforward solution to this problem, but one may try with the technique of Ridge Regression (*Judge, et al. 1980*), which yields biased estimators and yet does not always solve the problem. Of late, it has been suggested (*Paris, 2001*) that MEL estimator succeeds at solving the problem of severe multi-collinearity.

Multi-equation models (described as $YA + XB + U = 0$ where A is an $m \times m$ matrix and $m > 1$) are estimated by methods such as 2-Stage Least Squares, Limited Information Max Likelihood method, 3-Stage Least Squares, etc (see *Johnston 1991*).

III. Principal Components Analysis: This analysis attempts to replace $Z=X$ by a set of fewer (or equal in number) composite variables, say P , such that any P , say

$$p_i = \sum_{j=1}^k x_j a_{ji}; \quad i = 1, 2, \dots, k^*; \quad k^* \leq k. \text{ Here } x_j = (X_j - \bar{X}_j) / \sigma_{X_j}. \text{ Moreover, such derived}$$

composite variables are orthogonal among themselves or $P'P = I$. This is accomplished by using the eigenvectors of the inter-correlation matrix of X , say R , as weights vector a_i . More explicitly, let R be the inter-correlation matrix obtained from X ,

$\lambda = (\lambda_1, \lambda_2, \dots, \lambda_{k^*})$ be the eigenvalues of R and $a = (a_1, a_2, \dots, a_{k^*})$ be the eigen vectors

of R associated with λ . Then $p_i = \sum_{j=1}^k x_j a_{ji}$ is the i^{th} principal component of X . In applied

works, eigen vectors of R are normalized in such a manner that their squared (Euclidean) norm is equal to the eigenvalue associated with them. The eigenvalues of R are often used to indicate the percentage of variance in X explained by the principal components. Since eigenvalues are often obtained such that $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{k^*}$ the first principal component explains the largest part of the variance in X and so on.

To obtain the eigenvalues (and eigen vectors) of R , one may use power method or Jacobi's method. Power method would not be the viable method if any two eigenvalues are equal or more or less equal.

Principal components (especially the first) have often been used to construct composite indices in some applied works. This analysis also serves as a building block of Factor Analysis. If all the variables in X are measured in the same unit, one may use variance-covariance matrix of X to work out eigenvalues and eigen vectors. It is easier to interpret the derived variables in such cases. For application of Principal Component Analysis, one may refer to *Mishra & Ngullie, 2003a*.

IV. Canonical Correlation Analysis: If we have two variables (say, x and y, each in n observations) and we want to find out how they are associates, we often compute the coefficient of correlation $r(x,y)$, the cosine of the angle between two vectors, x and y. Now suppose, we have two sets of variables, $X=(X_1, X_2, \dots, X_p)$ and $Y=(Y_1, Y_2, \dots, Y_q)$. We want to find if these two sets are correlated. Could we somehow make two composite variables, say u and v, one by a linear combination of X and the other by a linear combination of Y and find $r(u, v)$, we could possibly correlate the sets of variables, X and Y. This is exactly done in the canonical correlation analysis.

First we regress all Y on all X and suppose we obtain the matrix of coefficients, say, A. Thus $A=[X'X]^{-1}X'Y$. Then we regress all X on all Y and obtain B, such that we have $B=[Y'Y]^{-1}Y'X$. Now we obtain, say, $C=AB=[X'X]^{-1}X'Y[Y'Y]^{-1}Y'X$. This is analogous to finding the product of regression coefficients of y on x and x on y. Next, we find out the eigenvalues and eigenvectors of C. Let the eigenvalues of C be $\lambda_1, \lambda_2, \dots, etc.$ and the eigenvectors be $a_1, a_2, \dots, etc.$ Now, λ_1 is the square of the largest correlation between the sets of variables X and Y. Similarly, λ_2 is the next largest squared correlation between X and Y and so on. The eigenvectors can be used to obtain the composite variables representing X and Y. For an application, reference may be made to *Mishra & Ngullie, 2003a*.

V. Discriminant Analysis: Discriminant Analysis is a (multi-variate) statistical method to decide whether two (or more) samples drawn from apparently different parents (with different mean values but identical dispersion) can be statistically distinguished from each other. The technique of Discriminant Analysis runs in five steps. These are:

(1). **Finding the vectors of sample means and differences:** Define Sample Mean (M_1 and M_2), Pooled mean (M) and the vector of differences between means ($D=M_1 - M_2$).

(2). **Finding Pooled Variance-Covariance matrix :** Using Pooled Mean (M) we find out the variance-Covariance Matrix (V) of the Pooled sample.

(3). **Solution for weights (W) such that $W = V^{-1} D$.** We invert the variance covariance matrix (V) to obtain V^{-1} .

(4). **Computation of δ_1, δ_2 and δ .** We define: $\delta_1 = M_1W$; $\delta_2 = M_2W$ and $\delta = MW$

(v) **Finding discriminant scores, $d = XW$.**

The discriminant function δ divides [d] score into two sets, one with $d \leq \delta$ and the other with $d > \delta$. If the value of the score (d) for an observation belonging to set #1 is conformal to its original membership, the discriminant function is powerful for the purpose. However, if there are many defections (members originally belonging to one set is misclassified as belonging to the other set), the discriminant function is not reliable.

An illustrative Application: Suppose we have collected time series data on three variables from two different states in the North-Eastern Region of India. We assume that the data have identical dispersion matrix, but different mean values.

Yield Rates of Rice, Pulses and Oilseeds in Assam													
Year	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998
Rice	1155	1043	1163	1060	1149	1313	1265	1308	1331	1350	1354	1336	1359
Pulses	471	419	461	420	451	428	456	468	522	546	534	572	547
Oilseed	500	358	499	446	530	530	578	476	474	530	509	512	549

Yield Rates of Rice, Pulses and Oilseeds in Nagaland												
Year	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997
Rice	1076	693	702	1048	1177	1227	1194	1304	1343	1229	1321	1093
Pulses	750	611	552	901	1033	1159	1071	824	909	752	1083	983
Oilseed	800	769	574	820	891	880	860	810	840	824	912	817

For Assam we have 13 observations and for Nagaland we have 12 observations. Can we discriminate between Assam and Nagaland on the basis of yield rates pertaining to the three major crops? We proceed to discriminant analysis as follows:

(i). **Finding the vectors of sample means and differences:** Define Sample Mean (M_1 and M_2), Pooled mean (M) and the vector of differences between means ($D=M_1 - M_2$).

Crops	Mean Yield Rates			D = Difference between Means
	Assam (M_1)	Nagaland (M_2)	Pooled (M)	$M_1 - M_2$
Rice	1245.077	1117.25	1183.72	127.8269
Pulses	484.2308	885.6667	676.92	-401.436
Oilseeds	499.3077	816.4167	651.52	-317.109

(ii). **Finding Pooled Variance-Covariance matrix :** Using Pooled Mean (M) we find out the variance-Covariance Matrix (V) of the Pooled sample.

**Variance-Covariance Matrix (V)
of Yield Rates (Assam & Nagaland Pooled Together)**

	Rice	Pulses	Oilseeds
Rice	31298.12	2613.498	-1839.13
Pulses	2613.498	57847.03	38381.6
Oilseeds	-1839.13	38381.6	29832.49

(iii). **Solution for weights (W) such that $W = V^{-1} D$.** We invert the variance covariance matrix (V) to obtain V^{-1} . The inverted variance-covariance matrix is as follows:

**Inverted Variance-Covariance Matrix (V)
of Yield Rates (Assam & Nagaland Pooled Together)**

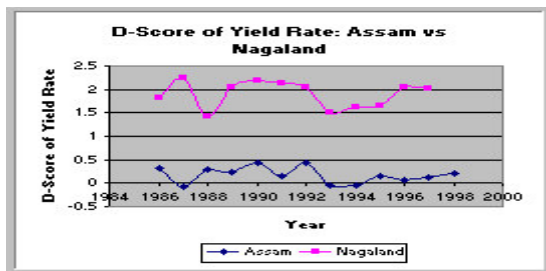
	Rice	Pulses	Oilseeds
Rice	35.39089	-20.8158	28.96275
Pulses	-20.8158	130.357	-168.997
Oilseeds	28.96275	-168.997	252.7321

When V^{-1} is multiplied by D we get weights (W) = [3695.759 -1400.434 -8600.047]’.

(iv). **Computation of δ_1 , δ_2 and δ .** We define: $\delta_1 = M_1W$; $\delta_2 = M_2W$ and $\delta = MW$.
From the data we have $\delta_1 = -370698.3$; $\delta_2 = -4132453$ and $\delta = -2176340$.

(v) **Finding discriminant scores, $d = XW$.**

Discriminant Scores (d) ; $\delta = -2176340$; *Relative Score = d/δ				
Year	Assam	Relative* Score	Nagaland	Relative* Score
1986	-691040	0.317524	-3953748	1.816696
1987	189068.3	-0.08687	-4907961	2.255144
1988	-638869	0.293552	-3115059	1.431329
1989	-506311	0.232643	-4440696	2.040442
1990	-943208	0.433392	-4759406	2.186885
1991	-304893	0.140095	-4656472	2.139588
1992	-934306	0.429301	-4483192	2.059969
1993	85014.44	-0.03906	-3300748	1.516651
1994	111593.6	-0.05128	-3533652	1.623667
1995	-333402	0.153194	-3597500	1.653004
1996	-121212	0.055695	-4477840	2.057509
1997	-266752	0.122569	-4363423	2.004936
1998	-464942	0.213635	-	-
<i>Mean Score</i>	<i>-370712</i>	<i>0.170337</i>	<i>-4132475</i>	<i>1.898819</i>
δ value (δ_1 and δ_2)	-370698.3		-4132453	



In this example, there is not a single case of misclassification. All observations belonging to Assam score larger than δ (= -2176340) and all observations belonging to Nagaland score less than δ . Relative score make the picture clearer, though one should be careful to note that the relative scores are

normalized with respect to δ . Only the values have changed their signs. That does not, however, affect discrimination. There is no year for Assam scoring > 0.44 (relative score) There is no year scoring < 1.43 in Nagaland. We conclude therefore that on the criteria the three yield rates (or rice, pulses and oilseeds) Assam can be discriminated from Nagaland.

VI. Factor Analysis: Factor Analysis is a multi-variate statistical method with a major objective of data reduction. It assumes that a few latent, fundamental and essential forces work in the hind side and manifest themselves into the empirically observed multivariate data on a complex of variables. These essential, but latent or unobservable forces are “factors”. Thus, the relationship between these latent factors and the observed data on the variables is that of “the essence” and “the manifestations”. Nevertheless, each individual variable that constitutes of the complex of empirical data has ‘noise’, specificity or errors of its own. Using suitable statistical and mathematical methods it is possible to extract these essential factors and in turn, they can be given a conceptual and theoretical meaning. Thus, Factor Analysis is a statistical method to extract the common, essential and latent variables that reflect themselves into a complex of empirically observed variables. These essential factors are very often much fewer in number than are their manifest variables. In this sense, factor analysis is a statistical method of reducing the dimensionality of data.

The method of Factor Analysis may take a number of different forms based on the following assumptions and techniques used. These different approaches to Factor analysis may conveniently be classified according to the following criteria.

1. Assumption regarding the relationship between the universe of variables and the sample variables drawn from the said universe: Latent factors may manifest themselves into innumerable many measurable variables, but in any particular empirical study, only a subset of the said universe of variables can be included in the analysis. That is to say that the sample over the variables is always a proper subset of the universe of variables.
2. Assumption regarding the relationship between the universe of individuals and the sample individuals drawn from the said universe: Empirical data on various variables (manifestations) are collected for a sample of individuals to whom the schedules are administered.
3. Assumption regarding the composition of variables in terms its explicability by other variables (complimentary set) in the sample: It may be assumed that every variable included in the analysis can be (partly) explained by the complementary set of variables included in the analysis.
4. Inter-relationship among the latent factors: It may be assumed that the latent factors (essence) are orthogonal among themselves or, alternatively, they are inter-correlated.
5. Optimization criterion: Factor Analysis has to depend on some optimization criterion, whether in terms of the minimization of some type of distance between the empirical data points and the inferred (estimated) points corresponding to

them or maximization of the probability of occurrence of the data points. In measuring the distance different criteria might be used.

6. Identifiability of factors and Reproducibility of the empirical inter-correlation matrix: On the one hand identifiability demands that the number of factors be as small as possible (parsimony) and such that each factor can be given some conceptual or theoretical meaning. On the other hand it is also required that the original empirical inter-correlation matrix among the variables is reproducible in terms of the factor loadings. These two requirements often compete with each other.

Methods to extract Initial Factors from the Observed Correlation Matrix: The complex of the criteria mentioned above gives rise to several strategies of factor analysis to extract initial factors from the observed correlation matrix. The general method of extraction may be expressed symbolically as the solution of the eigen-equation (characteristic equation) in which $\text{Det}(R_x - \lambda I) = 0$ is solved for λ and from $R_x V = \lambda V$ the vector V is obtained. Then V is used to obtain initial factor loadings. Here R is the inter-correlation matrix obtained from the variables included in the analysis. The real difference in the methods of extraction is in obtaining R_x from R . Obtaining R_x from R depends on two major ideas, first, the communality, generality, image or similar to these and the second the individuality, anti-image, uniqueness, residual or the like. The correlation matrix is decomposed into two parts, the common part and the uncommon part. The real issue is as to how one does the decomposition. Accordingly, we will see how R_x is defined differently in different methods of extraction.

1. Alpha Factoring: In this method it is assumed that the variables included in the analysis are only the samples from a universe of innumerable variables that the latent factors may manifest themselves into. Accordingly, the initial factors are extracted and later necessary rotations are applied to them. In Alpha factoring, the key emphasis is not on the statistical inference but its objective is to draw inference for an understanding of the underlying essence that give rise to the manifestation as empirically observed. In Alpha factoring $R_x = H^{-1}(R-U^2)H^{-1}$, where U is the diagonal matrix made up of the elements that are square root of the unique component for each variable (unexplained by other variables and so uncommon or unique to the variable) and H is the matrix of the square root of communalities.

2. Image factoring: Image analysis decomposes the variations in the variables into two parts, (i) image, and (ii) anti-image. That part of the variation of a variable which is predictable by a linear combination of all the other variables in the set (complimentary set) is called the image of the variable. It is the common part. On the other hand, the anti-image of the variable is the residual which is the unique part of the variable. Here $R_x = (R-S^2)R^{-1}(R-S^2)$, where S is the diagonal matrix whose elements are the standard deviation of the residuals of each variable that could not be explained by other variables. This S is the matrix of anti-image standard deviations or $s_j = \text{sqrt}(\Sigma(e_{ij})^2/n)$ for variable j .

3. Principal Axis factoring: This is one of the oldest methods of factor analysis to extract initial factors. An attempt is made to find out as many mutually orthogonal principal axes as needed. Each principal axis is obtained in such a manner that it minimizes the sum of distance between the observed points and the estimated points on the principal axis (distance measured by the length of the line **normal** to the principal axis and joining the observed point away from the principal axis and the expected point on the principal axis). This method uses the decomposition strategies of the Principal Components analysis as applied to the adjusted correlation matrix whose diagonal elements are replaced by corresponding estimates of communalities. These communalities are usually estimated by the highest absolute correlation in the corresponding row of a correlation matrix. In our general system here $R_x = R-h$, where h is a diagonal matrix of communalities. This method is gradually being replaced by the **Least Squares factoring**. In the Least Squares factoring, an attempt is made to minimize the residual correlation after extracting a given number of factors, and to assess the degree of fit between the reproduced correlation under the model and the observed correlations. Since the squared differences are minimized, the name follows. In the LS factoring the communalities are estimated at each iteration (not once for all as in the Principal Axis method) and a new R_x is found until the results are stable.

4. Maximum Likelihood procedure for initial factoring: It is assumed that the observed data comprise a sample from a population where a k -common factor model exactly applies, and where the distribution of variables and the factors is a multivariate normal. The exact loadings on each variable are unknown and to be estimated. The objective of ML method is to find the underlying population parameters that would have the greatest likelihood of producing the observed correlation matrix. In our general framework, here $R_x = U^{-1}(R-U^2)U^{-1}$, where U is the square root of the unique variance estimated at each stage of iteration. The initial iteration begins with the principal component analysis. In contrast with the Least Squares method where the sum of squared discrepancies (between initial and reproduced correlations) is minimized, in case of maximum likelihood method the likelihood of reproducing the observed (that is, initial) correlations by the estimated loadings is maximized. Further, in the LS method, assumption of normality of distribution is not necessary.

5. Principal Component analysis for initial extraction of factors: In terms of eigenvalues and eigenvectors, an observed or initial correlation matrix, R , can always be decomposed as $R = R_1 + R_2 + \dots + R_m$, where R_i is a cross-product of the i^{th} eigenvector (standardized) multiplied by the i^{th} eigenvalue. Conventionally eigenvalues are ordered according to magnitude and the first eigenvalue is the largest while the last is the smallest. Accordingly, R_1 reproduces larger part of R than do the subsequent R_s and that also in the order. Here we have $R_x = R$. Interestingly, the Principal Component method does not try to modify the matrix of inter-correlations in view of communalities, uniqueness, image, etc. In this sense, this method of initial factorization is quite different from others. However, due to its great power in decomposition of R matrix into orthogonal component matrices and finding out

eigenvalues, it is often used at the initial level of many methods of extraction. As a matter of fact, the famous theorem of Cayley-Hamilton (that every continuous function of a matrix is a function of its eigenvalues and eigenvectors) makes the Principal Component method so powerful.

The initial factoring step usually determines the minimum number of factors that can adequately account for the observed correlations, and provides the estimates of the communalities for each variable. With these factors (say m in number) and estimated communalities, it is possible to carry out some transformation over the factors (by rotation performed on them in the m -plane) so that simpler and more easily interpretable factors can be obtained. The output of such an attempt is the rotated factors.

Rotation of Factors for better interpretation: To proceed for rotation, one must, first assert whether the factors would be correlated (non-orthogonal or oblique) or they are orthogonal (spherical) among themselves. If one visualizes that the factors need not be orthogonal, one may proceed to oblique rotation (OBLIMIN or PROMAX). In cases when there is a good reason to hold that factors are orthogonal, one may try with methods like VARIMAX, QUARTIMAX or EQUAMAX rotation. It is fair not to impose orthogonality on the factors from the very beginning and look into the pattern obtained after carrying out the oblique rotation. If the factors are orthogonal, such an evidence will be available in the results of the oblique rotation. Alternatively, rotation can be carried out to obtain a pattern that is close to a given (target) matrix.

It is clear that an oblique solution is more general than an orthogonal rotation because it does not impose the restriction of orthogonality on the factors. Its supremacy over the orthogonal rotation methods lies in the fact that after carrying out the oblique rotation one may get the feel whether imposition of orthogonality relations on the factors from the very beginning could have been appropriate or not.

1. The Quartimax Rotation: Let us define $\mathbf{q}_i = [\sum_j (b_{ij})^4 - \{\sum_j (b_{ij})^2\}^2]/(m^2)$; where $j = 1, m$; $i = 1, n$. Further, let \mathbf{q} be defined as the sum of \mathbf{q}_i . That is $\mathbf{q} = \sum_i (\mathbf{q}_i)$; $i = 1, n$. Here m is the number of initial factors and n is the number of observations in the data set. Since communalities $(h_i)^2 = \sum_j (b_{ij})^2$ are already fixed and $m^2 = \text{constant}$, \mathbf{q} varies directly with $\sum_i \sum_j (b_{ij})^4$. Now axes are rotated in such a way that \mathbf{q} is maximized and in that process old (initial) factor loadings are replaced by new factor loadings. This is called the Quartimax rotation. An application of quartimax criterion usually results in emphasizing the simplicity of interpretation of the variables at the expense of simplicity of interpretation of factors. In general, fewer common factors add to simplicity in interpretation of variables while fewer variables with large loadings on each factor facilitate the interpretation of the factors and their identification. The Rotated factors obtained by this method are orthogonal among themselves.

2. The Varimax Rotation: Let us define $\mathbf{v}_j = [n\sum_i (b_{ij})^4 - \{\sum_i (b_{ij})^2\}^2]/(n^2)$; where $j = 1, m$; $i = 1, n$. Further, let \mathbf{v} be defined as the sum of \mathbf{v}_j . That is $\mathbf{v} = \sum_j (\mathbf{v}_j)$; $j = 1, m$. Here m is the number of initial factors and n is the number of observations in the data set. Now, unlike in case of quartimax rotation, $\sum_j (b_{ij})^2$ are no longer fixed but they are

variable due to summation being carried out over n individuals, the maximization of V is called the Varimax method of rotation. This method of rotation concentrates on simplifying the interpretation of factors rather than the variables as it was the case with the quartimax rotation method. The Rotated factors obtained by Varimax method are orthogonal among themselves.

3. The Equimax and the Biquartimax Rotation: A hybridization of quartimax and varimax rotation methods yields these methods. If one maximizes $\zeta = \alpha\mathbf{q} + \beta\mathbf{v}$, the compromise solution is obtained. Defining $\gamma = \beta/(\alpha+\beta)$, in special or limiting cases $\gamma = 0$ yields quartimax solution and $\gamma = 1$ yields varimax solution. In particular, maximization for $\gamma = m/2$ is called the Equimax solution and that for $\gamma = 0.5$ is called the biquartimax solution. The Rotated factors obtained by this method are orthogonal among themselves.

4. Indirect Oblimin Rotation: This method of rotation tries to simplify loadings on Reference Axes. The indirect Oblimin criterion is given by:

$$B = \sum_j [\sum_k \{n \sum_i a_{ij}^2 a_{ik}^2 - \gamma (\sum_i a_{ij}^2 \sum_i a_{ik}^2)\}]; k = 2, m; j = 1, k.$$

Iteratively, B is minimized. In the expression above, \mathbf{a}_{ij} are projections of i^{th} factor on j^{th} reference axis usually normalized by h_i^2 (communality) and γ refers to the degree of obliqueness, which can be altered to obtain more or less oblique solution. For $\gamma = 0$ this rotation is called **Quartimin** solution, while for $\gamma = 1$ it is called **Covarimin** solution. For $\gamma = 0.5$ it is called **Biquartimin** solution.

5. Direct Oblimin Rotation: This method of rotation tries to simplify loadings on primary factors. The direct Oblimin criterion is given by:

$$D = \sum_j [\sum_k \{ \sum_i b_{ij}^2 b_{ik}^2 - \delta (\sum_i b_{ij}^2 \sum_i b_{ik}^2) / n \}]; j, k = 1, m.$$

Iteratively, D is minimized. In the expression above, \mathbf{b}_{ij} are factor loadings in a pattern matrix and δ refers to the degree of obliqueness, which can be altered to obtain more or less oblique solution. For a unifactoral factor pattern, $\delta = 0$ identifies the correct pattern.

6. Promax Rotation: Promax rotation is a variant of Target matrix rotation, though in this case such a matrix is derived from the data itself and the analyst is not needed to supply any target matrix. The rationale behind the promax rotation is the observed fact that in practice the orthogonal solutions are not much different from oblique solutions. Therefore, if small loadings are reduced to near-zero loadings (that might be ignored), it is possible to construct a fairly good target matrix with much simpler structure. Having done that, one rotates the factors to be close to such an artificial target matrix.

The Choice of method: In practice, it is not at all easy to choose among the various methods of extraction of factors or rotation of factors. The first difficulty is regarding the decision as to how many factors are to be extracted. It is usually suggested that one should extract as many factors as the number of eigenvalues of the inter-correlation matrix exceeding unity. But when the number of variables included in the analysis is large (say over 100) and these variables are not strongly correlated with each other, this

criterion may suggest extraction of too many factors to be explicable or meaningful. The Principle of Parsimony of Factors is seriously violated in such a case. It is to be noted that the data origination from a relatively less developed population has only a poor communality. In case of such data the said criterion is utterly disappointing. Nevertheless, the irony is that extraction of factors and thereafter their rotation and identification (interpretation) depends greatly on the choice of the number of factors to be extracted.

Scree Test: It is a diagrammatical presentation of eigenvalues (of inter-correlation matrix) with the magnitude of the eigenvalue on the vertical axis and the serial number of the eigenvalue (integers of the index set) on the horizontal axis. This curve is usually (not always) very steep in the beginning but quickly tapers off. It clearly has an ‘elbow’ before which the slope is steep (negative) and after which it is poorly falling. It is a good heuristic to extract as many factors as suggested by this elbow projected on the horizontal axis. Nevertheless, one must not forget that factors so extracted have also to be interpreted and given some reasonable meaning within the available theoretical structure. One must make a number of trials and errors. Different methods of extraction and rotation must be experimented with.

Having decided tentatively as to how many factors are to be extracted, one has to choose the method of extraction. The most intriguing fact is that in most cases there is not much difference between the results obtained by one method of extraction and the others. It is more so when the number of individuals as well as number of variables in the data is large. In applying different methods of extraction of factors, one observes that factor loadings exhibit some alteration and their order is changed. The factors so extracted sometimes change their order (what is the first factor in one method of extraction turns out to be the second or the third factor in another method of extraction). New factors, however, do not emerge very often. In an explorative research, there is no much reason to hold as to the order of importance or strength of the underlying factors. When one method suggests one order and another method suggests another order, the researcher is faced with a real instance of indecision. Nevertheless, an attempt to derive a fixed (say, m) number of factors by different methods of extraction is always rewarding. It is true that in one method the list of variables loading on a particular factor is not exactly the same as in case of another method of extraction. Nevertheless, it is a general observation that these lists contain quite many common variables. Only a few are uncommon. Usually, they suggest adequately for proceeding to an attempt to giving some name (identification) and theoretical justification (interpretation) to those clusters of variables.

The next step is to go in for rotation. It is fair not to impose orthogonality on the factors in the very beginning. So, one may start with the ‘Direct Oblimin’ or ‘Promax’ solution. If factors are correlated, the solution would suggest that. In case there are no such symptoms observed, one may impose orthogonality conditions on the factors and try with such methods as ‘Quartimax’, ‘Varimax’ or ‘Equamax’. Alpha factoring may go well with ‘Quartimax’ and ‘Equamax’ solutions. Principal Component and Max Likelihood methods may go well with ‘Varimax’ rotation.

Finally, nothing can supersede judgment, interpretability and theoretical justifiability. To suppress small loadings and carefully deal with the bi-polar factors may be helpful in clear conceptualization. For application, reference may be made to *Mishra & Ngullie, 2003-b*.

VII. Cluster Analysis: The term cluster analysis actually encompasses a number of different classification algorithms. A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop *taxonomies*. There are three approaches to cluster analysis: (a) *joining (tree clustering)*, (b) *two-way joining (block clustering)*, and (c) *k-means clustering*.

1. Joining (Tree Clustering): The purpose of this method or approach is to join together objects into successively larger clusters, using some measure of similarity or **distance**. A typical result of this type of clustering is the hierarchical tree.

Hierarchical Tree: We begin with each object in a class by itself. Now imagine that, in very small steps, we "relax" our criterion as to what *is* and *is not* unique. Put another way, we lower our threshold regarding the decision when to declare two or more objects to be members of the same cluster. As a result we link more and more objects together and aggregate (amalgamate) larger and larger clusters of increasingly dissimilar elements. Finally, in the last step, all objects are joined together. When the data contain a clear "structure" in terms of clusters of objects that are similar to each other, then this structure will often be reflected in the hierarchical tree as distinct branches. As the result of a successful analysis with the joining method, one is able to detect clusters (branches) and interpret those branches.

(i) **Distance Measures:** The joining or tree clustering method uses the dissimilarities or distances between objects when forming the clusters. These distances can be based on a single dimension or multiple dimensions. The most straightforward way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances. If we had a two- or three-dimensional space this measure is the actual geometric distance between objects in the space (i.e., as if measured with a ruler). However, the joining algorithm does not "care" whether the distances that we use are actual real distances, or some other derived measure of distance that is more meaningful to us; and it is up to us to select the right method for our specific application.

a) *Euclidean distance.* This is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. It is computed as:

$$\text{distance}(x,y) = \{ \sum_{i} (x_i - y_i)^2 \}^{0.5}$$

b) *Squared Euclidean distance.* One may want to square the standard Euclidean distance in order to place progressively greater weight on objects that are further apart. This distance is computed as:

$$\text{distance}(x,y) = \sum_{i} (x_i - y_i)^2$$

c) *City-block (Manhattan or Absolute) distance.* This distance is simply the average difference across dimensions. In most cases, this distance measure yields results similar to the simple Euclidean distance. However, note that in this measure, the effect of single large differences (outliers) is dampened (since they are not squared). The city-block distance is computed as:

$$\text{distance}(x,y) = \sum_i |x_i - y_i|$$

d) *Chebychev distance.* This distance measure may be appropriate in cases when we want to define two objects as "different" if they are different on any one of the dimensions. The Chebychev distance is computed as:

$$\text{distance}(x,y) = \text{Max} (|x_i - y_i|)$$

e) *Power distance or Minkowski distance.* Sometimes we may want to increase or decrease the progressive weight that is placed on dimensions on which the respective objects are very different. This can be accomplished via the power distance. The power distance is computed as:

$$\text{distance}(x,y) = (\sum_i |x_i - y_i|^p)^{1/r}$$

where r and p are user-defined parameters.

f) *Percent disagreement.* This measure is particularly useful if the data for the dimensions included in the analysis are categorical in nature. This distance is computed as:

$$\text{distance}(x,y) = (\text{Number of } x_i \neq y_i)/i$$

(ii) **Amalgamation or Linkage Rules:** At the first step, when each object represents its own cluster, the distances between those objects are defined by the chosen distance measure. However, once several objects have been linked together, how do we determine the distances between those new clusters? In other words, we need a linkage or amalgamation rule to determine when two clusters are sufficiently similar to be linked together. There are various possibilities: for example, we could link two clusters together when any two objects in the two clusters are closer together than the respective linkage distance. Put another way, we use the "nearest neighbors" across clusters to determine the distances between clusters; this method is called single linkage. This rule produces "stringy" types of clusters, that is, clusters "chained together" by only single objects that happen to be close together. Alternatively, we may use the neighbors across clusters that are furthest away from each other; this method is called complete linkage. There are numerous other linkage rules such as these that have been detailed below.

a) *Single linkage (nearest neighbor).* In this method the distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the

different clusters. This rule will, in a sense, string objects together to form clusters, and the resulting clusters tend to represent long "chains."

b) Complete linkage (furthest neighbor). In this method, the distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors"). This method usually performs quite well in cases when the objects actually form naturally distinct "clumps." If the clusters tend to be somehow elongated or of a "chain" type nature, then this method is inappropriate.

c) Unweighted pair-group average. In this method, the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters. This method is also very efficient when the objects form natural distinct "clumps," however, it performs equally well with elongated, "chain" type clusters.

d) Weighted pair-group average. This method is identical to the unweighted pair-group average method, except that in the computations, the size of the respective clusters (i.e., the number of objects contained in them) is used as a weight. Thus, this method (rather than the previous method) should be used when the cluster sizes are suspected to be greatly uneven

e) Unweighted pair-group centroid. The centroid of a cluster is the average point in the multidimensional space defined by the dimensions. In a sense, it is the center of gravity for the respective cluster. In this method, the distance between two clusters is determined as the difference between centroids.

f) Weighted pair-group centroid (median). This method is identical to the previous one, except that weighting is introduced into the computations to take into consideration differences in cluster sizes (i.e., the number of objects contained in them). Thus, when there are (or one suspects there to be) considerable differences in cluster sizes, this method is preferable to the previous one.

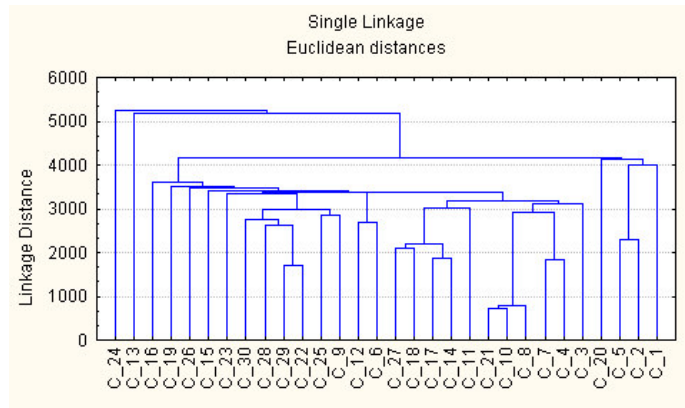
g) Ward's method. This method is distinct from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step. In general, this method is regarded as very efficient, however, it tends to create clusters of small size.

2. K-means Clustering: This method of clustering is very different from the Joining (Tree Clustering) and Two-way Joining methods. Suppose that we already have hypotheses concerning the number of clusters in our cases or variables. We may want to form exactly 3 clusters that are to be as distinct as possible. This is the type of research question that can be addressed by the k-means clustering algorithm. In general, the k-means method will produce exactly k different clusters of greatest possible distinction.

3. Two-way Joining :Previously, we have discussed this method in terms of "objects" that are to be clustered. In all other types of analyses described above the research question of interest is usually expressed in terms of cases (observations) or variables. It turns out that the clustering of both may yield useful results. For example, imagine a study where we have gathered data on different measures of physical fitness (variables) for a sample of heart patients (cases). We may want to cluster cases (patients) to detect clusters of patients with similar syndromes. At the same time, we may want to cluster variables (fitness measures) to detect clusters of measures that appear to tap similar physical abilities. In the Cluster Analysis, we can choose to cluster cases as well as variables.

Given the discussion in the paragraph above concerning whether to cluster cases or variables, one may wonder why not to cluster both simultaneously? Two-way joining is useful in (the relatively rare) circumstances when one expects that both cases and variables will simultaneously contribute to the uncovering of meaningful patterns of clusters. The difficulty with interpreting these results may arise from the fact that the similarities between different clusters may pertain to (or be caused by) somewhat different subsets of variables. Thus, the resulting structure (clusters) is by nature not homogeneous. This may seem a bit confusing at first, and, indeed, compared to the other clustering methods described (Joining or Tree Clustering and K-means Clustering), two-way joining is probably the one least commonly used. However, at many instances this method offers a powerful exploratory data analysis tool.

An Illustration of the Application of Cluster Analysis - Economics of Private Schooling Industry in Kohima, Nagaland: As an example to illustrate application of Cluster Analysis to applied economics, we present here the case of Private Schooling Industry as it operates in Kohima, Nagaland. There are 30 private schools. They charge different admission fees to the students of different standards. Their salary structures also differ.



Here we denote: X_1 , X_2 and X_3 as admission fees (annual) charges from the students of Lower, Secondary and Higher secondary standards respectively. X_4 stands for annual fees for computer exposure programme. Salaries (monthly) paid to Matriculate, Graduate, PG, Theology and Computer teachers are denoted by X_5 , X_6 , X_7 , X_8 and X_9

respectively. Salaries (per month) paid to Matriculate, Graduate and Grade IV non-teaching staff are denoted by X_{10} , X_{11} and X_{12} respectively.

Fees and Salaries Structure in Private Schooling Industry in Kohima												
Sl No.	Annual Fees Charged				Salaries of Teachers					Salaries of Staff		
	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	X ₈	X ₉	X ₁₀	X ₁₁	X ₁₂
1	1000	1000	3500	200	3500	4300	3500	4300	4300	2150	3400	2150
2	1000	1000	6000	150	3000	4500	5500	4500	4500	2900	0	2700
3	1300	1500	0	0	3500	3700	0	0	0	2900	0	1000
4	1100	1100	0	0	2750	3000	0	0	0	0	0	1000
5	1500	1500	4800	360	3200	4000	4800	4000	4000	1800	0	1800
6	515	530	0	420	2700	3000	3600	0	3600	2500	3000	0
7	750	750	0	0	1200	2200	0	0	0	0	0	700
8	1100	1210	0	0	2300	3000	0	3000	0	0	0	2100
9	1000	1000	0	350	3000	4500	5300	4500	4500	2900	0	2700
10	580	580	0	0	2500	3000	0	2800	0	0	0	1300
11	1200	1440	0	0	3000	3800	4000	0	0	2500	2800	1800
12	1250	1400	0	390	2800	3200	3500	0	3200	2500	3000	2400
13	2500	2800	0	150	4810	4570	0	0	4570	2625	0	1500
14	1020	1250	0	0	2400	3000	3100	0	0	0	0	1600
15	1300	1550	0	350	2600	3600	3635	3600	3600	3200	3200	1000
16	1425	1875	0	0	3450	3750	3930	3550	0	3400	0	2650
17	600	600	0	0	2050	3000	3500	0	0	0	0	0
18	1055	1300	0	480	1700	2200	2600	0	0	1800	0	1500
19	1200	1355	0	360	0	3863	4127	4500	3662	3259	0	1000
20	2450	3340	5800	480	2800	4300	6600	4700	4500	5618	0	3450
21	680	680	0	0	2300	3300	0	2800	0	0	0	1900
22	1850	3000	0	600	2500	3500	4000	3000	4000	0	0	2000
23	1100	1300	0	360	2000	3000	3800	0	3500	2000	0	1250
24	2075	3000	4500	250	3100	3800	4800	0	3800	1700	2800	800
25	1100	1450	0	200	3000	3500	3800	3000	3300	2800	0	1700
26	900	900	0	0	2572	3587	3638	3587	0	0	0	1740
27	800	800	0	0	2500	3500	3800	0	0	2000	0	1800
28	1750	1870	0	360	0	3500	4000	3500	3500	0	0	1200
29	1600	1750	0	360	2400	3000	3500	3200	3200	0	0	1800
30	1200	1350	0	360	2640	3580	3650	3580	3580	0	2540	2100

We want to know if the private schools (30 in number) have some similarity in matters of fees and salaries structure. Hence, we use cluster analysis. We choose Tree-Clustering approach with single linkage. The measure of distance used is Euclidean.

The dendrogram (a graphical presentation of tree-clustering) indicates that schools # 24, #13, # 20, #5, #2 and #1 are quite at distance from other schools, that make a closer cluster. For further details, reference may be made to *Mishra & Rio, 2004*.

References:

Intriligator, MD (1978). *Econometric Models, Techniques and Applications*, Prentice Hall, Inc. Englewood Cliffs, New Jersey.

Johnston, J (1991). *Econometric Methods*. McGraw Hill, New York.

Judge, G.G., W.E. Griffiths, R.C. Hill, T.C. Lee (1980) *The Theory and Practice of Econometrics*. John Wiley, New York.

Kendall, M G & A Stuart (1968). *The Advanced Theory of Statistics*. Charles Griffith, London.

Mishra, SK & M Dasgupta (2003). "Least Absolute Deviation Estimation of Multi-equation Linear Econometric Models : A Study based on Monte Carlo Experiments"
<http://www.skmishra.owns1.com>

Mishra, SK & M Ngullie (2003a). "Quality of Life in Dimapur, Nagaland (India)"
<http://www.skmishra.owns1.com>

Mishra, SK & M Ngullie (2003b). "Determinants of Quality of Life in Dimapur Nagaland (India)" <http://ssrn.com/author=353253>

Mishra, SK & K Rio (2004) "Economics of Private Schooling Industry in Kohima, Nagaland (India)" <http://ssrn.com/author=353253>

Paris, Q (2001) "Multicollinearity and maximum entropy estimators." *Economics Bulletin*, 3(11), pp. 1-9. At URL <http://www.economicsbulletin.com/2001/volume3/EB-01C20002A.pdf>