Class notes 9

Exploratory factor analysis

Covariance and correlation

Consider the two random variables X_1 and X_2. Let X_1 and X_2 be standardized in the following way

$$Z_1 = \frac{X_1 - \mu_{X_1}}{Std(X_1)}$$

$$Z_2 = \frac{X_2 - \mu_{X_2}}{Std(X_2)}$$

where μ_{X_1} and μ_{X_2} are the means of X_1 and X_2, respectively. Similarly, $Std(X_1)$ and $Std(X_2)$ express the standard deviations of X_1 and X_2. Recall that a standardized variable have mean of zero and variance of one. We can show that the correlation between the two variables can be written as

$$Corr(X_1, X_2) = \frac{Cov(X_1, X_2)}{Std(X_1) \cdot Std(X_2)} = Cov(Z_1, Z_2)$$

It is clear from the expression that the correlation is simply the covariance between standardized variables. The correlation value is bounded in the following way

$$-1 \leq Corr(X_1, X_2) \leq 1$$

where $Corr(X_1, X_2) = -1$ is perfect negative correlation, $Corr(X_1, X_2) = 0$ is no correlation and $Corr(X_1, X_2) = 1$ represents perfect positive correlation.

The covariance and correlation matrices

It is useful to consider an example with a smaller number of variables. Let X_1, X_2, X_3, X_4 and X_5 denote five random variables. Then the covariance matrix of $X_1, \ldots, X_5$ is given by

$$\mathbf{\Sigma} = \begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) & Cov(X_1, X_4) & Cov(X_1, X_5) \\ Cov(X_2, X_1) & Var(X_2) & Cov(X_2, X_3) & Cov(X_2, X_4) & Cov(X_2, X_5) \\ Cov(X_3, X_1) & Cov(X_3, X_2) & Var(X_3) & Cov(X_3, X_4) & Cov(X_3, X_5) \\ Cov(X_4, X_1) & Cov(X_4, X_2) & Cov(X_4, X_3) & Var(X_4) & Cov(X_4, X_5) \\ Cov(X_5, X_1) & Cov(X_5, X_2) & Cov(X_5, X_3) & Cov(X_5, X_4) & Var(X_5) \end{pmatrix}$$

and the correlation matrix is

$$\boldsymbol{R} = \begin{pmatrix} 1 & Corr(X_1, X_2) & Corr(X_1, X_3) & Corr(X_1, X_4) & Corr(X_1, X_5) \\ Corr(X_2, X_1) & 1 & Corr(X_2, X_3) & Corr(X_2, X_4) & Corr(X_2, X_5) \\ Corr(X_3, X_1) & Corr(X_3, X_2) & 1 & Corr(X_3, X_4) & Corr(X_3, X_5) \\ Corr(X_4, X_1) & Corr(X_4, X_2) & Corr(X_4, X_3) & 1 & Corr(X_4, X_5) \\ Corr(X_5, X_1) & Corr(X_5, X_2) & Corr(X_5, X_3) & Corr(X_5, X_4) & 1 \end{pmatrix}$$

Note that the variance is a special case of covariance, since $Var(X_1) = Cov(X_1, X_1)$. Moreover, the covariance between two variables is a symmetric measure. That is, $Cov(X_1, X_2) = Cov(X_2, X_1)$ (the order does not matter). It then follows that the covariance matrix is a symmetric matrix. The same applies to the correlation matrix.

The two matrices are estimated from the data by

$$\boldsymbol{S} = \begin{pmatrix} s_{X_1}^2 & s_{X_1,X_2} & s_{X_1,X_3} & s_{X_1,X_4} & s_{X_1,X_5} \\ s_{X_2,X_1} & s_{X_2}^2 & s_{X_2,X_3} & s_{X_2,X_4} & s_{X_2,X_5} \\ s_{X_3,X_1} & s_{X_3,X_2} & s_{X_3}^2 & s_{X_3,X_4} & s_{X_3,X_5} \\ s_{X_4,X_1} & s_{X_4,X_2} & s_{X_4,X_3} & s_{X_4}^2 & s_{X_4,X_5} \\ s_{X_5,X_1} & s_{X_5,X_2} & s_{X_5,X_3} & s_{X_5,X_4} & s_{X_5}^2 \end{pmatrix}$$

$$\widehat{\boldsymbol{R}} = \begin{pmatrix} 1 & r_{X_1,X_2} & r_{X_1,X_3} & r_{X_1,X_4} & r_{X_1,X_5} \\ r_{X_2,X_1} & 1 & r_{X_2,X_3} & r_{X_2,X_4} & r_{X_2,X_5} \\ r_{X_3,X_1} & r_{X_3,X_2} & 1 & r_{X_3,X_4} & r_{X_3,X_5} \\ r_{X_4,X_1} & r_{X_4,X_2} & r_{X_4,X_3} & 1 & r_{X_4,X_5} \\ r_{X_5,X_1} & r_{X_5,X_2} & r_{X_5,X_3} & r_{X_5,X_4} & 1 \end{pmatrix}$$

The general factor model

The general factor analysis model with p observed variable and q factors takes the form

$$\begin{aligned} X_1 &= \mu_{X_1} + \alpha_{11}F_1 + \alpha_{12}F_2 + \cdots + \alpha_{1q}F_q + \delta_1 \\ X_2 &= \mu_{X_2} + \alpha_{21}F_1 + \alpha_{22}F_2 + \cdots + \alpha_{2q}F_q + \delta_2 \\ X_3 &= \mu_{X_3} + \alpha_{31}F_1 + \alpha_{32}F_2 + \cdots + \alpha_{3q}F_q + \delta_3 \\ &\vdots \quad \vdots \qquad \vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \qquad \vdots \\ X_p &= \mu_{X_p} + \alpha_{p1}F_1 + \alpha_{p2}F_2 + \cdots + \alpha_{pq}F_q + \delta_p \end{aligned}$$

where $F_1, \dots, F_q$ are *unobserved* (or *latent*) random variables, also referred to as *common factors*. The common factors (at least at this point) are assumed standardized and mutually uncorrelated. The *specific* or *unique* factors $\delta_1, \dots, \delta_p$ are assumed to have zero mean and are mutually uncorrelated. The unique factors are additionally assumed to be uncorrelated with the common factors. The parameters in the system, given by $\alpha_{11}, \dots, \alpha_{pq}$, are so-called *factor loadings*. These loadings represent the relationships between the common factors and the observed variables. The right-hand-side of the equation system describes an underlying factor structure responsible for the correlation among the Xs. The model is such that the number of observed variables is larger than number of factors ($p > q$).

The model may also be written in terms of the standardized variables

$$
\begin{aligned}
Z_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1q}F_q + \delta_1 \\
Z_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2q}F_q + \delta_2 \\
Z_3 &= a_{31}F_1 + a_{32}F_2 + \cdots + a_{3q}F_q + \delta_3 \\
\vdots \quad & \quad \vdots \qquad \vdots \qquad\qquad \vdots \qquad \vdots \\
Z_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pq}F_q + \delta_p
\end{aligned}
$$

If a standardized solution is considered, the correlation matrix is used as input when performing the analysis.

To get a better understanding of how the factor model is tied to the correlation matrix, consider an example of four variables X_1, X_2, X_3 and X_4. The table below shows the correlation matrix

Correlations

		X1	X2	X3	X4
X1	Pearson Correlation	1	,860**	.058	.014
	Sig. (2-tailed)		.000	.197	.762
X2	Pearson Correlation	,860**	1	.002	-.043
	Sig. (2-tailed)	.000		.967	.342
X3	Pearson Correlation	.058	.002	1	,741**
	Sig. (2-tailed)	.197	.967		.000
X4	Pearson Correlation	.014	-.043	,741**	1
	Sig. (2-tailed)	.762	.342	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

In the table, one observe large significant correlations between X_1 and X_2 (0.86) and between X_3 and X_4 (0.74), whereas the remaining correlation values are small and insignificant. These observations suggest that a common source of variation is responsible for the correlation between X_1 and X_2, and a second common source is responsible for the correlation between X_3 and X_4. Thus, the pattern is indicative of a model of the form

$$
\begin{aligned}
Z_1 &= a_{11}F_1 + a_{12}F_2 + \delta_1 \\
Z_2 &= a_{21}F_1 + a_{22}F_2 + \delta_2 \\
Z_3 &= a_{31}F_1 + a_{32}F_2 + \delta_3 \\
Z_4 &= a_{41}F_1 + a_{42}F_2 + \delta_4
\end{aligned}
$$

where a_{12}, a_{22}, a_{31} and a_{41} are all close to zero.

Note that the correlation matrix in this example was constructed for the purpose of these notes. In practice, it may be difficult (if not impossible) to specify the structure of the factor model solely based on the correlations obtained from the correlation matrix.

Decomposing the correlation matrix

Factor analysis involves making choices about how many factors to include (choosing q), and which method to use when computing the loadings. It is therefore useful to introduce some additional concepts related to the correlation matrix (for the remaining of these notes, we will limit the analysis to only consider the correlation matrix).

A correlation matrix has a number of properties that are of interest to us. One property is that it can be decomposed. By applying a procedure known as *spectral decomposition*, one can write (after some mathematical manipulation) the observed (standardized) variables as a system of equations

$$\begin{aligned}
Z_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1p}F_p \\
Z_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2p}F_p \\
Z_3 &= a_{31}F_1 + a_{32}F_2 + \cdots + a_{3p}F_p \\
\vdots \quad & \quad \vdots \qquad\quad \vdots \qquad\qquad\quad \vdots \\
Z_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pp}F_p
\end{aligned}$$

The system of equations is different from the factor model stated previously. Firstly, the number of factors in the system is the same as the number of observed variables, and secondly, there are no unique factors present. It is useful to organize the loadings in a matrix in the following way

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}$$

Then, since the factors are standardized and mutually uncorrelated, the correlation matrix can be written in the simple form

$$\boldsymbol{R} = \boldsymbol{A}\boldsymbol{A}'$$

Eigenvalues

The procedure for decomposing the correlation matrix additionally provides a set of p (the number of observed variables) so-called *eigenvalues*. The eigenvalues are useful since they reveal information about the dimensionality of the data. If for instance, the largest eigenvalue is much larger than the smallest eigenvalue, the dimensionality is considerably smaller than p. If, on the other hand, the eigenvalues are near equal, the dimensionality is close to p.

A related notion is that the eigenvalues are informative about how much variation the individual factors contributes to the overall variation in the data. For instance, if the eigenvalue associated with the first factor is 2 and the overall variation in the data is 4, then the contribution of the first factor is $100(2/4)\% = 50\%$. The overall variation is here defined as the sum of the diagonal elements of the matrix being analyzed. In case of the correlation matrix, the overall variation is simply p (since we have ones on the diagonal).

The eigenvalues can be obtained from the elements of $\boldsymbol{A}$ in the following way

$$\begin{aligned}
\text{The largest eigenvalue} &= a_{11}^2 + a_{21}^2 + \cdots + a_{p1}^2 \\
\text{The second largest eigenvalue} &= a_{12}^2 + a_{22}^2 + \cdots + a_{p2}^2 \\
\vdots \quad & \vdots \quad \vdots \quad \vdots \quad \quad \vdots \\
\text{The } p\text{th largest eigenvalue} &= a_{1p}^2 + a_{2p}^2 + \cdots + a_{pp}^2
\end{aligned}$$

The equations are also known as the *sum of the squared loadings* (SSL).

Another, interesting property of eigenvalues is that their sum is equal to p (when analyzing the correlation matrix), from which it also follows that their average is 1.

Uniqueness

The system of equations presented before is not unique in the sense that other sets of loadings lead to the same correlation matrix. In fact, an infinite sets of loadings can be found. We briefly mention that the lack of uniqueness makes factor analysis a subject of controversy. To obtain a different solution, one can transform the initial solution (here represented by the matrix $\boldsymbol{A}$) by a procedure known as *rotation.*

Let the rotated solution be given by

$$\begin{aligned}
Z_1 &= b_{11}F_1 + b_{12}F_2 + \cdots + b_{1p}F_p \\
Z_2 &= b_{21}F_1 + b_{22}F_2 + \cdots + b_{2p}F_p \\
Z_3 &= b_{31}F_1 + b_{32}F_2 + \cdots + b_{3p}F_p \\
\vdots \quad & \quad \vdots \quad \quad \vdots \quad \quad \quad \vdots \\
Z_p &= b_{p1}F_1 + b_{p2}F_2 + \cdots + a_{pp}F_p
\end{aligned}$$

Again, we organize the coefficients in a loadings matrix

$$\boldsymbol{B} = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pp} \end{pmatrix}$$

As before, the correlation matrix is obtained by

$$\boldsymbol{R} = \boldsymbol{A}\boldsymbol{A}' = \boldsymbol{B}\boldsymbol{B}'$$

Since the rotated solution gives the same correlation matrix, it is equally valid as compared to the previous solution. Note that there are many procedures for rotation, providing different, but equally valid factor solutions. Later, we discuss the purpose of rotation and what can be accomplished by performing such rotation.

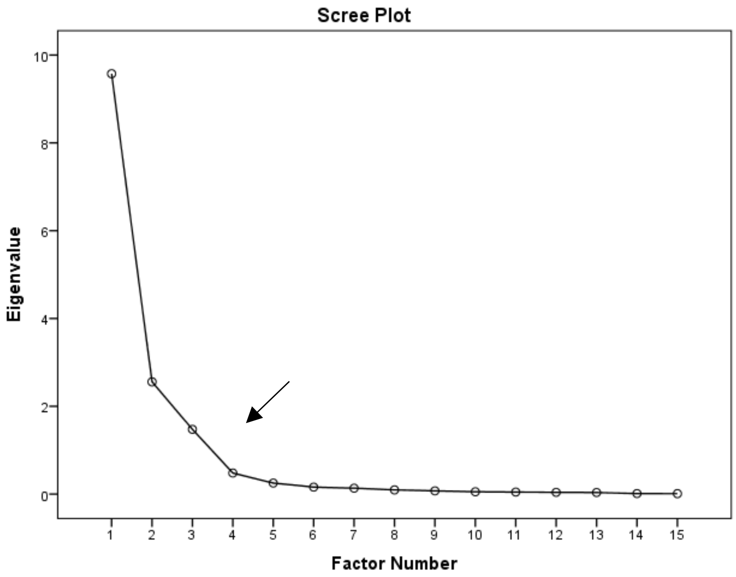
Specifying the factor model

How many factors to include

The main objective of factor analysis is to summarize the data into a smaller number of factors, and eventually arrive at a solution for which the factors can be given a meaningful interpretation.

To achieve this, an appropriate choice of the number of factors must be made. A possible approach is to let the choice be based on the size of the eigenvalues. A popular criteria is to choose the number of factors according to the number of eigenvalues larger than 1 (the average of the eigenvalues). This criteria works well in many cases. When it fails, it tend to give too many factors.

Another approach is to evaluate a so-called scree plot. The scree plot is a plot of the eigenvalues against the factor number (or component number). An example is given below



In the scree plot, one should look for the spot at which the curve turns flat (here indicated by the arrow). In this example, three factors seem appropriate. The three largest eigenvalues are 9.58, 2.56 and 1.48. The combined contribution of the first three factors to the overall variation is $100((9.58 + 2.56 + 1.48)/15)\% = 90.71\%$. Thus, quite some reduction in dimensionality is possible by a three-factor solution.

Although these approaches are helpful, one still needs to apply some human judgement in determining the appropriate number of factors. One may have to try a number of different solutions before reaching one that has a meaningful interpretation. The different solutions are evaluated based on the size of the loadings and how the observed variables group. For instance, if a factor has only small loadings, say 0.3 or less, we may try a solution with one factor less.

Communalities

For each equation in the system, a so-called *communality* is computed. A communality represents how much variance in the individual variables that is due to the q factors. The expression for the communality associated with Z_j (for $j = 1, \dots, p$) is

$$Comm_j = a_{j1}^2 + a_{j2}^2 + \cdots + a_{jq}^2 = 1 - Var(\delta_j)$$

where $0 \leq Comm_j \leq 1$. It is desirable that the communalities are large (close to one). The total variance due to the q-factor solution is the sum of the communalities.

Factor rotation

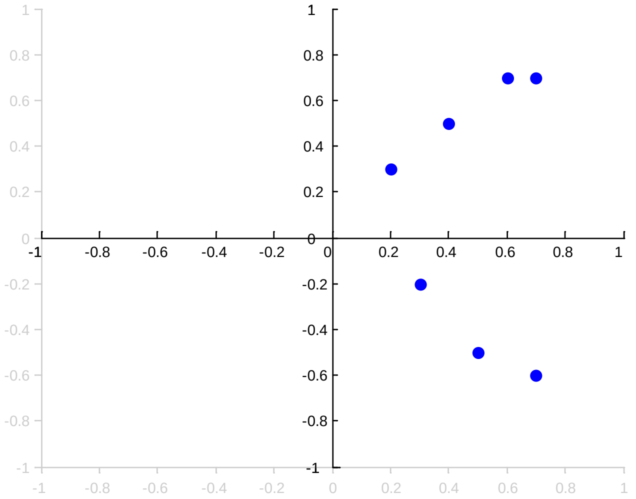
Rotation is an essential part of performing factor analysis. The purpose of rotation is to facilitate/improve the interpretability of the factor solution. As we just saw, not only one factor solution is valid for describing the correlation matrix. Rotation works by transforming the initial solution such that the observed variables load strongly on some factors and weakly on some other.

To illustrate the effect of rotation (in this case orthogonal rotation), consider an example of 7 observed variables and two factors. The table below shows the factors loadings before and after rotation

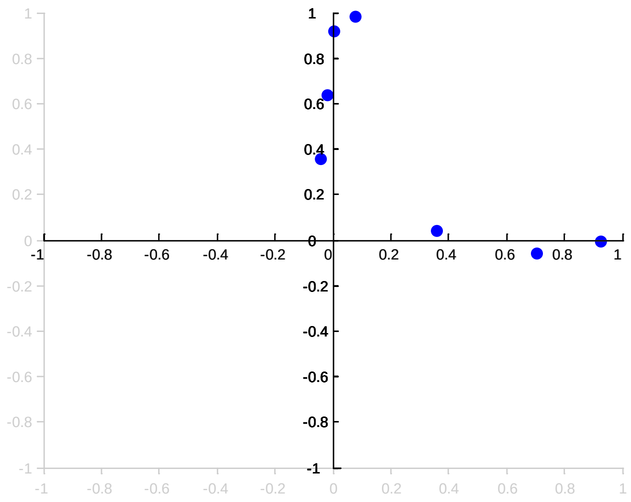
	Unrotated		Rotated (45°)	
	F1	F2	F1	F2
Z1 :	0.2	0.3	-0.043	0.36
Z2 :	0.4	0.5	-0.021	0.64
Z3 :	0.6	0.7	0.001	0.92
Z4 :	0.7	0.7	0.077	0.99
Z5 :	0.5	-0.5	0.705	-0.06
Z6 :	0.7	-0.6	0.922	0.00
Z7 :	0.3	-0.2	0.358	0.04

First, consider the loadings associated with factor 1. For $Z_1, \dots, Z_4$, the loadings are closer to zero after rotating the solution, while for $Z_5, \dots, Z_7$, the loadings are larger. This pattern indicates that factor 1 is the common (underlying) source of variation responsible for the correlation among $Z_5, \dots, Z_7$. In a similar fashion, we find that factor 2 is the common source of variation responsible for the correlation among $Z_1, \dots, Z_4$.

The two solutions are illustrated below. In the diagrams, the x-axis represents factor 1 and the y-axis represents factor 2. The solution is rotated exactly 45° degrees. Note that that the relative position of the points is not affected by the rotation.



Unrotated



Rotated

Two types of rotation:

1. *Orthogonal rotation*: the factors remain uncorrelated (just as in the unrotated solution). A popular method for orthogonal rotation is the *Varimax* procedure. The expression for how to compute the communalities remain the same as before.

2. *Oblique rotation*: when applying this form of rotation, one allows the factors to be correlated. As part of the solution, a factor correlation matrix is obtained. A popular method for oblique rotation is *Promax*. The expression for how to compute the communalities is not the same as before because of the factor correlation. Thus,

$$Comm_j \neq a_{j1}^2 + a_{j2}^2 + \cdots + a_{jq}^2$$

The effect of rotation (either orthogonal or oblique) is summarized in the following way:

- Rotation alters the interpretation of the factors (which is in fact what we want).
- The contribution (in terms of explained variance) by each factor (SSL) will change.
- The contribution (in terms of explained variance) by the factors taken together does not change.
- The communalities does not change.

If the interpretation of the factors suggests that the factors should be correlated, we apply the oblique rotation otherwise not.

Extraction (estimation)

Various procedures are available for estimating the factor loadings.

- *Principal component estimation* (not to be confused with Principal component analysis). Estimation works by spectral decomposing the correlation matrix (as previously described). The procedure is popular due to its simplicity, and it is typically the default option in the software. The procedure does not allow for testing the overall model fit.

Some estimation procedures, such as *Maximum Likelihood* and *Least Squares*, work by first defining a measure of distance between the correlation matrix obtained from the data and the correlation matrix implied by the model (a mathematical expression). Estimation is then performed by obtaining the factor loadings that minimize this distance.

- *Maximum Likelihood*: the maximum likelihood procedure is derived from the assumption that the distribution of the observed variables is multivariate normal. For just estimating the loadings, maximum likelihood may work well even if the normality assumption is not satisfied. If our aim is also to test overall model fit, the normality assumption must be satisfied. Note that this method is more sensitive to the data as compared to principal component estimation.

- *Least Squares*: estimators belonging to the class of procedures are the *Unweighted Least Squares* and *Generalized Least Squares.*

In practice, if the data is plenty, one often finds that the various estimation procedures provide solutions that are similar.

Factor scores

After performing the factor analysis, it is possible to obtain the factor scores. The factor scores are estimated data scores. The factor scores may be of interest in other statistical analysis. The procedure for how to obtain the factor scores is further described in Tabachnick and Fidell on p. 703 and 704.