Exploratory Factor Analysis Using SAS

In this document I explain how to use SAS to run exploratory factor analyses.

The data from this study are based on the Attitudes toward Scientists data from chapter 12 in the text. These data represent scores of 1974 respondents on the nine items shown on page 311 of the text and on the last page of this document. The data are in the file "Scientist. sas7bdat."

In SAS, factor and component analysis are obtained through **proc factor**. The default extraction method is principal components analysis; to obtain factor analysis results an extraction method keyword such as **uls, gls, ml**, or **prinit** must be specified.

```
proc factor data=Chap12.scientist  corr  residuals method=prinit  priors=smc
        nfactors=2  msa  rotate=oblimin  plot=scree;
var alone better boring nofun good help odd norelign nointrst;
run;
```

Prinit stands for iterated principal factors. This specification, combined with the specification **priors=smc**, yields the principal factors procedure described in the text. The specification **priors=smc** indicates that squared multiple correlations should be used for the initial communalities. These are then iterated in the iterated principal factors procedure.

The specification **corr** causes the correlation matrix to be printed in the output.

Residuals causes the residual correlation matrix to be printed.

Nfactors=2 specifies that 2 factors should be extracted.

MSA causes the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy to be printed.

Rotate=oblimin specifies an oblique oblimin rotation. Other rotation options can be found in the SAS help and documentation, and include varimax, quaritmax, promax, and quartimax.

Plot=scree causes the scree plot to be printed.

Running the SAS syntax above produces a great deal of output. In the sections below, I discuss selected tables.

The first part of the output is the matrix of correlations among the variables:

Correlations									
	alone	**better**	**boring**	**nofun**	**good**	**help**	**odd**	**norelign**	**nointrst**
alone	1.00000	0.07608	0.25161	0.25995	0.03105	-0.00971	0.21595	0.09581	0.35120
better	0.07608	1.00000	-0.13612	0.11342	0.46955	0.42452	0.13778	0.06718	0.12297
boring	0.25161	-0.13612	1.00000	0.26053	-0.13093	-0.23079	0.30320	0.08997	0.28647
nofun	0.25995	0.11342	0.26053	1.00000	0.04598	0.00661	0.38832	0.26424	0.45358
good	0.03105	0.46955	-0.13093	0.04598	1.00000	0.40162	0.02874	0.05136	0.04904
help	-0.00971	0.42452	-0.23079	0.00661	0.40162	1.00000	0.02741	0.04555	0.00594
odd	0.21595	0.13778	0.30320	0.38832	0.02874	0.02741	1.00000	0.31150	0.50361
norelign	0.09581	0.06718	0.08997	0.26424	0.05136	0.04555	0.31150	1.00000	0.29177
nointrst	0.35120	0.12297	0.28647	0.45358	0.04904	0.00594	0.50361	0.29177	1.00000

Although there are several pairs of variables with moderate correlations, it is difficult to see an overall pattern. This is where factor analysis can help us.

Shown below is Kaiser's Measure of Sampling Adequacy, usually known as the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. SAS provides a measure of sampling adequacy (MSA) value for each variable as well as an overall value.

Values range from 0 to 1, with higher values indicating greater amenability to factoring. According to Kaiser's criteria, the overall value of .750 shown below is between "middling" and "meritorious." The "middling" KMO value is not surprising, as correlations among variables measured on a four-point Likert scale, as these variables are, will be somewhat attenuated in comparison to correlations among variables measured on more continuous scales.

Kaiser's Measure of Sampling Adequacy: Overall MSA = 0.74997298								
alone	**better**	**boring**	**nofun**	**good**	**help**	**odd**	**norelign**	**nointrst**
0.78647147	0.67669225	0.76507770	0.82178701	0.68482477	0.70722209	0.76301243	0.80854021	0.75980570

The "Eigenvalues of the Reduced Correlation Matrix" are shown in the table below. These are the values obtained after extraction of the two specified factors. They are based on the reduced correlation matrix with communalities, rather than values of 1, on the diagonal.

The fact that these values were not obtained from the full correlation matrix with ones on the diagonal is the reason that the last few eigenvalues are negative. This pattern of negative values is typical of a reduced correlation matrix.

Eigenvalues of the Reduced Correlation Matrix: Total = 3.29879052 Average = 0.36653228				
	Eigenvalue	Difference	Proportion	Cumulative
1	1.92364790	0.54803355	0.5831	0.5831
2	1.37561435	1.20622899	0.4170	1.0001
3	0.16938536	0.11828951	0.0513	1.0515
4	0.05109585	0.03137110	0.0155	1.0670
5	0.01972476	0.02170284	0.0060	1.0730
6	-.00197809	0.01089331	-0.0006	1.0724
7	-.01287139	0.07253819	-0.0039	1.0685
8	-.08540958	0.05500906	-0.0259	1.0426
9	-.14041864		-0.0426	1.0000

The first two eigenvalues are much larger than the others, suggesting that there are 2 factors in the data.

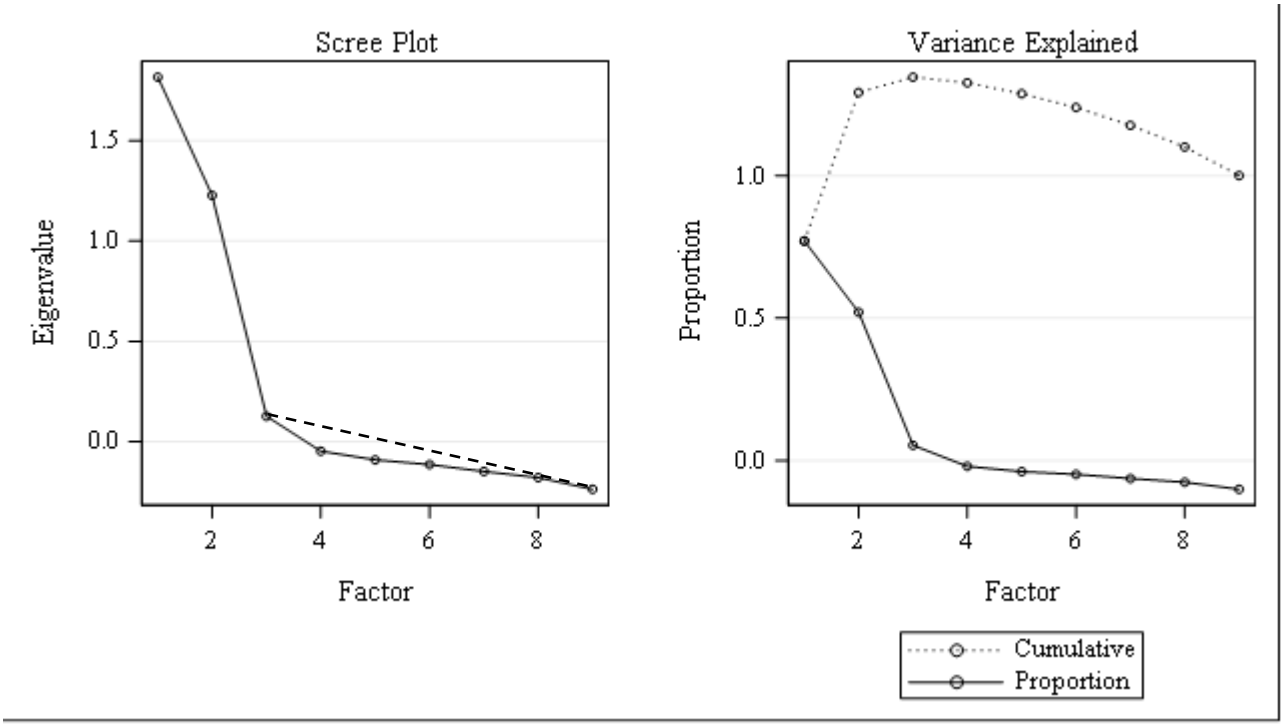
The proportion of variance accounted for, shown in the third column, is obtained by dividing each eigenvalue by the sum of the eigenvalues in the first column (approximately 3.3). This sum is not 9 (the number of variables) as it would be had the eigenvalues been obtained from the full (unreduced) correlation matrix.

The eigenvalues from the reduced matrix represent the *shared*, rather than the *total*, variance. The proportions in the third column therefore represent the proportion of the *shared,* rather than the *total* variance accounted for by each factor. Similarly, the cumulative values in the last column represent the proportion of cumulative shared variance represented by a given number of factors.

The final communalities, based on extraction of two factors, are shown below. The variables “alone” and “norelign” have the lowest communality values, consistent with the low values of their correlations with the other variables that can be seen in the variable correlation matrix.

Final Communality Estimates: Total = 3.299262								
alone	better	boring	nofun	good	help	odd	norelign	nointrst
0.17450207	0.50402473	0.27968430	0.37230642	0.40780376	0.40533885	0.44099334	0.14979504	0.56481376

The scree plot graphs the eigenvalue for each factor. SAS graphs both the eigenvalues (left-hand side) and the variance explained by each factor (right-hand side). The latter explained variance values show a similar pattern to the eigenvalues because they are simply the eigenvalues divided by their sum, as explained previously.



The number of factors is determined as the number before the line based on the eigenvalues levels off to become relatively straight. I have superimposed a dashed straight line along the eigenvalue line beginning at factor 3. Although the dots representing factors 3 – 9 do not fall directly on the straight line, they are fairly close. However, we may wish to examine a 3 factor solution.

One way of determining whether the correct number of factors has been extracted is to examine a matrix of residual correlations. These residuals are the differences between the observed, or actual, correlations and the correlations reproduced from the factor model (see pp.305-306 in the text for an explanation of the calculations for reproduced correlations).

If the number of factors extracted is incorrect, the factor model (in our example, a 2-factor model) will not be able to reproduce all the correlations sufficiently, and there will be some large residuals.

Some researchers use rough rules of thumb based on residual correlations to assess whether the number of factors is adequate. For example, if no more than 10% of the residuals correlations are greater than .05, the number of factors may be considered adequate.

A more useful way to use the residual correlations to assess model fit is to examine the variable pairs with large residuals in an attempt to determine why the model was unable to account for the observed correlation. Examples of this process are provided in Chapter 13 for confirmatory factor analysis, and the same can be done for exploratory factor analyses.

The reproduced correlations are shown in the off-diagonal and the uniquenesses (1-communality) are shown on the diagonal. Overall, the residual correlations for the 2-factor model are quite small, indicating 2 factors are probably sufficient.

Residual Correlations With Uniqueness on the Diagonal									
	alone	**better**	**boring**	**nofun**	**good**	**help**	**odd**	**norelign**	**nointrst**
alone	0.82550	0.01862	0.06646	0.00561	0.01567	0.00272	-0.06073	-0.06318	0.03776
better	0.01862	0.49598	0.00386	0.00181	0.02045	-0.01434	0.01338	-0.03416	-0.00999
boring	0.06646	0.00386	0.72032	0.00221	0.03669	-0.03412	0.02343	-0.05842	-0.03377
nofun	0.00561	0.00181	0.00221	0.62769	-0.00195	-0.00072	-0.01686	0.02970	-0.00497
good	0.01567	0.02045	0.03669	-0.00195	0.59220	-0.00274	-0.02613	-0.00750	-0.00582
help	0.00272	-0.01434	-0.03412	-0.00072	-0.00274	0.59466	0.01671	0.01213	0.00110
odd	-0.06073	0.01338	0.02343	-0.01686	-0.02613	0.01671	0.55901	0.05605	0.00459
norelign	-0.06318	-0.03416	-0.05842	0.02970	-0.00750	0.01213	0.05605	0.85020	0.00320
nointrst	0.03776	-0.00999	-0.03377	-0.00497	-0.00582	0.00110	0.00459	0.00320	0.43519

The pattern and structure matrices are shown next. Values in the pattern matrix are the correlations of the variables with each factor, holding constant or partialing out all other factors. For example, the pattern loading of alone with factor 1 is its correlation with factor 1, holding constant its correlation with factor 2.

Rotated Factor Pattern (Standardized Regression Coefficients)		
	Factor1	Factor2
alone	0.41778	-0.00230
better	0.12533	0.69591
boring	0.44834	-0.29102
nofun	0.60822	0.03666
good	0.02562	0.63749
help	-0.04092	0.63629
odd	0.66158	0.04417
norelign	0.37939	0.06826
nointrst	0.74966	0.03859

Values in the structure matrix are the correlations of the variables with the factors. For these coefficients, other factors are not held constant.

In this example, values of the pattern and structure coefficients are very similar. This is because of the low (.023) correlation between the 2 factors. This indicates that the factors are essentially uncorrelated, so partialing out the other factor has very little effect.

Factor Structure (Correlations)		
	Factor1	Factor2
alone	0.41773	0.00733
better	0.14138	0.69880
boring	0.44163	-0.28068
nofun	0.60907	0.05069
good	0.04033	0.63808
help	-0.02624	0.63535
odd	0.66260	0.05943
norelign	0.38097	0.07701
nointrst	0.75055	0.05588

Finally, the factor correlation matrix shows the very low correlation between the 2 factors. This suggests that an orthogonal rotation would have been appropriate for these data. However, oblique rotations allow the factors to be correlated as much or little as needed, so can accommodate both correlated and uncorrelated factors.

Inter-Factor Correlations		
	Factor1	**Factor2**
Factor1	1.00000	0.02307
Factor2	0.02307	1.00000

Attitudes Toward Scientists items

1. A scientist usually works alone. (ALONE)
2. Scientific researchers are dedicated people who work for the good of humanity (GOOD)
3. Scientists don't get as much fun out of life as other people do. (NOFUN)
4. Scientists are helping to solve challenging problems. (HELP)
5. Scientists are apt to be odd and peculiar people. (ODD)
6. Most scientists want to work on things that will make life better for the average person. (BETTER)
7. Scientists are not likely to be very religious people. (NORELIGN)
8. Scientists have few interests other than their work. (NOINTRST)
9. A job as a scientist would be boring. (BORING)