Exploratory factor analysis

It is important to distinguish between exploratory and confirmatory analysis. In an exploratory analysis, one wants to explore the empirical data to discover and detect characteristic features and interesting relationships without imposing any definite model on the data. An exploratory analysis may be structure generating, model generating, or hypothesis generating. In confirmatory analysis, on the other hand, one builds a model assumed to describe, explain, or account for the empirical data in terms of relatively few parameters. The model is based on *a priori* information about the data structure in the form of a specified theory or hypothesis, a given classificatory design for items or subtests according to objective features of content and format, known experimental conditions, or knowledge from previous studies based on extensive data.

Exploratory factor analysis is a technique often used to detect and assess latent sources of variation and covariation in observed measurements. It is widely recognized that exploratory factor analysis can be quite useful in the early stages of experimentation or test development. Thurstone's (1938) primary mental abilities, French's (1951) factors in aptitude and achievement tests and Guilford's (1956) structure of intelligence are good examples of this. The results of an exploratory factor analysis may have heuristic and suggestive value and may generate hypotheses which are capable of more objective testing by other multivariate methods. As more knowledge is gained about the nature of social and psychological measurements, however, exploratory factor analysis may not be a useful tool and may even become a hindrance.

Most studies are to some extent both exploratory and confirmatory since they involve some variables of known and other variables of unknown composition. The former should be chosen with great care in order that as much information as possible about the latter may be extracted. It is highly desirable that a hypothesis which has been suggested by mainly exploratory procedures should subsequently be confirmed, or disproved, by obtaining new data and subjecting these to more rigorous statistical techniques.

The basic idea of factor analysis is the following. For a given set of response variables $x_1, x_2, ..., x_p$ one wants to find a set of underlying latent factors $\xi_1, \xi_2, ..., \xi_k$, fewer in number than the observed variables. These latent factors are supposed to account for the intercorrelations of the response variables in the sense that when the factors are partialed out from the observed variables, there should no longer remain any correlations between these. If both the observed response variables and the latent factors are measured in deviations from the mean, this leads to the model (see Jöreskog, 1979)

$$x_i = \lambda_{i1}\xi_1 + \lambda_{i2}\xi_2 + ... + \lambda_{in}\xi_k + \delta_i, \qquad (1)$$

where δ_i, the unique part of x_i, is assumed to be uncorrelated with $\xi_1, \xi_2, ..., \xi_k$ and with δ_i for $j \neq i$.. The unique part δ_i consists of two components: a specific factor s_i and a pure random measurement error e_i. These are indistinguishable, unless the measurements x_i are designed in such a way that they can be separately identified (panel designs and multitrait-multimethod designs). The term δ_i is often called the *measurement error* in x_i even though it is widely recognized that this term may also contain a specific factor as stated above.

In a confirmatory factor analysis, the investigator has such knowledge about the factorial nature of the variables that he/she is able to specify that each measure x_i depends only on a few of the factors ξ_j. If x_i does not depend on ξ_j, λ_{ij} = 0 in (1). In a path diagram, this means that there is no one-way arrow from ξ_j to x_i. In many applications, the latent factor ξ_j represents a theoretical construct and the observed measures x_i are designed to be indicators of this construct. In this case there is only one non-zero λ_{ij} in each equation (1).

Equation (1) can be written in matrix form as

$$\mathbf{x} = \mathbf{\Lambda}\boldsymbol{\xi} + \boldsymbol{\delta}, \tag{2}$$

where $\mathbf{x}' = (x_1, x_2, ..., x_p)$, $\boldsymbol{\xi}' = (\xi_1, \xi_2, ..., \xi_k)$, $\boldsymbol{\delta}' = (\delta_1, \delta_2, ..., \delta_p)$. The matrix $\mathbf{\Lambda}$ of order $p \times k$ is called the factor matrix or the factor loadings matrix.

The assumption that $\boldsymbol{\delta}$ is uncorrelated with $\boldsymbol{\xi}$, implies that the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{x}$ is

$$\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Theta}, \tag{3}$$

where $\mathbf{\Phi}$ and $\mathbf{\Theta}$ are the covariance matrices of $\boldsymbol{\xi}$ and $\boldsymbol{\delta}$, respectively.

If $k > 1$ and there are no restrictions on $\mathbf{\Lambda}$, *i.e.*, in the exploratory case, the factors $\boldsymbol{\xi}$ are not uniquely defined, because one can make an arbitrary linear transformation of the factors. Let **T** be an arbitrary non-singular matrix of order $k \times k$ and let

$$\boldsymbol{\xi}^* = \mathbf{T}\boldsymbol{\xi} \quad \mathbf{\Lambda}^* = \mathbf{\Lambda}\mathbf{T}^{-1} \quad \mathbf{\Phi}^* = \mathbf{T}\mathbf{\Phi}\mathbf{T}'$$

Then we have identically

$$\mathbf{\Lambda}^*\boldsymbol{\xi}^* = \mathbf{\Lambda}\boldsymbol{\xi} \quad \mathbf{\Lambda}^*\mathbf{\Phi}^*\mathbf{\Lambda}^{*\prime} \equiv \mathbf{\Lambda}\mathbf{\Phi}\mathbf{\Lambda}'$$

This shows that at least k^2 independent conditions must be imposed on $\mathbf{\Lambda}$ and/or $\mathbf{\Phi}$ to make these identified. It is common to assume that the factors are uncorrelated and standardized, in which case **T** is restricted to be an orthogonal matrix. Such factors and factor loadings are only determined up to an orthogonal transformation.

Exploratory factor analysis is usually performed in two steps: First, estimate an arbitrary matrix $\mathbf{\Lambda}$; next, transform this matrix according to external criteria for simple structure to facilitate interpretation of the data. With TSLS estimation, one must use a different approach. If $\mathbf{\Lambda}$ has rank *k*, one can choose a transformation **T** such that there will be *k* rows of $\mathbf{\Lambda}$ that form an identity matrix. For simplicity of exposition, we assume that the variables in **x** have been ordered so that the first *k* rows of $\mathbf{\Lambda}$ form an identity matrix. The variables that correspond to the identity matrix are called reference variables and this solution is called a reference variables solution. In practice, it is best to choose as reference variables those variables that are the best indicators of each factor.

Partitioning $\mathbf{x}$ into two parts $\mathbf{x}_1(k \times 1)$ and $\mathbf{x}_2(q \times 1)$, where $q = p - k$, and $\boldsymbol{\delta}$ similarly into $\boldsymbol{\delta}_1(k \times 1)$ and $\boldsymbol{\delta}_2(q \times 1)$, (2) can be written

$$\mathbf{x}_1 = \boldsymbol{\xi} + \boldsymbol{\delta}_1 \tag{4}$$

$$\mathbf{x}_2 = \mathbf{\Lambda}_2\boldsymbol{\xi} + \boldsymbol{\delta}_2, \tag{5}$$

where $\boldsymbol{\Lambda}_2(q \times k)$ consists of the last $q = p - k$ rows of $\boldsymbol{\Lambda}$. The matrix $\boldsymbol{\Lambda}_2$ may, but need not, contain *a priori* specified elements. We say that the model is *unrestricted* when $\boldsymbol{\Lambda}_2$ is entirely unspecified and that the model is *restricted* when $\boldsymbol{\Lambda}_2$ contains *a priori* specified elements. For a more general discussion of restricted and unrestricted solutions, see Jöreskog (1969).

Solving (4) for $\boldsymbol{\xi}$ and substituting this into (5) gives

$$\mathbf{x}_2 = \boldsymbol{\Lambda}_2 \mathbf{x}_1 + \mathbf{u}, \qquad (6)$$

where $\mathbf{u} = \boldsymbol{\delta}_2 - \boldsymbol{\Lambda}_2 \boldsymbol{\delta}_1$. Each equation in (5) is of the form $y = \boldsymbol{\gamma}' \mathbf{x} + u$ but it is not a regression equation because $\mathbf{u}$ is correlated with $\mathbf{x}_1$, since $\boldsymbol{\delta}_1$ is correlated with $\mathbf{x}_1$.

Hägglund (1982) showed that instrumental variables can be obtained as follows. Let

$$x_i = \lambda_i' \mathbf{x}_1 + u_i,$$

be the *i*-th equation in (6), where λ_i' is the *i*-th row of $\boldsymbol{\Lambda}_2$, and let $\mathbf{x}_{(i)}(q-1 \times 1)$ be a vector of the remaining variables in $\mathbf{x}_2$. Then u_i is uncorrelated with $\mathbf{x}_{(i)}$ so that $\mathbf{x}_{(i)}$ can be used as instrumental variables for estimating (6). Provided $q \geq k+1$, this can be done for each $i = 1, 2, ..., q.$

The factors in a reference variables solution are neither standardized nor uncorrelated. After $\boldsymbol{\Lambda}$ has been estimated, the covariance matrix $\boldsymbol{\Phi}$ and the error covariance matrix $\boldsymbol{\Theta}$ can be estimated directly (non-iteratively) by unweighted or generalized least squares because the covariance structure in (3.14) is linear in $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$, see Browne (1974). Since most people prefer to interpret factors that are standardized, the solution is rescaled to standardized factors in the output.

The TSLS solution described here is used in LISREL to obtain starting values for the iterative methods, *i.e.*, the free elements of $\boldsymbol{\Lambda}_y$ and $\boldsymbol{\Lambda}_x$ in the model are estimated by the method described here and the elements of $\mathbf{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\Psi}$ are estimated by the method described previously. Note that this requires that $p \geq m+1$ and $q \geq n+1$, using the notation on p. 2 in Jöreskog & Sörbom (1996b). Whenever these conditions are not satisfied other methods are used to produce starting values.

1. Example: Exploratory Factor Analysis of Nine Psychological Variables

To illustrate exploratory factor analysis we use a classical data set. Holzinger & Swineford (1939) collected data on twenty-six psychological tests administered to 145 seventh- and eighth-grade children in the Grant-White school in Chicago. Nine of these tests are selected for this example. The nine selected variables and their intercorrelations are given in the table below.[1]

In a completely exploratory factor analysis, both the number of factors and the reference variables are unknown and must be determined from the data.

But to begin with, we shall assume that both of these are known and later show how to handle the completely exploratory case.

Table: Correlation matrix for nine psychological variables

[1] The correlations given here differ slightly from those given in Table 1.5 in Jöreskog & Sörbom (1996c) and Table 3.7 in Jöreskog & Sörbom (1996b), because the former have been computed from the raw scores (see file NPV1.PR2) whereas those in the previous tables have been copied from published correlation tables. There is no doubt that the correlations given here are the correct ones.

VIS PERC	1.000								
CUBES	0.326	1.000							
LOZENGES	0.449	0.417	1.000						
PAR COMP	0.342	0.228	0.328	1.000					
SEN COMP	0.309	0.159	0.287	0.719	1.000				
WORDMEAN	0.317	0.195	0.347	0.714	0.685	1.000			
ADDITION	0.104	0.066	0.075	0.209	0.254	0.178	1.000		
COUNTDOT	0.308	0.168	0.239	0.104	0.198	0.121	0.587	1.000	
S-C CAPS	0.487	0.248	0.373	0.314	0.356	0.222	0.418	0.528	1.000

We assume that there are three factors and we use VIS PERC, PAR COMP, and ADDITION as reference variables. Thus all factor loadings for these variables are known. To estimate the factor loadings for the other variables, we estimate the relationship between each of these and the reference variables using all the others as instrumental variables. This can be done with the following RG commands in LISREL, where the index numbers of the variables are used instead of the names of the variables:

```
RG 2 on 1 4 7 with 3 5 6 8 9
RG 3 on 1 4 7 with 2 5 6 8 9
RG 5 on 1 4 7 with 2 3 6 8 9
RG 6 on 1 4 7 with 2 3 5 8 9
RG 8 on 1 4 7 with 2 3 5 6 9
RG 9 on 1 4 7 with 2 3 5 6 8
```

If there are many variables and factors this is not practical. We have therefore made it completely automatic. In the SIMPLIS command language, simply write

```
Factor Analysis with 3 Factors
```

In the LISREL command language and in PRELIS, write

```
FA NF=3
```

LISREL will determine a suitable set of reference variables by a promax rotation of the maximum likelihood solution. Some users may prefer to use the unrotated solution, the varimax solution, or the promax solution, so all four solutions are given in the output.

Suppose the correlation matrix is stored in the file **NPV.KM**. Then a SIMPLIS command file for factor analyzing the variables in this table is (see file **NPV2.SPL**). All syntax for this example may be found in the **SIMPLIS Examples** folder.

```
Exploratory Factor Analysis of Nine Psychological Variables
Observed Variables
   'VIS PERC' CUBES LOZENGES 'PAR COMP' 'SEN COMP' WORDMEAN
    ADDITION COUNTDOT 'S-C CAPS'
Correlation Matrix from File NPV.KM
Sample Size 145
Factor Analysis with 3 Factors
End of Problem
```

The results are:

```
Maximum Likelihood Factor Analysis for 3 Factors

 Unrotated Factor Loadings

            Factor 1   Factor 2   Factor 3 Unique Var
            --------   --------   -------- ----------
 VIS PERC      0.520      0.162      0.453      0.499
    CUBES      0.326      0.083      0.383      0.740
 LOZENGES      0.488      0.079      0.469      0.535
 PAR COMP      0.808     -0.318     -0.073      0.241
 SEN COMP      0.793     -0.214     -0.151      0.302
 WORDMEAN      0.764     -0.301     -0.058      0.322
 ADDITION      0.418      0.541     -0.379      0.388
 COUNTDOT      0.419      0.709     -0.064      0.317
 S-C CAPS      0.573      0.435      0.161      0.456

Minimum Fit Function Chi-Squarewith 12 Degrees of Freedom = 9.38

 Varimax-Rotated Factor Loadings

            Factor 1   Factor 2   Factor 3 Unique Var
            --------   --------   -------- ----------
 VIS PERC      0.201      0.165      0.659      0.499
    CUBES      0.106      0.050      0.496      0.740
 LOZENGES      0.213      0.077      0.643      0.535
 PAR COMP      0.836      0.075      0.234      0.241
 SEN COMP      0.794      0.186      0.180      0.302
 WORDMEAN      0.787      0.066      0.232      0.322
 ADDITION      0.175      0.761     -0.042      0.388
 COUNTDOT     -0.004      0.782      0.267      0.317
 S-C CAPS      0.192      0.526      0.479      0.456

 Promax-Rotated Factor Loadings

            Factor 1   Factor 2   Factor 3 Unique Var
            --------   --------   -------- ----------
 VIS PERC      0.678      0.039      0.031      0.499
    CUBES      0.531     -0.013     -0.051      0.740
 LOZENGES      0.670      0.064     -0.059      0.535
 PAR COMP      0.074      0.848     -0.045      0.241
 SEN COMP      0.007      0.805      0.086      0.302
 WORDMEAN      0.083      0.796     -0.049      0.322
 ADDITION     -0.184      0.131      0.793      0.388
 COUNTDOT      0.200     -0.145      0.774      0.317
 S-C CAPS      0.431      0.041      0.446      0.456
```

Factor Correlations

	Factor 1	Factor 2	Factor 3
Factor 1	1.000		
Factor 2	0.444	1.000	
Factor 3	0.344	0.261	1.000

Reference Variables Factor Loadings Estimated by TSLS

	Factor 1	Factor 2	Factor 3	Unique Var
VIS PERC	0.708	0.000	0.000	0.499
CUBES	0.538	-0.031	-0.075	0.740
	(0.22)	(0.15)	(0.16)	
	2.417	-0.215	-0.456	
LOZENGES	0.674	0.042	-0.086	0.535
	(0.26)	(0.15)	(0.18)	
	2.604	0.286	-0.480	
PAR COMP	0.000	0.871	0.000	0.241
SEN COMP	-0.033	0.808	0.128	0.302
	(0.15)	(0.11)	(0.11)	
	-0.212	7.081	1.127	
WORDMEAN	0.013	0.819	-0.007	0.322
	(0.15)	(0.11)	(0.11)	
	0.085	7.357	-0.062	
ADDITION	0.000	0.000	0.782	0.388
COUNTDOT	0.415	-0.298	0.731	0.317
	(0.19)	(0.14)	(0.21)	
	2.205	-2.164	3.412	
S-C CAPS	0.557	-0.061	0.413	0.456
	(0.19)	(0.13)	(0.14)	
	2.857	-0.454	2.847	

Factor Correlations

	Factor 1	Factor 2	Factor 3
Factor 1	1.000		
Factor 2	0.543	1.000	
Factor 3	0.240	0.284	1.000

The first solution is the unrotated solution computed using the maximum likelihood procedure described by Jöreskog (1967) and in more detail by Jöreskog (1977). The second solution is the varimax solution of Kaiser (1958). Both of these are orthogonal solutions, *i.e.*, the factors are uncorrelated. The third solution is the promax solution of Hendrickson & White (1964). This is an oblique solution, *i.e.*, the factors are correlated. The varimax and the promax solutions are transformations of the unrotated solution and as such they are still maximum likelihood solutions. The fourth solution is the TSLS solution

obtained in reference variables form as described earlier. The reference variables are chosen as those variables in the promax solution that have the largest factor loadings in each column. This gives VIS PERC, PAR COMP, and ADDITION as reference variables. The advantage of the TSLS solution is that standard errors can be obtained for all the variables except for the reference variables. This makes it easy to determine which loadings are statistically significant or not. The standard errors are given in parentheses below the loading estimate and the *t*-values are given below the standard errors. A simple rule to follow is to judge a factor loading statistically significant if its *t*-value is larger than 2 in magnitude. On the basis of the TSLS solution one can formulate a hypothesis for confirmatory factor analysis by specifying that all non-significant loadings are zero. This hypothesis should be tested on an independent sample.

The same result will be obtained with the following LISREL command file (see file **NPV3.LIS**):

```
Factor Analysis of Nine Psychological Variables
DA NI=9 NO=145  MA=KM
LA
   'VIS PERC' CUBES LOZENGES 'PAR COMP' 'SEN COMP' WORDMEAN
    ADDITION COUNTDOT 'S-C CAPS'
KM=NPV.KM
PC NC=3
OU NS
```

The reference variables solution given in the output is a TSLS solution. It is not a maximum likelihood solution. For exploratory factor analysis, the TSLS solution is often quite sufficient. However, one can obtain the maximum likelihood solution for the reference variables representation using the SIMPLIS command file **NPV4.SPL** or the LISREL command file **NPV5.LIS**. These files are not shown here. Because the correlation matrix is analyzed in these examples, the standard errors for the factor loadings are slightly incorrect, see Cudeck (1989). Correct standard errors can be obtained, still using the correlation matrix, by the more complicated LISREL command file **NPV6.LIS**. Alternatively, the problem of incorrect standard errors can be avoided entirely by analyzing the covariance matrix instead of the correlation matrix. To obtain standardized factor loadings put SC on the OU command in LISREL or on an Options line in SIMPLIS.

If the correlation matrix has not been computed but raw data is available, one can use the following PRELIS command file (see **NPV7A.PRL**) to obtain the same result directly (here it is assumed that the raw data is in the file **NPV.RAW**):

```
Factor Analysis of Nine Psychological Variables
Data NI = 9
Labels
'VIS PERC' CUBES LOZENGES 'PAR COMP' 'SEN COMP'
WORDMEAN  ADDITION COUNTDOT 'S-C CAPS'
Rawdata=NPV.RAW
Continuous 'VIS PERC' - 'S-C CAPS'
FA NF=3
Output MA=KM
```

This will also save the computed correlation matrix in a file so that if one wants to reanalyze the data with a different number of factors one can do this with LISREL rather than PRELIS. Regardless of whether PRELIS or LISREL is used, one can read in a large number of variables and use an SE command to select a subset of variables to factor analyze.

2. Number of Factors

There is no unique way to determine the number of factors. This is best done by the investigator who knows what the variables (are supposed to) measure. Then the number of factors can be specified *a priori* at least tentatively. Many procedures have been suggested in the literature to determine the number of factors analytically. One of them is to continue

to extract factors until there no longer are at least three large loadings in each column of the varimax solution, say. The question is what is meant by large. The TSLS reference variables solution offers an answer to this question by considering as large those loadings which are statistically significant. Our procedure for deciding on the number of factors is based on statistical fit.

If the number of factors is not specified, *i.e.*, if with 3 Factors or NF = 3 is omitted, PRELIS or LISREL will try to determine a suitable number of factors using a decision procedure based on a number of fit criteria for maximum likelihood factor analysis for $k = 0, 1, ..., k_{\max}$ where $k_{\max}$ is the largest number of factors for which a factor solution can be obtained. Note that only the solution for the number of factors determined in this way will appear in the output. If one wants a solution with a larger or smaller number of factors than that determined by this procedure, one must redo the analysis and specify the number of factors.

For our NPV example (see file **NPV7B.PRL**), the decision table for deciding the number of factors is based on the values in the table below.

Table: Fit statistics for deciding the number of factors

k	c_k	d_k	P_k	$\triangle c_k$	$\triangle d_k$	$P_{\triangle c}$	ρ_k
0	488.91	36	0.000				0.295
1	175.49	27	0.000	313.41	9	0.000	0.195
2	61.70	19	0.000	113.79	8	0.000	0.124
3	9.38	23	0.670	52.32	7	0.000	0.000
4	2.59	6	0.858	6.79	6	0.341	0.000

The quantities c_k, d_k, P_k, $\triangle c_k$, $\triangle d_k$, $P_{\triangle c}$, and ρ_k, are defined as follows.

$$c_k = [n - (2p+5)/6 - 2k/3][\ln\left|\hat{\boldsymbol{\Sigma}}\right| - \ln|\mathbf{S}|],\ k = 0,1,...,k_{\max} \tag{7}$$

$$d_k = [(p-k)^2 - (p-k)]/2, \quad k = 0,1,...,k_{\max} \tag{8}$$

$$P_k = \Pr\{\chi^2_{d_k} > c_k\},\ k = 0,1,...,k_{\max} \tag{9}$$

$$\triangle c_k = c_k - c_{k-1}, \quad k = 0,1,...,k_{\max} \tag{10}$$

$$\triangle d_k = d_k - d_{k-1},\ k = 0,1,...,k_{\max} \tag{11}$$

$$P_{\triangle c} = \Pr\{\chi^2_{d_k} > \triangle c_k\},\ k = 0,1,...,k_{\max} \tag{12}$$

$$\rho_k = \sqrt{[c_k - d_k / nd_k]},\ k = 0,1,...,k_{\max} \tag{13}$$

Here c_k is the chi-square statistic for testing the fit of *k* factors, see Lawley & Maxwell (1971, pp. 35–36). If the model holds and the variables have a multivariate normal distribution, this is distributed in large samples as χ^2 with d_k degrees

of freedom.[2] The *P*-value of this test is P_k , *i.e.*, the probability that a random χ^2 with d_k degrees of freedom exceeds the chi-square value actually obtained. For reasons stated elsewhere (see, *e.g.*, Jöreskog & Sörbom, 1996b, p. 28, or Jöreskog & Sörbom, 1996c, p. 122), it is better to regard these quantities as approximate measures of fit rather than as test statistics. $\triangle c_k$ measures how much better the fit is with *k* factors than with $k - 1$ factors.

$\triangle d_k$ and $P_{\triangle c}$ are the corresponding degrees of freedom and *P*-value. ρ_k is Steiger's (1990) *Root Mean Squared Error of Approximation* (RMSEA) which is a measure of population error per degree of freedom, see Browne & Cudeck (1993) or Jöreskog & Sörbom (1996c).

LISREL investigates these quantities for $k = 0,1,...,k_{\max}$ and determines the smallest acceptable *k* with the following decision procedure: If $P_k > .10$, *k* factors are accepted. Otherwise, if $P_{\triangle c} > .10$, $k - 1$ factors are accepted. Otherwise, if $\rho_k < .05$, *k* factors are accepted. If none of these conditions are satisfied, *k* is increased by 1.

The first criterion, $P_k > .10$, guarantees that one stops at *k* if the overall fit is good. The second criterion, $P_{\triangle c} > .10$, guarantees that one will not increase the number of factors unless the improvement in fit is statistically significant at the 10% level. The third criterion, $\rho_k < .05$, is the Browne–Cudeck guideline (Browne & Cudeck, 1993, p. 144). This guarantees that one does not get too many factors in large samples. This procedure may not give a satisfactory answer to the number of factors in all respects, but at least there will not be a tendency to overfit, *i.e.*, to take too many factors.

For the values in the table above the decision will be *k* = 3 factors, because for *k* = 2, P_k and $P_{\triangle c}$ are too small and ρ_k is too large, but for *k* = 3, P_k is acceptable.

In the output (see file **NPV7B.OUT**), the decision table is given as:

```
Decision Table for Number of Factors

 Factors      Chi2   df      P       DChi2 Df     PD      RMSEA
 -------      ----   --      -       ----- --     --      -----
    0        488.91  36    0.000                          0.295
    1        175.49  27    0.000    313.41   9    0.000    0.195
    2         61.70  19    0.000    113.79   8    0.000    0.124
    3          9.38  12    0.670     52.32   7    0.000    0.000
```

3. Factor Scores

To obtain factor scores for the factors in the TSLS reference variables solution, add the keyword FS on the FA command in PRELIS. Factor scores can only be obtained by PRELIS because it requires raw data.

The factor scores are computed by an extension of a formula given by Anderson & Rubin (1956). These factor scores are unbiased estimates of the factors and their sample covariance matrix is exactly equal to the estimated covariance or correlation matrix of the reference variables factors. The other two commonly used methods for estimating factor scores, *i.e.*, the regression method and Bartlett's method, do not have these properties.

[2] For *k* = 0 this is a test of the hypothesis that the variables are uncorrelated. If this hypothesis cannot be rejected, it is meaningless to do a factor analysis.

To obtain factor scores just add FS on the FA command in files **NPV7A.PRL** or **NPV7B.PRL**, see file **NPV7C.PRL**. The factor scores are saved as a plain text (ASCII) file with the same name as the input file, but with suffix FSC. Thus, in this case the factor scores will be saved in the file **NPV7C.FSC**. This can be read in free format.

Possible uses of these factor scores are

- Select subgroups of individuals on the basis of the factor scores
- Rank the individuals on the basis of the factor scores for one factor
- Correlate the factor scores with some external variable
- Compute scores for the unique (error) variables $\boldsymbol{\delta}$

Here we illustrate how to merge the factor scores with the scores on the observed variables and compute the joint correlation matrix of the observed variables and the factor scores. The PRELIS command file to do this is (see file **NPV7D.PRL**):

```
Merging Observed Variables and Factor Scores and
Computing Joint Correlation Matrix
Data NI = 9,3
Labels
'VIS PERC' CUBES LOZENGES 'PAR COMP' 'SEN COMP'
WORDMEAN  ADDITION COUNTDOT 'S-C CAPS'
Factor_1 Factor_2 Factor_3
Rawdata=NPV.RAW,NPV7C.FSC
Continuous 'VIS PERC' - Factor_3
Output MA=KM
```

Note that the correlations among the three factors are exactly the same as the correlation matrix for the factors of the TSLS reference variables solution, see any of the files **NPV2.OUT** to **NPV6.OUT**.