Tutorials and worked examples for simulation, curve fitting, statistical analysis, and plotting.

https://simfit.org.uk
https://simfit.silverfrost.com

Factor analysis seeks to explore the relationships between multivariate observations with m variables in terms of a set of k hypothetical factors, where $k < m$. It is widely used in social and psychological research where the factors could be things such as intelligence which are difficult to quantify and model, but it is not used much in the physical sciences where the construction of deterministic models is preferred where possible.

Example 1

From the main SIMFIT menu choose [Statistics], [Multivariate], then [Factor analysis] and read in the default test file `g03caf.tf1` which contains the following correlation matrix from a sample of 211 subjects where 9 variables were measured. Actually, due to the symmetry and unit diagonals, only the strict lower or strict upper triangle is needed, but the SIMFIT data input requires a full matrix because the factor analysis procedure can also read in a data matrix then calculate the correlation matrix interactively.

1	0.523	0.395	0.471	0.346	0.426	0.576	0.434	0.639
0.523	1	0.479	0.506	0.418	0.462	0.547	0.283	0.645
0.395	0.479	1	0.355	0.270	0.254	0.452	0.219	0.504
0.471	0.506	0.355	1	0.691	0.791	0.443	0.285	0.505
0.346	0.418	0.270	0.691	1	0.679	0.383	0.149	0.409
0.426	0.462	0.254	0.791	0.679	1	0.372	0.314	0.472
0.576	0.547	0.452	0.443	0.383	0.372	1	0.385	0.680
0.434	0.283	0.219	0.285	0.149	0.314	0.385	1	0.470
0.639	0.645	0.504	0.505	0.409	0.472	0.680	0.470	1

This matrix is discussed in the book *Factor Analysis as a Statistical Method by D.N.Lawley and E.A.Maxwell London Butterworths (2nd Edition) 1971* which must be consulted in order to understand the following results.

Results from analysis of test file *g03caf.tf1*	
Number of variables	9
Transformation	Untransformed
Matrix type	Input correlation matrix
Number of factors	3
Replicates	Unweighted for replicates
$F(\hat{\Psi})$	0.0350
Test statistic TS	7.1494
Degrees of Freedom	12 (Number of cases = 211)
$P(\chi^2 \geq TS)$	0.8476

Eigenvalues	Communalities	$\hat{\Psi}$
15.968	0.54954	0.45046
4.3577	0.57293	0.42707
1.8475	0.38345	0.61655
1.1560	0.78767	0.21233
1.1190	0.61947	0.38053
1.0271	0.82308	0.17692
0.92574	0.60046	0.39954
0.89508	0.53846	0.46154
0.87710	0.76908	0.23092

Residual correlations

```
 0.0004
-0.0128   0.0220
 0.0114  -0.0053   0.0231
-0.0100  -0.0194  -0.0162   0.0033
-0.0046   0.0113  -0.0122  -0.0009  -0.0008
 0.0153  -0.0216  -0.0108   0.0023   0.0294  -0.0123
-0.0011  -0.0105   0.0134   0.0054  -0.0057  -0.0009   0.0032
-0.0059   0.0097  -0.0049  -0.0114   0.0020   0.0074   0.0033  -0.0012
```

Factor loadings by columns

```
 0.6642  -0.3209  -0.0735
 0.6888  -0.2471  -0.1933
 0.4926  -0.3022  -0.2224
 0.8372   0.2924  -0.0354
 0.7050   0.3148  -0.1528
 0.8187   0.3767   0.1045
 0.6615  -0.3960  -0.0778
 0.4579  -0.2955   0.4914
 0.7657  -0.4274  -0.0117
```

Example 2

Test file `g03ccf.tf1` contains the following correlation matrix that is also discussed by Lawley and Maxwell. It is from an analysis of 220 students on the six subjects indicated in column 1. They suggest that "*the fact that all the correlations between the variates are positive indicates that students who get scores above average on any one of the subjects tend also to get scores above average on the other subjects.*"

Gaelic	1	0.439	0.410	0.288	0.329	0.248
English	0.439	1	0.351	0.354	0.320	0.329
History	0.410	0.351	1	0.164	0.190	0.181
Arithmetic	0.288	0.354	0.164	1	0.595	0.470
Algebra	0.329	0.320	0.190	0.595	1	0.464
Geometry	0.248	0.329	0.181	0.470	0.464	1

The next table shows the results from analysis of this correlation matrix for two factors.

Results from analysis of test file *g03ccf.tf1*

Number of variables	6
Transformation	Untransformed
Matrix type	Input correlation matrix
Number of factors	2
Replicates	Unweighted for replicates
$F(\hat{\Psi})$	0.1088
Test statistic TS	2.3346
Degrees of Freedom	4 (Number of cases = 220)
$P(\chi^2 \geq TS)$	0.6754

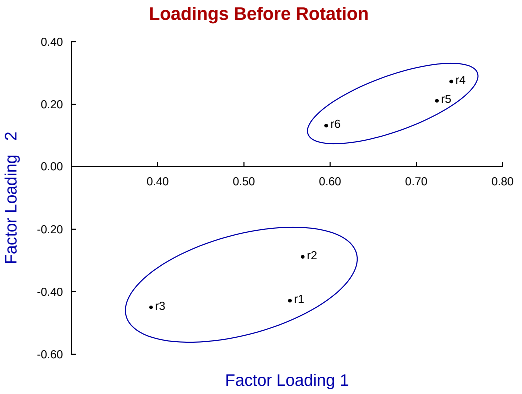
Eigenvalues	Communalities	$\hat{\Psi}$
5.6142	0.48983	0.51017
2.1428	0.40593	0.59407
1.0923	0.35627	0.64373
1.0264	0.62264	0.37736
0.9908	0.56864	0.43136
0.8905	0.37179	0.62821

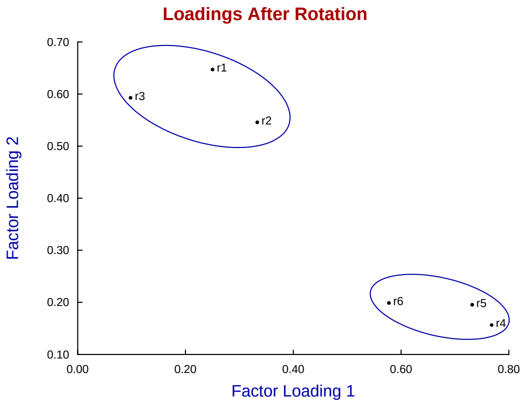
Factor loadings by columns	
0.55332	-0.42856
0.56816	-0.28832
0.39218	-0.44996
0.74042	0.27280
0.72387	0.21131
0.59536	0.13169

The score coefficients are now shown but also a further possibility should be mentioned. As the factors are only unique up to rotation, it is possible to perform a Varimax or Quartimax rotation to calculate a rotation matrix R before working out the score coefficients, which may simplify the interpretation of the observed variables in terms of the unobservable variables.

Factor score coefficients	
Method	Regression
Rotation	None
0.19318	-0.39203
0.17035	-0.22649
0.10852	-0.32621
0.34950	0.33738
0.29891	0.22861
0.16881	0.09783

The next figures illustrate the rows from the loading matrix labeled as $r1, r2, \cdots, r6$ both before and after a Varimax rotation with $\gamma = 1$ and reflection of the y-axis and indicating the presence of two clusters.





Many workers find it convenient to rotate loadings in this way until all are positive so that the relative magnitudes and potential groupings can be visualized more easily. The example illustrated above indicates that factor 2 is what is known as a bi-polar factor with approximately half positive and half negative, but that the obvious grouping is still preserved by rotation.

It should be pointed out that this procedure may also require the use of reflection of axes in order to achieve positive loadings, as in the present case where the second set of loadings were reflected by the automatic technique provided by SimFIT to do such transformations interactively.

Theory

This technique is used when it is wished to express a multivariate data set in m manifest, or observed variables, in terms of k latent variables, where $k < m$. Latent variables are variables that by definition are unobservable, such as social class or intelligence, and thus cannot be measured but must be inferred by estimating the relationship between the observed variables and the supposed latent variables. The statistical treatment is based upon a very restrictive mathematical model that, at best, will only be a very crude approximation and, most of the time, will be quite inappropriate. For instance, Krzanowski (in *W.J.Krzanowski Principles of Multivariate Analysis, Oxford, revised edition, 2000*) explains how the technique is used in the psychological and social sciences, but then goes on to state

> *At the extremes of, say, Physics or Chemistry, the models become totally unbelievable. p477*
> *It should only be used if a positive answer is provided to the question, "Is the model valid?" p503*

However, despite such warnings, the technique is now widely used, either to attempt to explain observables in terms of hypothetical unobservables, or as just another technique for expressing multivariate data sets in a space of reduced dimension. In this respect it is similar to principal components analysis, except that the technique attempts to capture the covariances between the variables, not the variances. If the observed variables x can be represented as a linear combination of the unobservable variables or factors f, so that the partial correlation $r_{ij.l}$ between x_i and x_j with f_l fixed is effectively zero, then the correlation between x_i and x_j can be said to be explained by f_l. The idea is to estimate the coefficients expressing the dependence of x on f in such a way that the the residual correlation between the x variables is a small as possible, given the value of k.

The assumed relationship between the mean-centered observable variables x_i and the factors is

$$x_i = \sum_{j=1}^{k} \lambda_{ij} f_j + e_i \text{ for } i = 1, 2, \dots, m, \text{ and } j = 1, 2, \dots, k$$

where λ_{ij} are the loadings, f_i are independent normal random variables with unit variance, and e_i are independent normal random variables with variances ψ_i. If the variance covariance matrix for x is Σ, defined as

$$\Sigma = \Lambda\Lambda^T + \Psi,$$

where Λ is the matrix of factor loadings λ_{ij}, and Ψ is the diagonal matrix of variances ψ_i, while the sample covariance matrix is S, then maximum likelihood estimation requires the minimization of

$$F(\Psi) = \sum_{j=k+1}^{m} (\theta_j - \log \theta_j) - (m - k),$$

where θ_j are eigenvalues of $S^* = \Psi^{-1/2} S \Psi^{-1/2}$. Finally, the estimated loading matrix $\hat{\Lambda}$ is given by

$$\hat{\Lambda} = \Psi^{1/2} V (\Theta - I)^{1/2},$$

where V are the eigenvectors of S^*, Θ is the diagonal matrix of θ_i, and I is the identity matrix.

The proportion of variation for each variable x_i accounted for by the k factors is the communality $\sum_{j=1}^{k} \lambda_{ij}^2$, the Psi-estimates are the variance estimates, and the residual correlations are the off-diagonal elements of

$$C - (\Lambda\Lambda^T + \Psi)$$

where C is the sample correlation matrix. If a good fit has resulted and sufficient factors have been included, then the off-diagonal elements of the residual correlation matrix should be small with respect to the diagonals (listed with arbitrary values of unity to avoid confusion). Subject to the normality assumptions of the model, the minimum dimension k can be estimated by fitting sequentially with $k = 1$, $k = 2$, $k = 3$, and so on, until the likelihood ratio test statistic

$$TS = [n - 1 - (2m + 5)/6 - 2k/3]F(\hat{\Psi})$$

is not significant as a chi-square variable with $[(m - k)^2 - (m + k)]/2$ degrees of freedom. Note that data for factor analysis can be input as a general n by m multivariate matrix, or as either a m by m covariance or correlation matrix. However, if a square covariance or correlation matrix is input then there are two further considerations: the sample size must be supplied independently, and it will not be possible to estimate or plot the sample scores in factor space, as the original sample matrix will not be available.

It remains to explain the estimation of scores, which requires the original data of course, and not just the covariance or correlation matrix. This involves the calculation of a m by k factor score coefficients matrix Φ, so that the estimated vector of factor scores $\hat{f}$, given the x vector for an individual can be calculated from

$$\hat{f} = x^T \Phi.$$

However, when calculating factor scores from the factor score coefficient matrix in this way, the observable variables x_i must be mean centered, and also scaled by the standard deviations if a correlation matrix has been analyzed. The regression method uses

$$\Phi = \Psi^{-1}\Lambda(I + \Lambda^T \Psi^{-1}\Lambda)^{-1},$$

while the Bartlett method uses

$$\Phi = \Psi^{-1}\Lambda(\Lambda^T \Psi^{-1}\Lambda)^{-1}.$$