Advanced Methods for Determining the Number of Factors

Horn's (1965) *Parallel Analysis* (PA) is an adaptation of the Kaiser criterion, which uses information from random samples. The rationale underlying PA is that factors from real data with a valid underlying factor structure should have larger Eigenvalues than those derived from random data having the same sample size and number of variables. In the first step of PA, 1.000 datasets are randomly generated which have the same number of observations and number of variables as the original dataset. Factor analysis is then run on each of the 1.000 datasets, resulting in 1.000 sets of Eigenvalues (each set includes as many Eigenvalues as there are variables in the original dataset). Next, the 95th percentile is calculated for the largest second largest, third largest etc. Eigenvalue in the set. Researchers can now plot the Eigenvalues from the original dataset and the 95th percentile Eigenvalues from the random data. As a consequence, researchers should retain only those factors whose Eigenvalues are greater than the Eigenvalues from the random data.

Similar to the PA is the *Broken-Stick* (B-S) criterion which is also based on randomly generated Eigenvalues. The rationale underlying B-S can be best explained by the following metaphor: if a stick is randomly broken into p pieces (where p is the number of variables in the data set), b1 would represent the mean size of the largest piece in each set of broken sticks, b2 would represent the mean size of the second largest piece, and so forth. Accordingly, the B-S model assumes that if the total variance in the data is randomly divided among all factors, the expected distribution of the Eigenvalues will follow a B-S distribution. Therefore, researchers should always extract those factors from the dataset whose Eigenvalues exceed those generated by the B-S model.

Velicer (1976) suggested the Minimum Average Partial (MAP) test, which is based on the average partial correlations between the variables after successively removing the effect of the factors. The factor with the highest Eigenvalue is removed first and its effect on the correlations between the items is excluded. Next, the factor with the second highest Eigenvalue is excluded and so on. In each step, the (squared) average partial correlations between the items are computed which will initially decrease but, after several steps, will increase again. Researchers should retain the number of factors from the data, which minimizes the (squared) average partial correlations.

You might ask yourself which of the procedures works best. Some studies suggests that the PA and MAP approaches are the best when deciding on the number of factors as their results are accurate and show little variation (Zwick and Vellicer, 1986, Henson and Roberts 2006). The Kaiser criterion is known to overestimate the actual number of factors. Considering that many statistical packages, such as IBM SPSS Statistics, rely on the Kaiser criterion, you should use multiple procedures (classic as well as these more advanced techniques including PA and

MAP) to get an indication of the best number of factors. The final decision should - in line with factor analysis' exploratory character - always also consider whether the solution can be readily interpreted.

Let's use these procedures on our soccer fan satisfaction data. Go to http://www.guide-market-research.com/chapters/chapter8.html and save the file *Syntax number of factors.zip* to your computer which includes the three syntax files *Parallel Analysis.sps*, *Broken Stick.sps*, and *MAP Test.sps*. Next, go to► File ► Open ► Syntax, go to the download folder and select Parallel Analysis.sps. This will open a window similar to Fig. A8.1 which shows the syntax of the Parallel Analysis procedure.

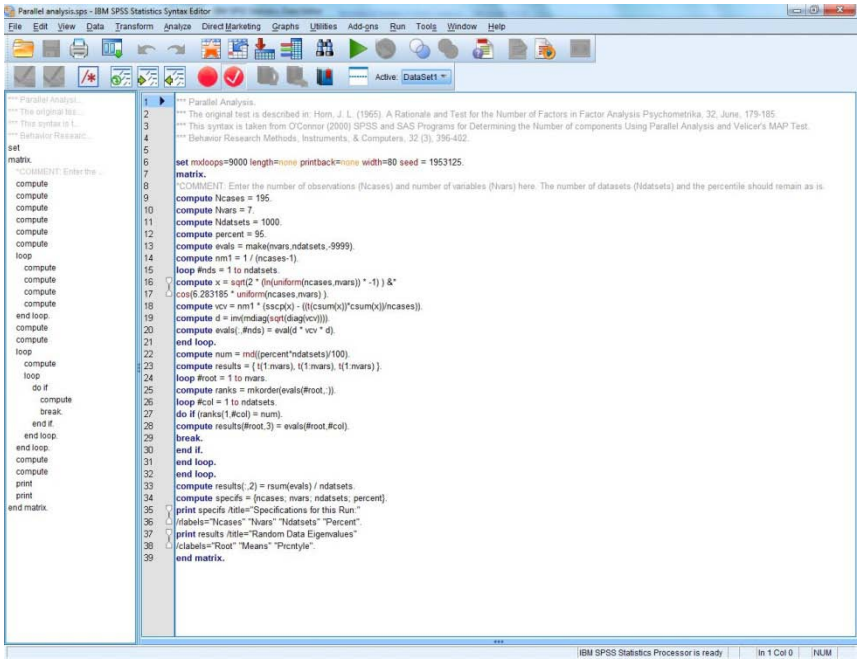

Fig. A8.1 Syntax window for the Parallel Analysis

Under **compute Ncases = 195.** (line 9), you need to specify the number of observations in the original dataset. Similarly, under **compute Nvars = 7.** (line 10), you need to specify the number of variables. As the numbers correspond to our original dataset *soccer_fan_satisfaction.sav*, we can simply leave the syntax as is. To run the syntax go to ► Run ► All which will produce an output similar to Fig. A8.2.

```
Run MATRIX procedure:

Specifications for this Run:
Ncases     195
Nvars        7
Ndatsets  1000
Percent     95

Random Data Eigenvalues
          Root         Means      Prcntyle
   1.000000000   1.275437613   1.371961515
   2.000000000   1.159887270   1.231234219
   3.000000000   1.069624900   1.123115085
   4.000000000    .991713955   1.041706752
   5.000000000    .918806320    .970513262
   6.000000000    .838245959    .896949710
   7.000000000    .746283983    .818614133

------ END MATRIX -----
```

Fig. A8.2 Parallel Analysis output

In the column labeled **Prcntyle**, we can see the 95th percentile for each of the factors. Note that the numbers are almost certainly going to look different in your analysis as their computation is based on a random process. We can now plot the percentile values against the original Eigenvalues from our factor analysis. Fig A8.3 shows such a plot using Excel. As we can see, for the first two factors, the Eigenvalues from the random data are clearly lower than those from the original data. For three factors, however, the picture is not as clear-cut. Comparing the numerical values, we learn that the 95th percentile Eigenvalue of the third factor from the random data (1.123) is slightly below the Eigenvalue of the third factor from the original data (1.135). Therefore, based on the Parallel Analysis results, we would opt for a three-factor solution.

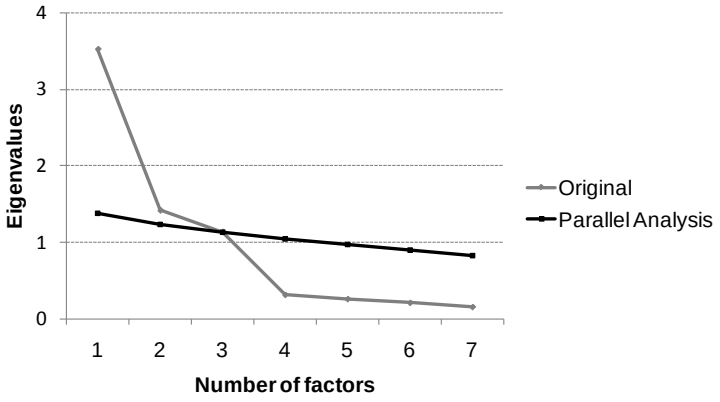

Fig. A8.3 Parallel Analysis results

Let's now run the B-S procedure by opening the corresponding syntax file (*Broken Stick.sps*). Taking a closer look at the syntax file, we can see that there are two figures which we need to adjust according to our dataset. In line 24 where it says **SAMPLE 6 from 10000.**, we need to indicate one minus the number of variables in our dataset (i.e., 7-1=6) as described in the comment just below. Similarly, in line 68 (**COMPUTE eigenvalue=percentage*7**), we need to indicate the number of variables (this time, however, not minus one). Once you are done (again, in the case of our soccer fan satisfaction example, there are no adjustments necessary), click ► Run ► All.

IBM SPSS Statistics will open an output window (Figure A8.4) which shows the Eigenvalues generated from the B-S analysis (the program also sets up a new .sav file, which contains the same information). Again, the numbers in your analysis will likely look different because of the random nature of the approach. Analogous to the Parallel analysis, you can now plot these Eigenvalues against those from the original analysis. Fig. A8.5 shows the corresponding results.

eigenvalue

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	.25	1	14.3	14.3	14.3
	.36	1	14.3	14.3	28.6
	.48	1	14.3	14.3	42.9
	.80	1	14.3	14.3	57.1
	.98	1	14.3	14.3	71.4
	1.13	1	14.3	14.3	85.7
	3.01	1	14.3	14.3	100.0
	Total	7	100.0	100.0	

Fig. A8.4 Results of the B-S analysis (I)

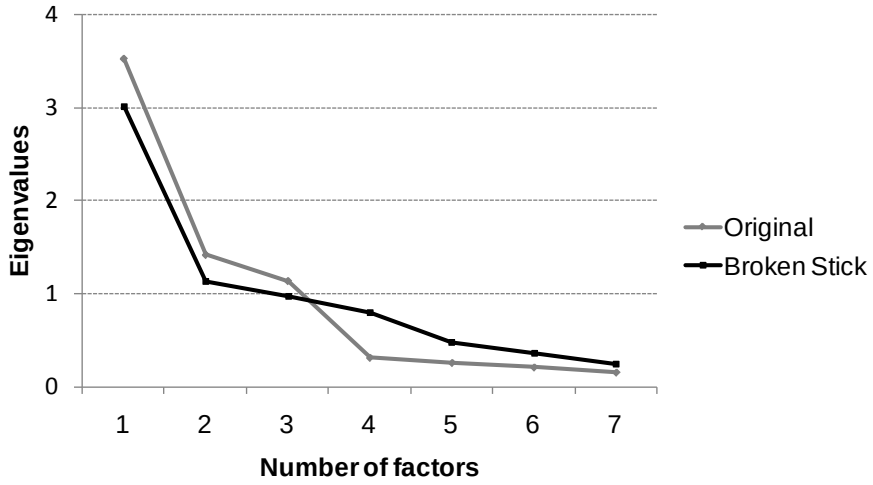

Fig. A8.5 Results of the B-S analysis (II)

Just like in the Parallel Analysis, the B-S results suggest a three-factor solution as for the Eigenvalues from the B-S distribution are lower than those from the original analysis for factors one, two, and three.

Lastly, let's open the *MAP Test.sps* file in the syntax window. Just as before, we need to make some minor adjustments to the syntax. Specifically, in lines 8

and 9, we need to indicate the variables that we use in our factor analysis. In our case, we use the variables x1, x2, x3, x4, x5, x6, and x7. As these appear one after another in the original dataset, we can simply write **x1 to x7**. Now run the syntax by clicking on ► Run ► All.

IBM SPSS Statistics will open the output window which shows the following result (Fig. A8.6).

```
Run MATRIX procedure:

MGET created matrix CR.
The matrix has 7 rows and 7 columns.
The matrix was read from the record(s) of row type CORR.

Eigenvalues
   3.520269746
   1.415206256
   1.135416791
    .311526689
    .256648480
    .207806052
    .153125985

Velicer's Average Squared Correlations
    .000000000    .212233985
   1.000000000    .161130383
   2.000000000    .177411457
   3.000000000    .137104785
   4.000000000    .273701889
   5.000000000    .477745867
   6.000000000   1.000000000

The smallest average squared correlation is
   .1371047853

The number of components is
  3

------ END MATRIX -----
```

Fig. A8.6 MAP Test result

Under **Velicer's Average Squared Correlations**, we can see the average partial correlations after removing each of the factors. As indicated in the output, the minimum correlation of 0.137 is achieved for a three-factor solution.

Taken jointly, the results of the three analyses all provide clear support for a three-factor solution. This is in accordance with the results from the classic approaches as discussed in the book in Chapter 8. Note, however, that this is not always the case, especially when a high number of variables is involved. When divergences occur it is best to rely on PA or MAP test.

References:

Hayton, J.C., Allen, D. G., Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: a tutorial on Parallel Analysis. Organizational Research Methods, 7(2): 191-205.

Henson, R.K., Roberts, J.K. (2006). Use of exploratory factor analysis in published research: common errors and some comment on improved practice. Educational and Psychological Measurement, 66(3): 393-416.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2): 179-185.

Velicer, W.F. (1976) Determining the Number of Components from the Matrix of Partial Correlations. Psychometrika, 41(3): 321-327.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of Five Rules for Determining the Number of Components to Retain. Psychological Bulletin, 99(4): 432-442.

Further reading:

Peres-Neto, R./Jackson, D. a./Somers, K. M. (2005). "How Many Principal Components? Stopping Rules for Determining the Number of Non-trivial Axes Revisited," Computational Statistics & Data Analysis, 49 (4): 974-997.