Scientific Software International

# Exploratory factor analysis: population example

## Contents

## 1. Introduction

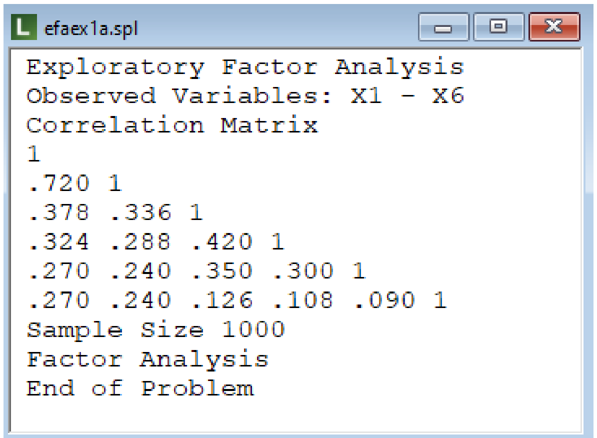
For a given set of manifest response variables one wants to find a set of underlying factors, fewer in number than the observed variables, that account for the inter-correlations of the response variables in the sense that when factors are held constant, no correlation should remain between response variables. In a factor analysis model, underlying latent factors are assumed to account for the correlations between observed variables, in contrast with principal component analysis where the components are supposed to account for maximum variance. Factor analysis typically incorporates more domain specific assumptions about the underlying structure and solves eigenvectors of a slightly different matrix. It should also be noted that factor analysis is a model which can be tested.

In this example, we a hypothetical example. The covariance matrix shown below is assumed to be a population covariance matrix rather than a sample covariance matrix as it has been specifically constructed so as to satisfy a factor analysis model with two factors.

$$\begin{bmatrix} 1.000 & & & & & \\ 0.720 & 1.000 & & & & \\ 0.378 & 0.336 & 1.000 & & & \\ 0.324 & 0.288 & 0.420 & 1.000 & & \\ 0.270 & 0.240 & 0.350 & 0.300 & 1.000 & \\ 0.270 & 0.240 & 0.126 & 0.108 & 0.090 & 1.000 \end{bmatrix}$$

## 2. Exploratory factor analysis

The FA command is used to request a factor analysis, as shown in the syntax file below. No specific number of factors are requested in this file. All files used here can be found in the **MVABOOK\Chapter6** folder. Corresponding LISREL syntax is given in the file **efaex1b.lis**.



```
Exploratory Factor Analysis
Observed Variables: X1 - X6
Correlation Matrix
1
.720 1
.378 .336 1
.324 .288 .420 1
.270 .240 .350 .300 1
.270 .240 .126 .108 .090 1
Sample Size 1000
Factor Analysis
End of Problem
```

The output file contains the following decision table. The chi-square for two factors is 0, indicating that 2 factors describe the data perfectly. This is in line with the intentional construction of the correlation matrix analyzed here.

```
Total Variance = 6.000 Generalized Variance = 0.248

 Largest Eigenvalue = 2.570 Smallest Eigenvalue = 0.277

 Condition Number = 3.044


Maximum Likelihood Factor Analysis

Decision Table for Number of Factors

Factors       Chi2    df       P        DChi2  Ddf      PD       RMSEA
-------       ----    --       -        ----- ---       --       -----
   0       1387.99    15    0.000                                0.303
   1        177.62     9    0.000     1210.37    6     0.000     0.137
   2          0.00     4    1.000      177.62    5     0.000     0.000
```

For our current example, 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**

| $k$ | $c_k$ | $d_k$ | $P_k$ | $\Delta c_k$ | $\Delta d_k$ | $P_{\Delta c}$ | $\rho_k$ |
|---|---|---|---|---|---|---|---|
| 0 | 1387.99 | 15 | 0.000 | | | | 0.303 |
| 1 | 177.62 | 9 | 0.000 | 1210.37 | 6 | 0.000 | 0.137 |
| 2 | 0.00 | 4 | 0.000 | 177.62 | 5 | 0.000 | 0.000 |

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

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

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

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

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

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

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

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

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 $\chi^2$ with $d_k$ degrees of freedom.[1] The *P*-value of this test is $P_k$ , *i.e.*, the probability that a random $\chi^2$ 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. $\rho_k$ 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 $k = 0,1,\ldots,k_{\max}$ 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.

[1] 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.

The first solution is the unrotated solution computed using the maximum likelihood procedure described by Jöreskog (1967) and in more detail by Jöreskog (1977).

```
Unrotated Factor Loadings

            Factor 1   Factor 2 Unique Var
            --------   -------- ----------
       X1      0.889     -0.138      0.190
       X2      0.791     -0.122      0.360
       X3      0.501      0.489      0.510
       X4      0.429      0.419      0.640
       X5      0.358      0.349      0.750
       X6      0.296     -0.046      0.910

Minimum Fit Function Chi-Square with 4 Degrees of Freedom = 0.00
```

The second solution is the varimax solution of Kaiser (1958). Both of these are orthogonal solutions, *i.e.*, the factors are uncorrelated.

```
Varimax-Rotated Factor Loadings

            Factor 1   Factor 2 Unique Var
            --------   -------- ----------
       X1      0.854      0.285      0.190
       X2      0.759      0.253      0.360
       X3      0.221      0.664      0.510
       X4      0.190      0.569      0.640
       X5      0.158      0.474      0.750
       X6      0.285      0.095      0.910
```

The third solution is the promax solution of Hendrickson & White (1964). This is an oblique solution, *i.e.*, the factors are correlated.

```
Promax-Rotated Factor Loadings

            Factor 1   Factor 2 Unique Var
            --------   -------- ----------
       X1      0.867      0.059      0.190
       X2      0.771      0.052      0.360
       X3      0.014      0.692      0.510
       X4      0.012      0.593      0.640
       X5      0.010      0.494      0.750
       X6      0.289      0.019      0.910

      Factor Correlations

                  X1         X2
            --------   --------
               1.000
               0.541      1.000
```

The varimax and the promax solutions are transformations of the unrotated solution and as such they are still maximum likelihood solutions. Finally, a reference variables rotation with a factor correlation of 0.6 is given.

```
Reference Variables Factor Loadings Estimated by TSLS

            Factor 1   Factor 2 Unique Var
            --------   -------- ----------
       X1      0.900      0.000      0.190

       X2      0.800     -0.000      0.360
              (0.13)     (0.12)
               6.096     -0.001

       X3      0.000      0.700      0.510

       X4     -0.000      0.600      0.640
              (0.08)     (0.14)
              -0.000      4.302

       X5     -0.000      0.500      0.750
              (0.07)     (0.11)
              -0.000      4.354

       X6      0.300     -0.000      0.910
              (0.07)     (0.10)
               4.462     -0.000


         Factor Correlations

                  X1         X2
            --------   --------
               1.000
               0.600      1.000
```