STATISTICAL SUPPORT
SWISS CENTER FOR AFFECTIVE SCIENCES

How many components in PCA?

E-mail distributed on 08-07-2024

Dear all,

Principal component analysis (PCA) transforms a set of variables into an equal number of orthogonal components, along directions sorted from largest to smallest variance (Fig. 1, left panel). The transformation is derived from the data's covariance or correlation matrix, and may never involve the explicit calculation of the transformed variables, if the researcher merely wants to interpret how the original variables *load* on the components. As well, although the original formulation of PCA does not require the transformed variables to be reduced, many researchers keep only the first few components (for interpretation or subsequent analyses). For this month's stat support, I would like to discuss a useful rule for how many components to retain, known as **parallel analysis.**

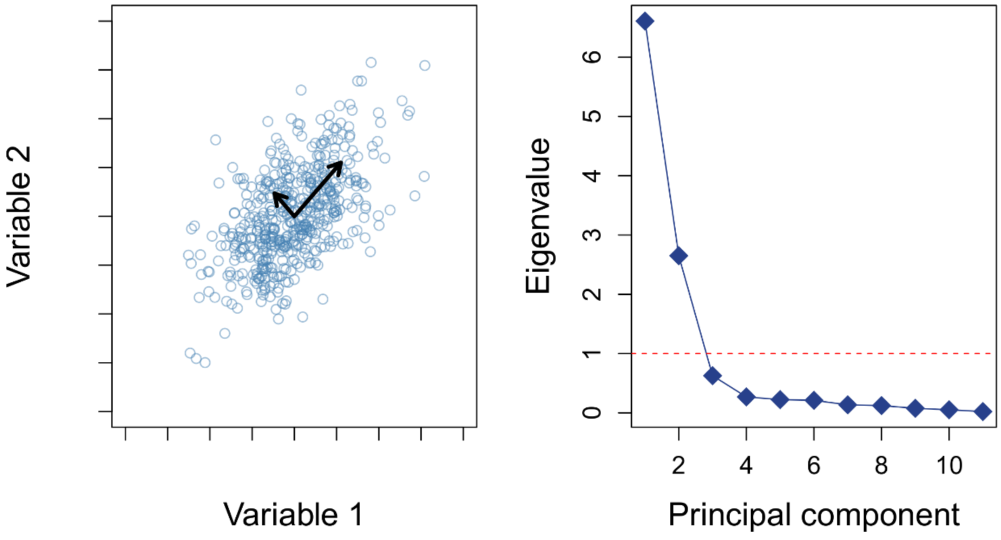

Figure 1. Left panel: Orthogonal directions of variance for 2-dimensional data. Right panel: Scree plot of principal components by their size.

1. Traditional criteria

When the interest is preservation rather than reduction, the number of components to keep is often based on the cumulative variance explained by them, with rules of thumb requiring at least 80% or as much as 95% of the variance explained, with the assumption that any remaining components in the PCA solution reflect merely noise. When the interest is in reduction rather than preservation, other criteria may be used to decide how many components to retain, such as the **scree-criterion**, where components are plotted by their eigenvalue (≈size), and we look for visual evidence of a dent or "elbow" where these eigenvalues suddenly flatten out (Fig. 1, right panel). However, this may fail when there is no clear evidence of an elbow, or more than one.

The scree plot also shows a dashed line at the value 1. Another criterion for component retention is to keep all components with an eigenvalue larger or equal to 1, and is known as **Kaiser's rule.**[1] In theory, for a set of completely uncorrelated variables, the component eigenvalues should all be exactly equal to 1. Unfortunately, this property is only true asymptotically for infinitely large data sets. For finite data sets, especially small ones, components with an eigenvalue larger than 1 are always observed, even for randomly generated data. Table 1 shows a simulation for the first principal component, for increasing sample size and variable size, when averaged over 1000 repetitions. Even for the largest data set (N=1000, Var=5), the first component is larger than 1, with much larger values observed for smaller data sets, especially when the number of variables is large.

Table 1. Average eigenvalue of the first PC for increasing sample size and variable size (1000 repeats). Data consisted of random Gaussian noise.

	Variables			
N	*5*	*10*	*20*	*50*
25	1.60	2.16	3.06	5.26
50	1.41	1.78	2.34	3.66
100	1.28	1.53	1.90	2.71
200	1.20	1.37	1.61	2.13
500	1.13	1.23	1.37	1.67
1000	1.09	1.16	1.26	1.46

When we look at how many components exceed a size of 1 for the same simulation, the average number for these data sets is between 2 and 24, despite the fact that the data are completely random! This phenomenon is often cited as an example of **small-sample bias**, and affects not just principal components but many other statistics that rely on asymptotic theory (e.g., chi-square tests, Akaike's information criterion). However, when the bias is known, it can be exploited to improve Kaiser's rule, which is the principle behind Horn's parallel analysis (Horn, 1965)

2. Parallel analysis

The idea behind parallel analysis is simple. Rather than retaining all PCs with an eigenvalue larger than 1, we only retain PCs whose eigenvalue exceeds the expected eigenvalue for random data of the same dimensions. Fig. 2 shows an example for a random data set with 50 observations and 20

[1] Only applies for PCAs conducted on correlation matrices, not covariance matrices.

variables. For original eigenvalues (red line), the elbow criterion or Kaiser's rule might keep between 3 and 8 components. Horn's parallel analysis, on the other hand (black line; 9999 simulations), selects only one component, with its adjusted eigenvalue extremely close to 1 Note that, although the third component exceeds the adjusted threshold, it is customary to stop retaining components after the first one that fails to exceed its threshold.

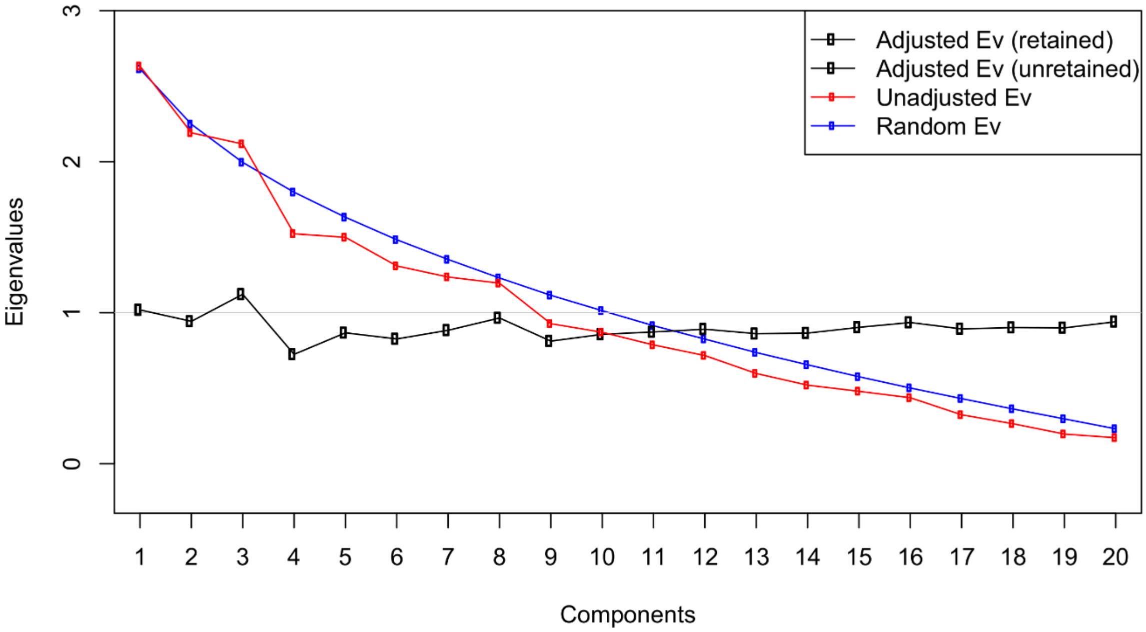

Figure 2. Horn's parallel analysis for a random data set with 50 observations and 20 variables, using 9999 simulations and the 95th percentile as adjustment value.

The parallel analysis illustrated in Fig. 2 has a twist, however. Recall that we should compare our original eigenvalues to "expected" eigenvalues for random data. This expected value need not be the mean across simulations. One might choose the 50th or even 95th percentile of the simulated eigenvalues (Glorfeld, 1995). The latter is depicted in Fig. 2, and yields a far more conservative decision criterion than the mean. In fact, this procedure resembles a permutation test, and provides a somewhat "inferential" way to determine component "significance". One could go one step further and permute the observed data to derive the parallel analysis thresholds, but this appears to have a negligible impact on the outcome, compared to using completely random data.

In R, it is easy to derive the critical thresholds with a simple for-loop, although several packages also supply off-the-shelf functions for parallel analysis. Package `psych` has `fa.parallel`, package `nFactors` has `parallel`, and the package `paran` is entirely dedicated to parallel analysis with the `paran` function. The above graph was generated with `paran`, which can be run directly on your data set, with arguments to control the number of repetitions, the threshold limit, and a switch for factor analysis instead of PCA. Finally, the package `PCDimension` offers a proper permutation test version of parallel analysis with its `rndLambdaF` function.

3. More criteria

While parallel analysis will perform better at component selection than Kaiser's rule, it is advised to always inspect a combination of criteria, as no criterion has been shown to consistently outperform others in the literature, or they may fail in special cases (Jolliffe, 2002). The R package `PCDimension` has the most comprehensive set of criteria to compare from, including some that we did not discuss here, such as Bartlett's test, the Broken-Stick criterion, and Bayesian approaches to component selection. The package vignette provides a quick overview of how to use all of them. I decided to focus on parallel analysis for this newsletter due to its intuitive appeal.

Note finally that component selection should also be guided by interpretability, not just statistical rules and models. If a component does not have a loading pattern with a coherent or plausible interpretation, it may represent a data artefact.

References

Glorfeld, L. W. 1995. An Improvement on Horn's Parallel Analysis Methodology for Selecting the Correct Number of Factors to Retain. *Educational and Psychological Measurement*. 55(3): 377–393.

Horn J. L. 1965. A rationale and a test for the number of factors in factor analysis. *Psychometrika*. 30: 179–185

Jolliffe , I.T. (2002). Principal Component Analysis (2nd edition). New York : Springer-Verlag.

--

Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79