Non Graphical Solutions for the Cattell's Scree Test

Gilles Raîche, Université du Québec à Montréal
Martin Riopel, Université du Québec à Montréal
Jean-Guy Blais, Université de Montréal

1.   Introduction

Several strategies had been proposed to determine the number of components to retain following a principal components analysis of a correlation matrix. Most of these rely on the analysis of the eigenvalues of the correlation matrix and on numerical solutions. For example, Kaiser's eigenvalue greater than unity rule, parallel analysis, or hypothesis significance tests, like the Bartlett test, allow to make use of numerical criteria of comparison or statistical significance criteria. Apart of these numerical solutions, Cattell proposed the scree test, a graphical strategy to determine the number of components to retain. With the Kaiser's rule, the scree test is probably one the most used strategy and is included in almost all statistical software dealing with principal components analysis. Unfortunately, it is generally recognized that the graphical nature of the Cattell's scree test don't favorise agreement in the number of components to retain. To palliate this problem, some numerical solutions are proposed. A first family of numerical solutions deals with the acceleration of the plot of the eigenvalues, while a second family, more in the spirit of Cattell, deals formally with the scree part of this plot.

Because these solutions are linked to the Kaiser-Guttman rule and the parallel analysis, these are first introduced. Following is the presentation of the Cattell's scree test and its usual interpretations. Finally, the two families of numerical solutions to the scree test are

d4escribed. The presentation of each of these strategy will be accompanied by an example.

## 2.   Classical solutions to determine the number of components to retain

2.1   Kaiser-Guttman rule

According to Guttman (1954) and Kaiser (1960), when a correlation matrix is factorized, it makes no sense to retain components that explain less variance than the original standardized variables. So, principal components eigenvalues equal or less than 1.00 are excluded from the analysis and the number of factors to retain are determined accordingly. The Kayser-Guttman number of components to retain (n_{K-G}) is computed by the following equation:

$$n_{K-G} = Count(\lambda_i > 1)$$ **Equation 1.**

Where λ_i is the ith eigenvalue.

Figure 1 shows a graphical representation of the application of the Kaiser-Guttman rule to 11 eigenvalues obtained from a Pearson correlation matrix. An horizontal line is drawn at 1,00 to illustrate the decision point. With these eigenvalues, four components would be retained.
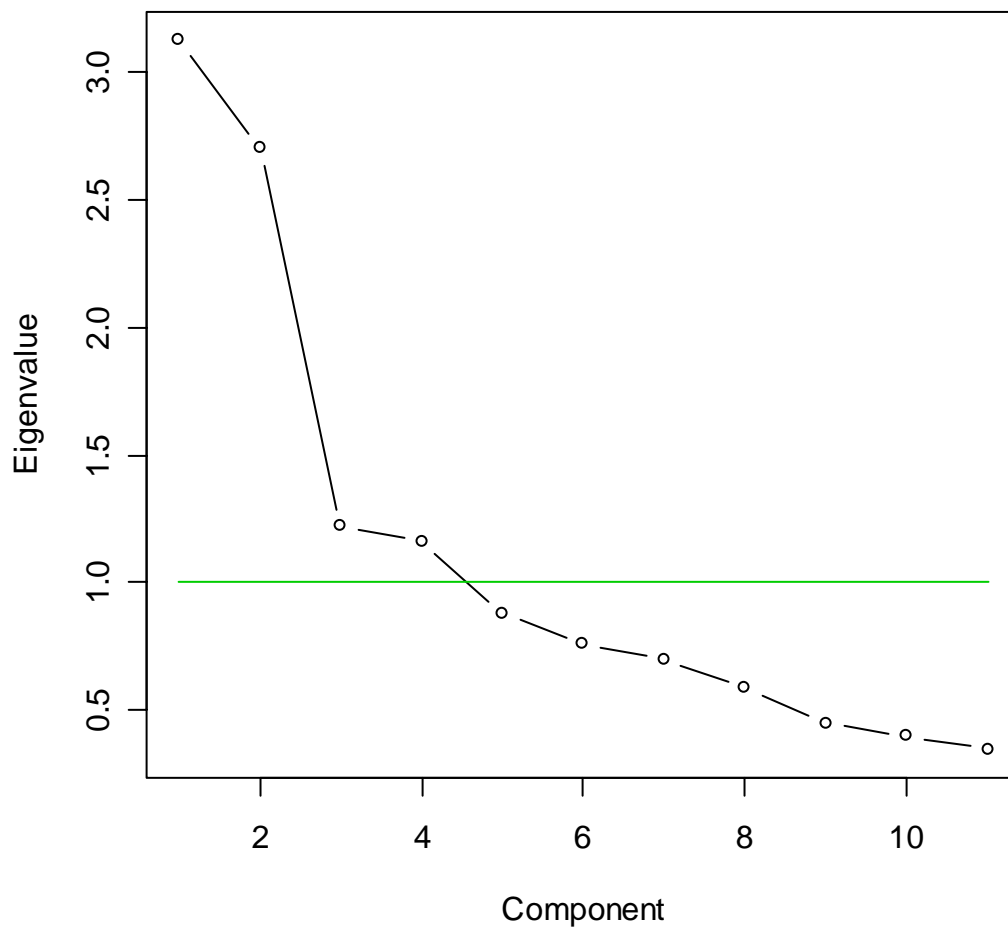
Figure 1. Illustration of the Kaiser-Guttman rule

2.é2   Parallel Analysis

Horn (1965), like Montanelli and Humphrey (1976), indicated that the Keyser-Guttman rule is justified only in the asymptotic case. It can be applied only when the ratio of the number of observations to the number of variables approaches infinity. If practice, this ratio is a finite one and the Keyser-Guttman rule can be greatly misleading. So, they proposed a modification of the Keyser-Guttman rule where the criterion for each eigenvalue is different, and also superior to 1.00.

The proposition is to retain only eigenvalues of the principal components that are superior or equal to the average of the eigenvalues obtained from *k* correlation matrices computed with *n* random observations. If a principal components analysis have to be done on a *p*

variables Pearson correlation matrix computed from n observations, the strategy used for parallel analysis is:

1. Generate n random observations according to a N(0,1) distribution independently for p variates;
2. Compute the Pearson correlation matrix;
3. Compute the eigenvalues of the Pearson correlation matrix;
4. Repeat steps 1 to 3 k times;
5. Compute a location statistic (LS_i) on the p vectors of k eigenvalues : mean, median, 5th centile, 95th centile, etc.
6. Replace the value 1.00 by the location statistic in the Kaiser-Guttman formula.

These steps are summarized by equation 2:

$$n_{Parallel} = Count(\lambda_i > LS_i)$$ **Equation 2.**

A graphical representation of the application of a parallel analysis is shown at figure 2. The 95th centile is used as the location statistic. 100 replications of a Pearson correlation matrix obtained from 30 random N(0,1) observations are generated. With the same eigenvalues that are shown at figure 1, the decision is now to retain only two components. The parallel analysis generally gives more parsimonious solutions than the Kaiser-Guttnan rule.
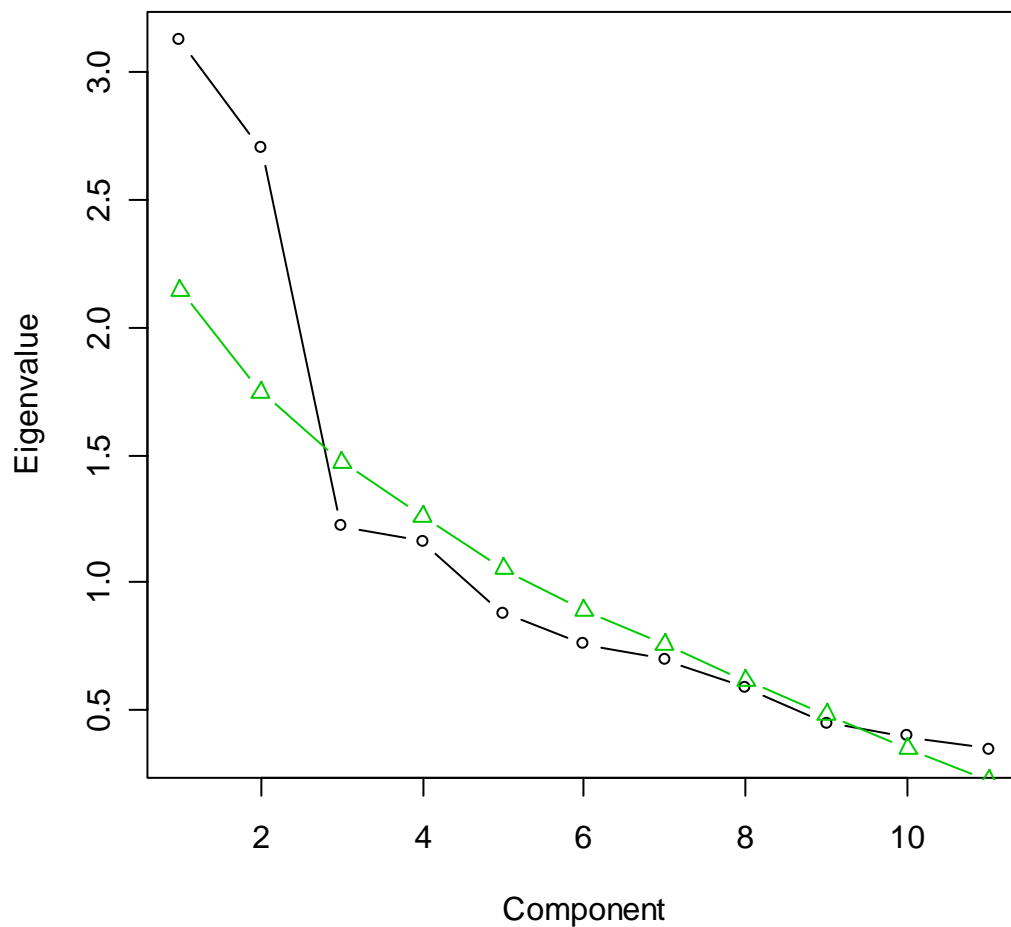
Figure 2. Illustration of a parallel analysis ($LS_i = 95^{th}$ centile, 30 observations, 100 replications)

2.3 Scree Test

The Cattell's scree test (Cattell, 1966), compared to all the other rules for determining the number of components to retain is a visual inspection device. Many challenged its validity (Hoyle and Duvall, 2004; Zwick and Velicer, 1986). For example, Zwick and velicer reported that the inter-rater reliability estimates could be so low than 0.61, with a mean of only 0.81. Despite it subjective basis, the Cattell's scree test is one of the most used strategy to determine the number of components to retain.

Using the scree test consist to graph the eigenvalues with the component number and to look at one of two informations. The first, and the one proposed by Cattell, is to find

where the graph seems to behave randomly, like rocks falling on a scree down a hill. So, according to Cattell, we have to find the line representing this scree. The number of components corresponds to the number of eigenvalues preceding this scree.

Frequently this scree is appearing where the slope of the hill change drastically to generate the scree. It is why many choose the criterion eigenvalue where the slope change quickly to determine the number of components. It is what Cattell named the elbow. So, they look for the place where the positive acceleration of the curve is at his maximum.

Would it be with the location of the scree, or with the location of the position of the maximum positive acceleration, the eigenvalues of the components retained have to be equal or superior to 1.00 in the Kaiser-Guttman rule. It is also possible, and recommended according to us, to replace the Kaiser-Guttman rule by the parallel analysis criterion.

With the same eigenvalues used for figure 1 and 2, a graphical representation of a Cattell's scree test is shown at figure 3. The scree seems to appear at the third eigenvalue, like the quick change of the slope of the curve. According to this test, two components are retained. The same conclusions about the number of components are obtained that the one observed with the parallel analysis.
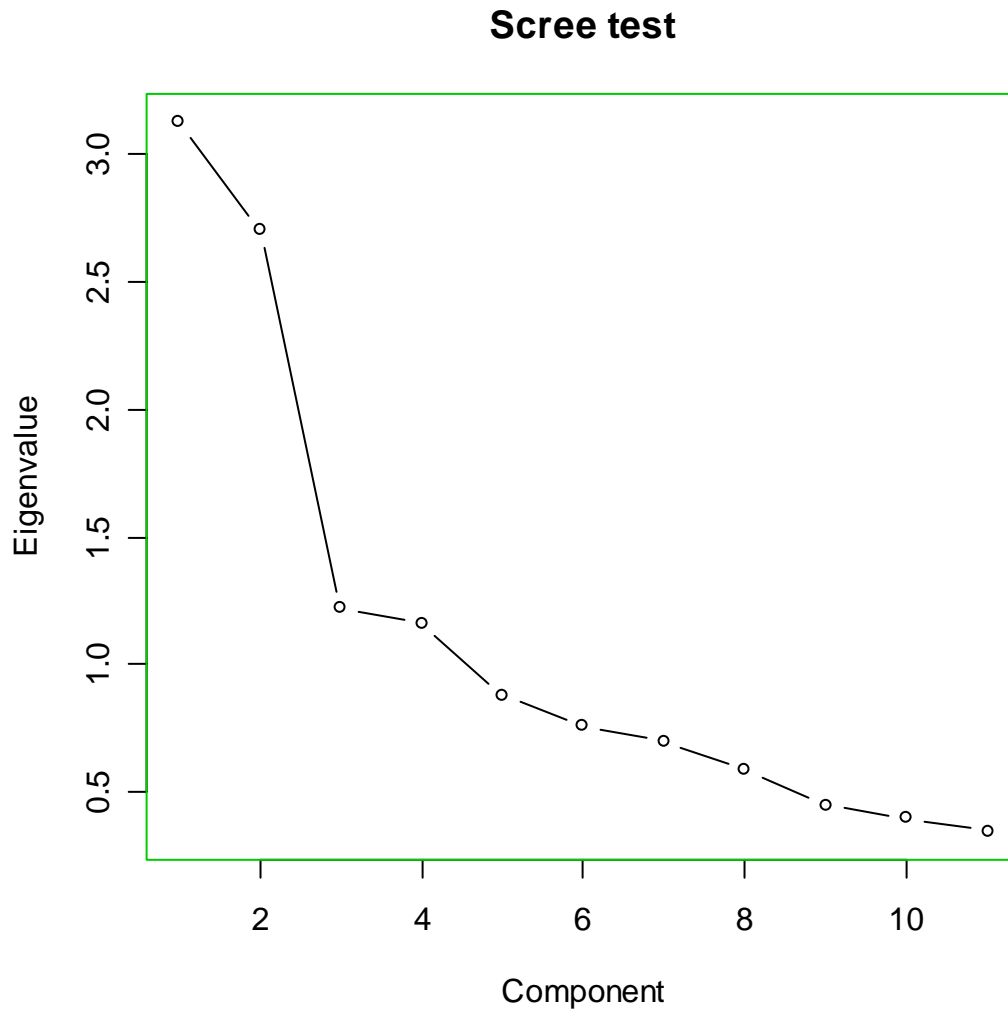
Scree test

Figure 3. Example of a Cattell's Scree Test

3.   Non graphical solutions to the Cattell's scree test

3.1   Scree Test Optimal Coordinate (n_{oc})

To determine the location of the scree, in fact, one has to experiment eigenvalue by eigenvalue with lines going from the coordinates of the last eigenvalue through each of the preceding coordinates. This way one can project each lines to the preceding eigenvalue and verify if this observed eigenvalue is superior or not to the estimated projected eigenvalue. The number of principal components to retain corresponds to the last observed eigenvalue that is superior or equal to the estimated predicted eigenvalue.

Raîche, Riopel, and Blais (2006)

This strategy can be easily automated by computing *p*-2 two-points regression models, and verifying if the observed eigenvalue is, or not, superior or equal to the estimated predicted one by these models. The last of these positive verifications, beginning at the second eigenvalue, and without interruption of the verification, is retained to determine the number of principal components to retain. Of course, the value has to be superior or equal to 1.00, by the Kaiser-Guttman rule (equation 3), or to location statistics criterion, by a parallel analysis (equation 4).

$$n_{oc} = \text{Count} \left[(\lambda_i \geq 1) \text{ and } (\lambda_i \geq predicted(\lambda_i)) \right]$$ **Equation 3.**

$$n_{oc} = \text{Count} \left[(\lambda_i \geq LS_i) \text{ and } (\lambda_i \geq predicted(\lambda_i)) \right]$$ **Equation 4.**

3.2   Scree Test Acceleration Factor (n_{af})

The acceleration factor put emphasis on the coordinate where the slope of the curve change abruptly. At each of the *i* eigenvalue (from 2 to *p*-1), the second derivative ($f''(i)$) is approximated (Yakovitz and Szidarovszky, 1986) by equation 5:

$$f''(i) = \frac{f(i+h) - 2f(i) - f(i-h)}{h}$$ **Equation 5.**

But because *h* correspond to the step in this function and is always equal to 1.00, the function simplify to:

$$f''(i) = f(i+1) - 2f(i) - f(i-1)$$ **Equation 6.**

According to what was explained before for the application of the visual inspection of the Cattell's scree test, the number of principal components to retain is determined by the eigenvalue preceding the coordinate where there is a maximal acceleration factor. At the same time, again, the value of the observed eigenvalue has to be superior or equal to 1.00, by the Kaiser-Guttman rule (equation 7), or to location statistics criterion, by a parallel analysis (equation 8).

Raîche, Riopel, and Blais (2006)

$$n_{af} = \text{If}\left[(\lambda_i \geq 1) \text{ and } (i \equiv \max(af))\right] \qquad \textbf{Equation 7.}$$

$$n_{af} = \text{If}\left[(\lambda_i \geq LS_i) \text{ and } (i \equiv \max(af))\right] \qquad \textbf{Equation 8.}$$

3.3   Comparison of the non graphical solutions

Again with the same 11 eigenvalues, the optimal coordinates and the acceleration factor rules are applied. In each case these solutions are coupled with a 95th centile parallel analysis. Table 1 and figure 4 summarize and picture the results. According to table 1, the optimal coordinate solution shows that the third observed eigenvalue is inferior to the estimated predicted eigenvalue. So the number of components to retain is equal to two (3-1) if the second eigenvalue is equal or superior to the location statistic of the corresponding parallel analysis. It is the case. At figure 4, to picture this solution, a green line go through the pth eigenvalue to the first estimated predicted eigenvalue.

Table 1. Comparison of the determination of the of factors to retain by the optimal coordinates, the acceleration factor, the parallel analysis, and the Keyser-Guttman rule ($n = 30$, $p = 11$, $k = 100$, location statistic = 95th centile)

Component	Eigenvalue	Parallel Analysis	Optimal Coordinate	Acceleration Factor
1	3.12	2.15	2.96	*na*
2	2.70	1.75	1.33	-1.06
3	1.22	1.47	1.28	1.42
4	1.16	1.26	*na*	*na*
5	0.88	1.05	*na*	*na*
6	0.76	0.89	*na*	*na*
7	0.70	0.76	*na*	*na*
8	0.59	0.62	*na*	*na*
9	0.45	0.48	*na*	*na*
10	0.40	0.35	*na*	*na*
11	0.35	0.23	*na*	*na*

Table 1 and figure 4 show that the acceleration factor is also maximum at the third eigenvalue. So, here the same conclusions are drawn with these two non graphical solutions to the scree test.
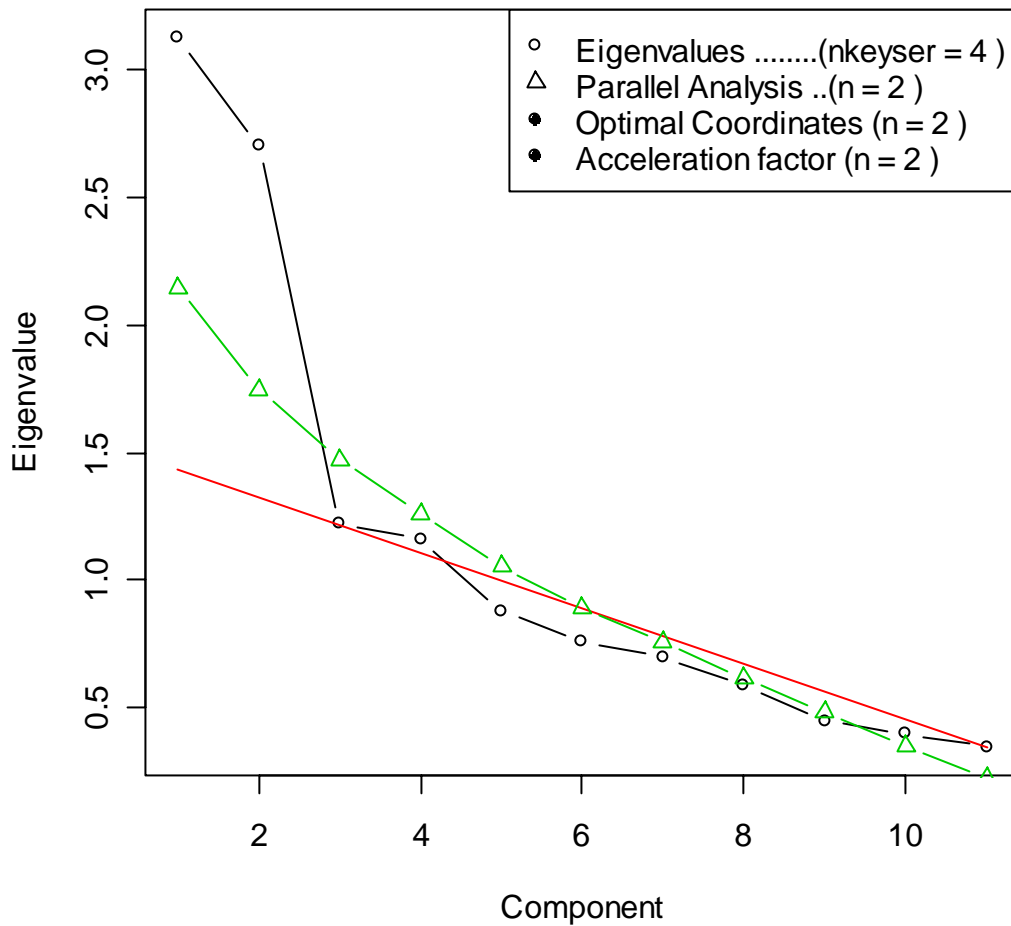
Figure 4. Comparison of the determination of the of factors to retain by the optimal coordinates, the acceleration factor, the parallel analysis, and the Keyser-Guttman rule ($n = 30$, $p = 11$, $k = 100$, location statistic = 95th centile)

Raîche, Riopel, and Blais (2006)

4.   Conclusion

To palliate the subjective weekness of the Cattell's scree test, two families of non graphical solutions to this test were presented. These solutions are easy to apply and, like the parallel analysis solution, seems to give parsimonious solutions compared to the Kaiser-Guttman rule. The next step would be to simulate Pearson correlation matrices from a predetermined factorial structure and to compare the number of components retained by the optimal coordinate and the acceleration factor with other classical strategies: parallel analysis, Kaiser-Guttman rule, Cattell's scree test, minimum average partial correlation (velicer, 1976), Bartlett chi-squared test (Bartlett, 1950), Thompson bootstrap (1988), etc.

5.   References

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology, 3*, 77-85.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1*, 245-276.

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19*, 149-162.

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

Hoyle, R. H. and Duvall, J. L. (2004). Determining the number of factors in exploratory and confirmatory factor analysis. Dans D. Kaplan (Éd.) : *The Sage handbook of quantitative methodology for the social sciences*. Thousand oaks, CA : Sage.

Kaiser, H. F. (1960). The application of electronic computer to factor analysis. *Educational and Psychological Measurement, 20*, 141-151.

Montanelli, R. G. and Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal : A Monte Carlo study. *Psychometrika, 41*, 341-348.

Thompson, B. (1988). Program FACSTRAP: A program that computes bootstrap estimates of factor structure. *Educational and Psychological Measurement, 48*, 681-686.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*, 321-327.

Yakovitz, S. and Szidarovszky, F. (1986). *An introduction to numerical computation*. New York, NJ: Macmillan.

Zwick, W. R. and Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*, 432-442.