The Philippine Statistician
1992, Vol. 41,
Pages 69-81

FACTOR ANALYSIS: AN ANALYSIS OF VARIABLE INTERDEPENDENCE

by

Violeta C. Menil
De La Salle University

Abstract

Factor analysis was used to successfuly extract 3 factors that determine consumers' choice of car. Practical issues related to the use of Factor Analysis were satisfactorily resolved in this application.

1. INTRODUCTION

Factor Analysis is a statistical tool which reduces the number of variables under consideration to a more manageable number. Through the factor analytic technique, the number of variables for further research can be minimized while at the same time maximizing the amount of information in the analysis.

The essence of factor analysis is shown in Figure 1.1. Here there are fourteen (14) variables V1,V2,...,V14 which "load" on four unobservable common factors. The variables V3, V7, V10 and V14 are grouped together, meaning that they are highly correlated with one another and constitute the first factor. Similarly, variables V2, V5 and V13 define a second separate factor; variables V4 and V12 define a third factor and variables V1, V6, V8, V9 and V11 contribute to form the fourth factor. Therefore, each subset of variables can be thought of as reflecting a latent underlying dimension. So, instead of having to deal with the fourteen variables separately, we now need to consider only the four factors as defined in the figure.

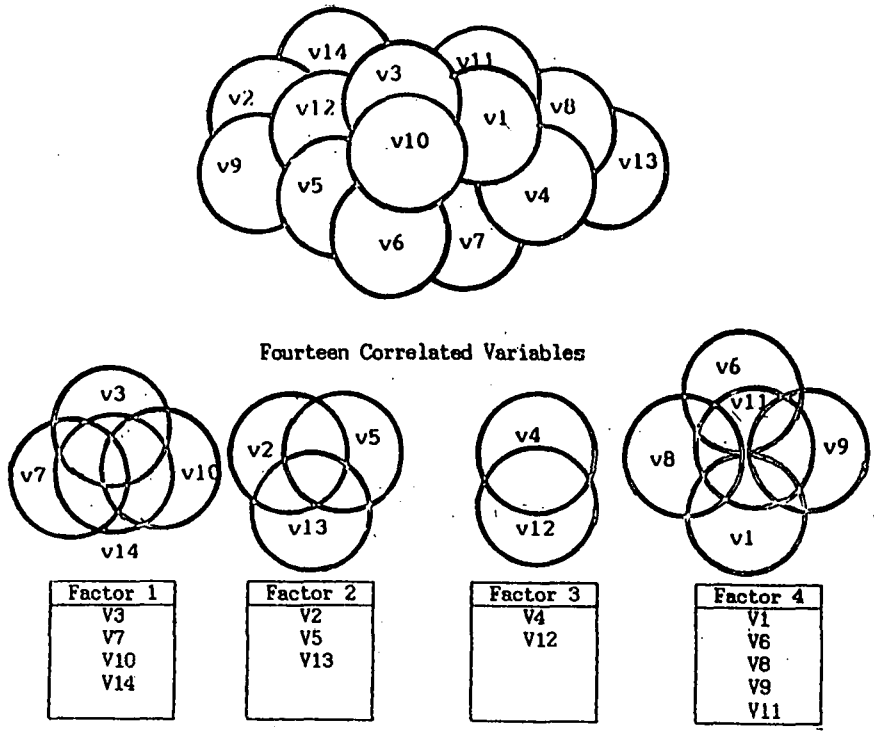

Fig. 1.1: Fourteen variables reduced to four factors

The Basic Model

The basic common factor-analytic model is usually expressed as

$$X = Lf + e \tag{1.1.1}$$

where

$X = p$-dimensional vector of observed responses,

$$x' = [x_1, x_2, \ldots, x_p]$$

$f = q$-dimensional vector of unobservable variables called common factors,

$$f' = [f_1, f_2, \ldots, f_q]$$

$e = p$-dimensional vector of unobservable variables called unique factors,

$$e' = [e_1, e_2, \ldots, e_p]$$

$L = pxq$ matrix of unknown constants called factor loadings

$$L = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1q} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pq} \end{bmatrix}$$

There are p unique factors and it is generally assumed that the unique part of each variable is uncorrelated with each other or with their common part; that is

$$E(ee') = \Psi = \begin{bmatrix} \Psi_1 & 0 & \cdots & 0 \\ 0 & \Psi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \Psi_p \end{bmatrix}$$

and

$$\mathrm{Cov}(e, f') = 0$$

The model given by (1.1.1) along with the imposed assumptions implies that the covariance matrix of the response vector X denoted by Σ can be expressed as

$$\Sigma = L\Phi L' + \Psi \tag{1.1.1}$$

where L and Ψ are as previously defined and

$$\Phi = \begin{bmatrix} 1 & & & \\ \phi_{21} & 1 & & \\ \phi_{31} & \phi_{32} & \ddots & \\ \vdots & \vdots & & \ddots \\ \phi_{q1} & \phi_{q2} & \cdots & \phi_{,q-1} \quad 1 \end{bmatrix}$$

The qxq symmetric matrix Φ has elements ϕ_{ij}, $i,j = 1, 2, ..., q$ which give the covariances (correlations) between the common factors. Note that since each column of L may be scaled arbitrarily, we have assumed, without loss of generality, that the common factors have unit variances. Therefore the diagonal elements of Φ have been replaced with ones. If we further assume that the factors themselves are uncorrelated, then

$$\Phi = I$$

and thus (1.1.2) becomes

$$\Sigma = LL' + \Psi \tag{1.1.3}$$

Given a particular Σ, certain conditions must be met for the factorization in (1.1.3) to exist and if it does, to be unique. The issue of identifiability and uniqueness of parameter estimates often poses difficulty in the context of the common factor analytic model. The total number of parameters in need of estimation is the number of factor loadings, namely pq. There are $^1/_2 p(p+1)$ equations. Generally, the requirement for identification is that the number of parameters be less than the number of equations, so that

$$pq + p < \tfrac{1}{2}p(p+1) \quad \text{or} \quad q < \tfrac{1}{2}(p-1)$$

Therefore, q should be fairly small compared to p. Unfortunately, this does not guarantee that a solution will exist.

It is however important to note that in the case of exploratory factor analysis, if $q > 1$ and a solution exists it is not generally unique. Using (1.1.3), we see that any orthogonal rotation of factors in the relevant q-space will give a new set of factors which will also satisfy the conditions of equation (1.1.3). To illustrate, let T be an orthogonal matrix of order (qxq). We now have

$$(LT)(LT)' = LTT'L' = LL' \tag{1.1.4}$$

However, even though the loading in L and LT are different, their ability to generate the given covariances in Σ is the same.

The basic model given by (1.1.1) may be written alternatively as

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1q} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2q} \\ \vdots & \vdots & & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pq} \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_q \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_p \end{bmatrix} \tag{1.1.5}$$

or

$$X_i = \sum_{j=1}^{q} \lambda_{ij} f_j + e_i \tag{1.1.6}$$

A set of equations like those in (1.1.5) is called a factor pattern. For simplicity the factor pattern is usually shown in tabular form in which only the factor coefficients are listed. The pxq matrix of factor loadings (where $p < q$) with the factor designations as column is referred to as the pattern matrix. We are also interested in the correlation between the variables and the common factors. A matrix of such correlations is called a factor structure matrix. Both structure and pattern are needed for a complete solution. However, though in general the elements of a structure matrix are different from the coefficients of a pattern matrix, in the case of uncorrelated and standardized factors, the two are identical - the factor loadings λ_{ij} refer to the correlations between the jth factor f_j and the ith variable X_i.

Though the common factor- analytic model has been developed in terms of the variance - covariance matrix of the observed responses, the original variables are usually standardized so that the basic input to a common factor analysis is the correlation matrix. Denoting the correlation matrix by **R**, we can rewrite (1.1.2) as

$$\mathbf{R} = L\Phi L' + \Psi \tag{1.1.7}$$

where the qxq symmetric matrix Φ contains the correlation between the common factors. The product matrix $L\Phi L'$ is called the common factor correlation matrix. Equation (1.1.5) could also be written as equivalent to a linear factor model in (1.1.8)

$$\begin{aligned} X_1 &= \lambda_{11}f_1 + \lambda_{12}f_2 + \cdots + \lambda_{1q}f_q + e_1 \\ X_2 &= \lambda_{21}f_1 + \lambda_{22}f_2 + \cdots + \lambda_{2q}f_q + e_2 \\ &\vdots \\ X_p &= \lambda_{p1}f_1 + \lambda_{12}f_2 + \cdots + \lambda_{1q}f_q + e_p \end{aligned} \tag{1.1.8}$$

Note that each equation in (1.1.8) partitions the variable X_i into two uncorrelated parts

$$X_i = c_i + e_i \tag{1.1.9}$$

where $c_i = \lambda_{i1}f_1 + \lambda_{i2}f_2 + \ldots + \lambda_{iq}f_q$ is that part of each variable that is common to the other p-1 variables, and e_i is that part of each variable that is unique.

Because the common and unique parts of a variable are assumed uncorrelated and because the common factors have unit variance, we can partition the variance of X_i into

$$\text{var}(X_i) = \text{var}(c_i) + \text{var}(e_i) \tag{1.1.10}$$

where var(c_i) and the var(e_i) represent the common variance and the unique variance of X_i, respectively. The common variance of a variable is also called the communality of the variable. By communality of a variable is meant that portion of a variable's total variance that is accounted for by the common factors. Letting h_i^2 to represent the communality of the ith variable, we can write

$$\text{Var}(X_i) = h_i^2 + \Psi_i \tag{1.1.11}$$

where var(e_i) = Ψ from the basic model given in (1.1.1). Note that

$$\text{Var}(c_i) = \sum_{j=1}^{q} \lambda_{ij}^2 = h_i^2 \tag{1.1.12}$$

is simply the sum of the squared elements in the ith row of L. The unique variance of a variable, Ψ, is called the uniqueness of the variable and reflects the extent to which the common factors fail to account for the variance of the variable - it is the portion left unexplained by the common factors.

The total contribution of factor f_j to the total variance of the entire set of variables is given by the eigenvalues of the factor f_j, which can be obtained by computing

$$\begin{aligned} {}_j &= \sum_{i=1}^{p} \lambda_{ij}^2 \\ V_j &= \lambda_j'\lambda_j \end{aligned} \tag{1.1.13}$$

where λj denotes the jth column of L. Equation (1.1.13) is nothing more than the squared factor loadings, $\sum_{i=1}^{p} \lambda_{ij}$ for $j=1,2,\ldots,q$. The total contribution of all the common factors to the total variance among all the variables is the total communality, defined as

$$V = \sum_{j=1}^{q} V_j \tag{1.1.14}$$

The variance among all the variables that is accounted for by a factor f_j as a percentage of that accounted for by all the factors is given by

$$V_c = \frac{V_j}{V} \tag{1.1.15}$$

Thus, the total variance can be written as

$$\begin{aligned} \text{Total Variance} &= \text{tr}(\Sigma) \\ &= \sum_{j=1}^{q} V_j + \sum_{i=1}^{p} \Psi_i \\ &= \sum_{i=1}^{p} \sum_{j=1}^{q} \lambda_{ij}^{\ 2} + \sum_{i=1}^{p} \Psi_i \end{aligned} \tag{1.1.16}$$

Common Factor Analysis versus Principal Component Analysis

The present study utilizes a model that represents a large set of observed variables by some smaller set that still preserves the essential information. Horst (1965) and van de Geer (1971) discussed principal components analysis (PCA) as one method of dealing with this problem. Principal components analysis is the most commonly employed methods of a wide class of data reduction procedure typically called components analysis. A second class of procedure called factor analysis has been employed for the same problem. In principal components analysis:

1) p linear compounds are needed to account for the total variance of p variables.

2) The sum of the variances of all p principal components is equal to the sum of the variances of the original variables.

In principal components analysis, the unobservable factors are expressed as functions of the observable variables as in (1.2.1)

$$\begin{aligned} PC_{(1)} &= w_{(1)1}X_1 + w_{(1)2}X_2 + \cdots + w_{(1)p}X_p \\ PC_{(2)} &= w_{(2)1}X_1 + w_{(2)2}X_2 + \cdots + w_{(2)p}X_p \\ &\vdots \\ PC_{(m)} &= w_{(m)1}X_1 + w_{(m)2}X_2 + \cdots + w_{(m)p}X_p \end{aligned} \tag{1.2.1}$$

In principal components analysis, the total variation contained in the set of variables is considered. In contrast, with the common factor-analytic model interest centers on that part of the total variance that is shared by the variables. The common factor-analytic model assumes that a variable consists of common and unique parts. The common part of a variable is that part of the variable's variation that is shared with the other variables, whereas the unique part of a variable is that part of the variable's variation that is specific, to that variable alone. Thus, one important distinction between principal components and common factor analysis comes from the amount of variance analyzed. In factor analysis a formal model

is specified describing each original variable in terms of a linear function of a small amount of unobservable common factors and a single latent unique factor. Its algebraic representation is as follows:

$$\begin{aligned} X_1 &= v_{1(1)}CF_{(1)} + v_{1(2)}CF_{(2)} + \ldots + v_{1(m)}CF_{(m)} + e_1 \\ X_2 &= v_{2(1)}CF_{(1)} + v_{2(2)}CF_{(2)} + \ldots + v_{2(m)}CF_{(m)} + e_2 \\ &\vdots \\ X_p &= v_{p(1)}CF_{(1)} + v_{p(2)}CF_{(2)} + \ldots + v_{p(m)}CF_{(m)} + e_p \end{aligned} \qquad (1.2.2)$$

Here, there are m ($<p$) common factors, denoted by $CF_{(i)}$, $i = 1, 2, \ldots,$ m where the $v_{j(i)}$, $j = 1, 2, \ldots, m$, give the weight of the ith common factor associated with the jth observable variable, and the e_j, $j = 1, 2, \ldots, p$, are the unique factor effects.

In contrasting equations (1.2.1) and (1.2.2), we find that in one case the inherently unobservable factor is a function of its indicators (principal components analysis), whereas in the other case, the indicators are a function of the unobservables (common factor analysis). In general expressing an unobservable as a function of its indicators is not equivalent to expressing the indicators as a function of the unobservable. Empirically, the parameter' effects will be different since in principal components analysis there is no error term (see eq. 1.2.1). Conceptually, the absence of error term implies that the observable variables are measured without error and that the unobservable latent principal component is a perfect linear combination of its measures.

2. APPLICATION

Fig. 2.1 presents a schematic representation of the factor analysis results of 2500 responses gathered in Metro Manila, i.e., Manila, Makati, Pasay, Paranaque, Caloocan and Quezon City. Of the 2500 respondents, 52.8 percent were male respondents, 47.1 percent were female respondents and .1 percent represented the third sex. Also out of 2500 respondents, 51.9 percent were below 25 years of age, 35.1 percent were between 25 to 45 years old and 13 percent were above 45 years old.

The analysis considered consumers' ratings of the importance of 14 variables in choosing a car. The 14 variables were patterned after Kachigan's (1982) study which are: V1 - low cost repairs; V2 - variety of colors; V3 - roomy interior; V4 - good gas mileage; V5 - good handling; V6 - modern looking; V7 - high resale value; V8 - comfortable; V9 - large engine; V10 - sleek appearance; V11 - easy to drive; V12 - eye catching; V13 - large trunk space; V14 - easy to park.

The 14 variables can be characterized by three latent underlying dimensions relation to (1) ease of handling, (2) stylishness and (3) cost efficiency. Thus, instead of having to understand 14 variables, we have simplified matters to the extent that now only three factors need be considered in characterizing the underlying structure of the car data.

Table 1 shows the initial statistics for each factor. The total variance explained by each factor is listed under the column labeled eigenvalue. Table 1 shows that almost 54 percent of the total variance is attributable to the first three factors. The remaining eleven factors together account for only 46.3 percent. Thus, a model with three factors may be adequate to represent the data.

Fig. 2.2 is a plot of the total variance associated with each factor. The plot shows a distinct break between the steep slope of the large factors and the gradual trailing off the rest of the factors. The gradual trailing off is called scree, (Catell, 1965) because it resembles the rubble that forms at the foot of a mountain. Experimental evidence indicates that the scree begins at the Kth factor, where K is the true

number of factors. From the scree plot, it again appears that a three-factor model should be sufficient for the car data.

Table 2 shows the factor loadings before rotation of the axes. Each row of Table 2 contains the coefficients used to express a standardized variable in terms of the factors. These coefficients are called factor loadings since they indicate how much weight is assigned to each factor. Factors with large coefficients (in absolute value) for a variable are closely related to the variable. For example, Factor 1 is the factor with the largest loading for the V8 (Comfortable) variable.

When the estimated factors are uncorrelated with each other (orthogonal) the factor loadings are also the correlations between the factors and the variables. Thus, the correlation between V7 (High resale value) and Factor 1 is .60124. Similarly, there is a very small correlation (-.00410) between V7 and Factor 2. The unrotated factor matrix is difficult to interpret. Table 2 shows that the car data are heavily loaded on factor 1 than on factors 2 and 3. The loadings on factors 2 and 3 are mostly bipolar. Variables 3, 4, 5,6, 7, 8, 11, 12, 13 and 14 load highly on factor 1; while variables 2 and 9 load highly on factor 2. Finally, variable 1 load highly on factor 3. Variables loading highly on a factor are underlined.

To judge how well the three factor model describes the original variables, the proportion of variances explained by the 3 factor model was computed. Since the factors are uncorrelated, the communality of the variable which is the proportion of variance explained by the common factors is just the sum of variance explained by each factors.

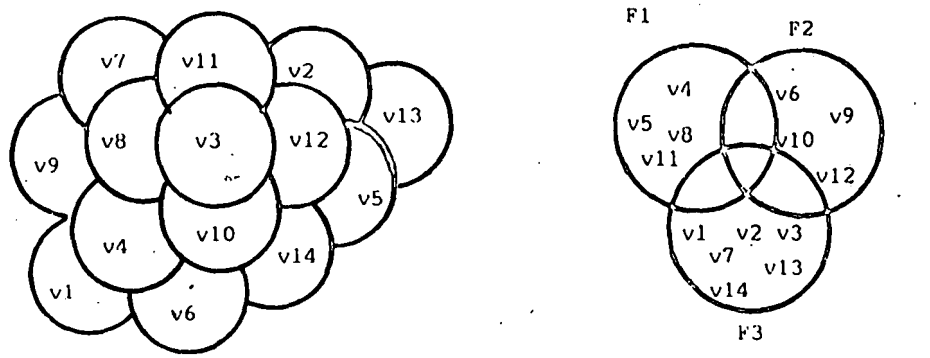

Fig. 2.1: Schematic representation of 14 variables reduced to only three factors

Consider for example variable 7. Factor 1 accounts for 36.15 percent of the variance for this variable. This is obtained by squaring the correlation coefficient for Factor 1 and V7 $(.60124)^2$. Similarly, Factor 2 explains .001681 percent of the variance, and Factor 3 accounts for 2.55 percent of the variance. The total percentage of variance in V7 accounted by this three factor model is therefore $[(.60124)^2 + (.15964)^2]$ or (36.15% + .00168% + 2.55%) = 38.7% * (see Table 4).

Table 3 shows the factor loadings after a varimax rotation. The rotation phase of factor analysis attempts to transform the initial matrix into one that is easier to interpret. Also, rotation redistributes the explained variance for the individual factors. Each variable's highest (absolute) loading is underlined in the table. Interpretations will be based on those variables loading highest on a given factor. It is to be

Table 1
Initial Factors

Initial Statistics

Variable	Communality	*	Factor	Eigenvalue	Pct. of Var.	Cum. Pct.
V1	1.0000	*	1	4.73044	33.8	33.8
V2	1.0000	*	2	1.75275	12.8	46.3
V3	1.0000	*	3	1.03479	7.4	53.7
V4	1.0000	*	4	0.92694	6.6	60.3
V5	1.0000	*	5	0.81298	5.8	66.1
V6	1.0000	*	6	0.72966	5.2	71.3
V7	1.0000	*	7	0.67399	4.8	76.2
V8	1.0000	*	8	0.59669	4.3	80.4
V9	1.0000	*	9	0.55710	4.0	84.4
V10	1.0000	*	10	0.51312	3.7	88.1
V11	1.0000	*	11	0.46988	3.4	91.4
V12	1.0000	*	12	0.43114	3.1	94.5
V13	1.0000	*	13	0.40958	2.9	97.4
V14	1.0000	*	14	0.36095	2.6	100.0

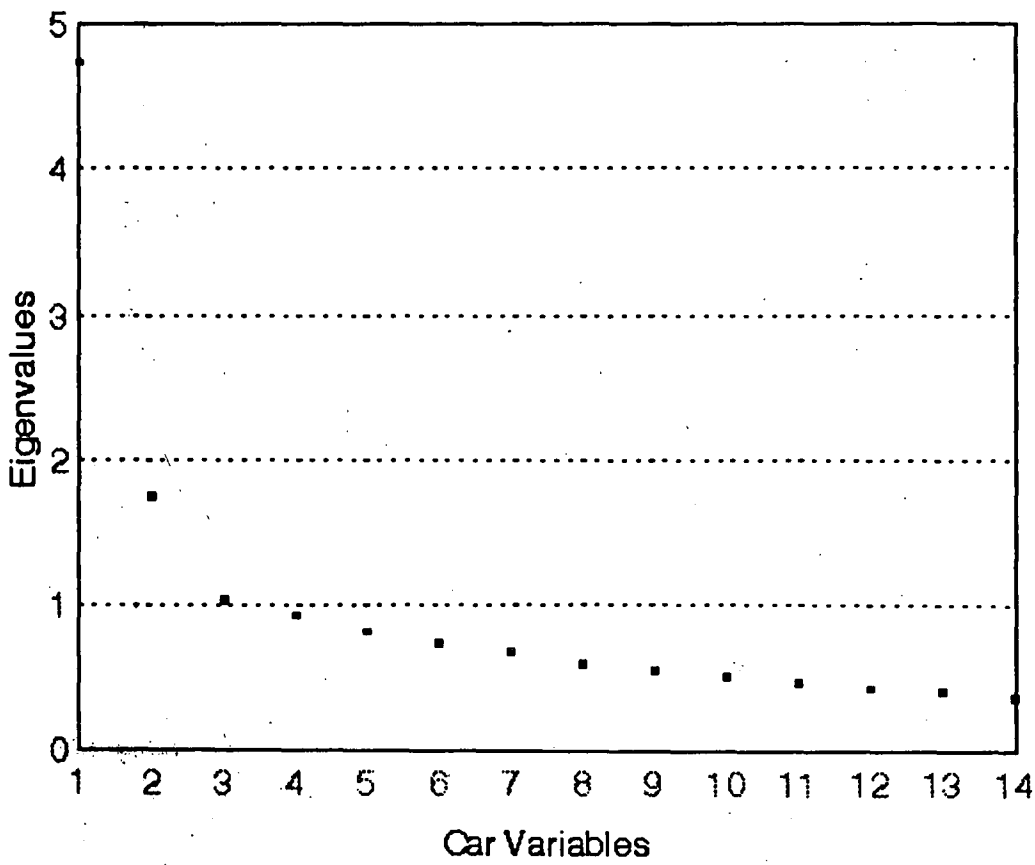

Fig. 2.2: Scree test based on Eigenvalues

Table 2

Factor Matrix

	FACTOR 1	FACTOR 2	FACTOR 3
V1	0.46361	-0.14777	0.58575
V2	0.40162	0.52629	0.20390
V3	0.60465	-0.01690	0.19097
V4	0.63323	-0.50960	0.05694
V5	0.66208	-0.34612	-0.11933
V6	0.63782	0.15374	-0.41734
V7	0.60124	-0.00410	0.15964
V8	0.63958	-0.46851	-0.17715
V9	0.48765	0.49049	0.05649
V10	0.60004	0.34856	-0.28143
V11	0.61746	-0.44052	0.13655
V12	0.56732	0.40756	-0.36313
V13	0.59930	0.32802	0.31758
V14	0.55837	0.03429	0.17705

Table 3
Factor Loading After Rotation

Factor Matrix

	FACTOR 1	FACTOR 2	FACTOR 3
V1	0.30165	-0.18157	0.67522
V2	-0.16036	0.42462	0.52329
V3	0.37255	0.20137	0.47224
V4	0.77702	-0.01369	0.24489
V5	0.70945	0.20463	0.16490
V6	0.38386	0.67422	0.05206
V7	0.36685	0.22551	0.44896
V8	0.79515	0.14954	0.07298
V9	-0.05007	0.53258	0.44207
V10	0.19098	0.69763	0.19384
V11	0.75254	0.13194	0.10449
V12	0.14175	0.76387	0.12740
V13	0.09284	0.34093	0.66540
V14	0.30740	0.21666	0.45041

noted that variables with higher loadings are to be considered as having greater influence. Also, the loading of each factor and the correlation of the variables with the factors is known as factorial validity of the variables. Factorial validity is essentially the correlation of the variables with whatever is common to a group of variables.

Table 3 shows that the variables loading highly on Factor 1 are V4 (good gas mileage), V5 (good handling), V8 (comfortable) and V11 (easy to drive). Thus, factor 1 might be interpreted as something like ease of handling.

The second factor loads highly on V6 (modern looking), V9 (large engine), V10 (sleek appearance) and V12 (eye catching). These describe the dimension of stylishness.

Finally, factor 3 is loading highly on V1 (low cost repairs), V2 (variety of colors), V3 (roomy interior), V7 (high resale value), V13 (large trunk space) and V14 (easy to park). The last factor is associated with cost efficiency. Thus, the car data maybe fairly characterized and attributed to these three factors: ease of handling, stylishness and cost efficiency. Table 4 reveals the final three selected factors.

Table 4 shows the communalities for the variables together with the percentage of variance accounted for by each of the retained factors. Recall that in Table 1, about 53.7% of the total variance is accounted for by the first three eigenvalues. Again, it appears safe to conclude that in terms of variance explained, three factors sufficiently capture the variance structure of the original data.

Factor Scores

Table 4
Selected Factors

Final Statistics

Variable	Communality	*	Factor	Eigenvalue	Pct. of Var.	Cum. Pct.
V1	0.57988	*	1	4.73044	33.8	33.8
V2	0.47985	*	2	1.75275	12.8	46.3
V3	0.40236	*	3	1.03479	7.4	53.7
V4	0.66392	*				
V5	0.57239	*				
V6	0.60463	*				
V7	0.38699	*				
V8	0.65995	*				
V9	0.48157	*				
V10	0.56074	*				
V11	0.59396	*				
V12	0.61982	*				
V13	0.56762	*				
V14	0.34430	*				

Since one of the goals in factor analysis is to reduce a large number of variables to a smaller number of factors, it is often desirable to estimate factor scores for each case. Recall that a factor can be estimated as a linear combination of the original variables. Therefore, for case K, the score for the jth factor is estimated as

$$\hat{F}_{jk} = \sum_{i=1}^{p} W_{ji} X_{ik}$$

where X_{ik} is the standardized value of the ith variable for case k, and W_{ji} is the factor score coefficient for jth factor and the ith variable. Except for principal components analysis, exact factor scores cannot be obtained. Estimates are obtained instead.

3. SOME PRACTICAL ISSUES

Issues that surface in real-life applications of Factor Analysis are: (1) Choosing the appropriate data (variables) as input to the Factor Analysis module; (2) Interpreting the underlying construct or dimension (factor); and (3) What to do with cases that have missing values.

Choosing the appropriated data (variables) as input to a Factor Analysis module

One of the criteria in choosing variables for the factor analysis model is that the variables must be correlated with each other. If the correlations between variables are small, then it is possible that the variables do not share common factors. In this case, the use of factor analysis model is inappropriate. Bartlett's test of Sphericity can be used to test the hypothesis that the correlation matrix is an identity matrix(i.e., all diagonal terms are 1 and all off-diagonal terms are 0) or in other words, the variables are uncorrelated. The test requires that the data be a sample from a multivariate normal population. If the test statistic for sphericity (based on a chi-square transformation of the determinant of the correlation matrix) is large and the corresponding significance level is small, then it appears unlikely that the population correlation matrix is an identity. This result would be appropriate for a factor analysis. However, if the hypothesis that the population correlation matrix is an identity cannot be rejected because the observe significance level is large, then one should reconsider the use of a factor model. Another indicator of the strength of relationships among the variables is the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy where the magnitudes of the correlation coefficients are being compared with the magnitudes of the partial correlation coefficients. It is computed as

$$KMO = \frac{\sum_{i \neq j} \Sigma r_{ij}^2}{\sum_{i \neq j} \Sigma r_{ij}^2 + \sum_{i \neq j} \Sigma a_{ij}^2}$$

Table 5

KMO and Bartlett's Test Results

Kaiser-Meyer-Olkin (KMO Measure of Sampling Adequacy = 0.8742	
Bartlett's Test of Sphericity = 9437.0868	**Significance = 0.0000**

where r_{ij} is the simple correlation coefficient between variables i and j and a_{ij} is the partial correlation coefficient between variables i and j. If the sum of the squared partial correlation coefficients between all pairs of coefficients, the *KMO* measure is close to 1. Small values for the *KMO* measure indicate that a factor analysis of the variables may not be a good idea, since correlations between pairs of variables cannot be explained by the other variables. Kaiser (1974) characterizes measures in the .90's marvelous, in the .80's as meritorious, in the .70's as average, in the .60's as mediocre, in the .50's as miserable, and below .50 as unacceptable. Table 5 shows the *KMO* index of sampling adequacy and the Bartlett's test for th car data in this study.

Table 5 reveals that the overall *KMO* is close to 0.9 and the Bartlett's test is highly significant, so it is safe to say that factor analysis is appropriate for the car data.

Interpreting Factors

A problem in interpreting the factors in deciding when a borderline loading should be considered significant or salient. Using the salient loadings to cluster the variables may aid the interpretation. "Salient" has been used on an intuitive basis to identify high loading. More technically, a salient loading is one that is sufficiently high to assume that a relationship exists between the variable and the factor. In addition, it usually means that the relationship is high enough so that the variable can aid in interpreting the factor and vice versa. In some situations however, factors may occur that present problems for interpretation. Such problems arise when the variables are not sufficiently understood, when the factor includes such a wide range of variables that they cannot be readily integrated, or when the factor is poorly defined. Poorly defined factors are generally those that do not have several salient loadings by variables that load on specific factors. Without a unique set of variables loading the factor, there is no real basis for interpreting the factor. There is, of course no reason to interpret all factors. It therefore follows that only factors well defined by interpretable variables are being examined.

Cases that have missing values

In cases where some individuals may have scores on all variables except one or two, the following may be suggested. the missing element can be replaced by calculating the mean for the variable from the individuals who do have scores on it. The mean is then used as the best estimate of the missing element. Or another individual can be selected at random and his score on the variables is used to replace the missing element. The latter procedure will, on the average, leave both the mean and variance unaffected. The former procedure leaves only the mean unaffected. Both procedures reduce correlations with other variables.

To maintain the variable's correlations with the other variables, a multiple regression analysis is used. The other variables are used to predict the variable with a missing element. The regression analysis is computed only with those individuals who have scores on all the variables to be used in the analysis. The regression weights are then applied to the known scores to estimate the missing score. A regression analysis is computed for every variable that has one or more missing scores.

The regression procedure has been found more effective for estimating what the correlation matrix would have been if all score had been available (Timm, 1970). The other procedures against which it was compared included dropping the individuals who had a missing score, replacing the missing score with the mean, and a procedure based on principal components. The regression procedure was best whether 1% or 20% of the data were missing (Gorsuch, 1983).

Computer programs are available that calculate each correlation coefficient from only the individuals who have the necessary scores. The resulting correlation matrices should be factored only if the number of individuals is quite similar for all elements. If the number of individuals varies widely form element to element, the coefficients may be sufficiently incompatible to prevent a factor analysis or distort the results.

REFERENCES

Dillon, W. R., et al **Multivariate Analysis, Methods and Applications.** New York: John Wiley and Sons, c. 1984.

Gorsuch, R. L. **Factor Analysis.** 2nd Ed., New Jersey: Lawrence Erlbaum Associates, c. 1983

Johnson, R. A., et al. **Applied Multivariate Statistical Analysis**. 2nd Ed. New Jersey: Prentice-Hall International, Inc., c. 1988.

Norusis, M. J. **SPSS/PC+ Base Manual V 2.0 for the IBM PC/XT/AT and PS/2**, c. 1988.

Norusis, M. J. **Advanced Statistics V 2.0 for the IBM PC/XT/AT and PS/2**, c. 1988.

Timm, N. H. The estimation of variance-covariances and correlation matrices from incompletedata **Psychometrika**. 1970, 35(4), 417-437.

Johnson, R. A., et al. **Applied Multivariate Statistical Analysis**. 2nd Ed. New Jersey: Prentice-Hall International, Inc., c. 1988.

Norusis, M. J. **SPSS/PC+ Base Manual V 2.0 for the IBM PC/XT/AT and PS/2**, c. 1988.

Norusis, M. J. **Advanced Statistics V 2.0 for the IBM PC/XT/AT and PS/2**, c. 1988.

Timm, N. H. The estimation of variance-covariances and correlation matrices from incompletedata **Psychometrika**. 1970, 35(4), 417-437.