**Title**
Implementing Horn's Parallel Analysis for Principal Components Analysis and Factor Analysis

**Authors**
Alexis Dinno
Center for Tobacco Control Research and Education
530 Parnassus Ave, Suite 366
San Francisco, CA
94143-1390

e-mail: adinno@post.harvard.edu

**Abstract**
I present an implementation of Horn's parallel analysis criteria for factor or component retention in common factor analysis or principal components analysis in Stata, called **paran**. The program permits classical parallel analysis and more recent extensions to it for the **pca** and **factor** commands. **paran** provides a needed extension to Stata's built-in factor and component retention criteria.

**Keywords**
parallel analysis, factor analysis, principal components analysis, factor retention, component retention, dimensional analysis

**Introduction**
A method for factor or component retention is implemented in the Stata program **paran**, based on classical parallel analysis, (Horn, 1965) and recent Monte Carlo extensions to it. (Glorfeld, 1995) A critical aspect of principal components analysis (PCA) or factor analysis (FA) is the researcher decision of how many factors to retain, which can be motivated by a desire to reduce the analytic dimensionality of observed data, as when multiple scores are combined into a single scale, or by a desire to unpack the structure underlying the covariance of observed data, as in exploratory FA. (Vellicer and Jackson, 1990; Preacher and MacCallum, 2003) While there exist a number of criteria for retaining factors or components, a strong consensus has developed in the literature endorsing parallel analysis as among the the most accurate methods. (Humphreys & Montanelli, 1976; Silverstein, 1977; Zwick & Velicer, 1985; Cota, Longman, Holden, & Fekken, 1993; Jackson, 1993; Glorfeld, 1995; Velicer, Eaton, & Fava, 2000; Hayton, Allen, & Scarpello, 2004; Lance, Butts, & Michels, 2006) The choice of retention method is important, as different methods are more or less likely to overestimate or underestimate the number of factors or components.

Horn developed parallel analysis after considering the Kaiser rule, in which one retains eigenvalues are greater than 1 for principal components, or 0 for common factors (Kaiser, 1960). Kaiser formulated this rule following a formal treatment by Guttman, demonstrating that in a population of $P$ variables of infinite size, eigenvalues greater than 1 form a theoretical lower bound on the number of components (or 0 for factors) that can produce a correlation structure among the $P$ variables through linear combination (Guttman 1954). Put another way, a PCA of uncorrelated data would be expected to produce $P$ eigenvalues exactly equal to 1 in uncorrelated data of infinite size. Horn reasoned that in a finite sample of size $N$, one would expect to see eigenvalues greater than and less than one simply due to "sample bias." Indeed, the poor performance of the Kaiser rule has resulted in its panning in the methodological literature (Silverstein, 1977; Zwick and Velicer, 1986; Jackson,

1993; Glorfeld, 1995; Velecer et al., 2000). Horn reasoned that this bias in the Kaiser rule could be corrected by generating a "sufficiently large" number *K* of uncorrelated random data of the same number of *N* and *P* as the observed data, performing a parallel PCA or FA on each, and averaging the results. The bias estimate is thus this average eigenvalue minus one for each component. By subtracting this bias estimate from the eigenvalues from a PCA or FA on the observed data, one retains those adjusted eigenvalues greater than one. (Horn, 1965) More recently, it has been suggested that a more conservative approach would be to generate a large number of random data sets (e.g. 5,000), and use the 95$^{th}$ or 99$^{th}$ percentile, rather than the mean. (Glorfeld, 1995) The **paran** program implements parallel analysis and Glorfeld's extension to it.

**Syntax**

**paran** follows a similar syntax to **pca** and **factor**:

**paran** *varlist* [*weight*] [**if** *exp*] [**in** *range*] [**, iterations(#) centile(#) quietly status factor(***factor_type***) citerate(#) protect(#) all graph copyleft**]

Both *aweights* and *fweights* are allowed.

**Options**

**iteration(#)** Users may set the number of contrast datasets to evaluate. The default value is 30 * the number of variables, and values less than 1 are ignored. For large datasets with large numbers of variables many iterations may be time consuming. The program indicates that every ten iterations of this process have been completed with a dot. The greater the number of iterations the more accurate the estimates of sample bias will be.

**centile(#)** This option specifies that supplied centile value is to be used is used instead of the mean (assumed median, since the distribution is symmetrical) in estimating bias. Values above the mean/median, such as the 95th percentile, are more conservative estimates of chance bias in PCA calculation of eigenvalues from sample data. Values of centile must be greater than 0 and less than 100. Non-integer values will be rounded to the nearest integer value. Running **paran** without this option uses the mean value (very close to **centile(50)**). (see Glorfeld, 1995)

**quietly** Users may set this switch to suppress output for the PCA or factor analysis. This option is only used if a *varlist* is specified in the **paran** command.

**status** This option permits the user to turn off *status*. Default is on.

**factor(***factor_type***)** You must select one of the factor estimation types: **pf**, **pcf**, **ipf**, or **ml** (for principal factors, principal components factors, iterated principal factors, or maximum likelihood factors, respectively). If you specify anything but one of these four abbreviations, you will be warned and the program will halt. CAVEAT: Conducting parallel analysis using factor methods other than **pf** is unorthodox. Interpret such results at your own risk.

**citerate(#)** This option is for the iterated principal factor type.

**protect(#)** This option is for the maximum likelihood factor type.

**all** Users can request that all components or factors be reported, not just those with unadjusted eigenvalues greater than one. The default is not to report all components or factors.

**graph** This option draws a graph of the observed eigenvalues, the random eigenvalues, and the adjusted eigenvalues much like the graphs presented by Horn in his 1965 paper.

**color** This option renders the graph in color with unadjusted eigenvalues drawn in red, adjusted eigenvalues drawn in black, and random eigenvalues drawn in blue, and all lines drawn solid. Without the color option, the graph is rendered in black and white, and the line connecting the unadjusted eigenvalues is dashed, the line connecting the random eigenvalues is dotted, and the line connecting the adjusted eigenvalues is solid.

**copyleft paran** is free software, licensed under the GPL. The **copyleft** option displays the copying permission statement for **paran**. The full license can be obtained by typing:

. net describe **paran**, from (http://www.doyenne.com/stata)

and clicking on the click here to get link for the ancillary file.

## Example

A simulated data set is included with this distribution. It contains 250 observations across 20 variables that have been defined by four random factors plus an amount of noise unique to each measurement. The common variance in these data have been constrained to 0.5 of the total variance. A classical parallel analysis of a PCA performed on these data is obtained by typing:

```
. paran X1-X20, all graph
```

The resulting graph is shown in Figure 1. The default uses 30 times the number of variables iterations, or 600 in this case.

## Technical Notes

Horn (1965) suggested that the simulated data sets be normal with mean of zero, and unit variance. Thompson & Daniel (1996) asserted that data for the simulation be of the same "rank" as the observed data. More recently Hayton, Allen, & Scarpello (2004) urged a parameterization of the random data to approximate the distribution of the observed data with respect to the middle ("mid-point") and the observed min and max. However, PCA and FA standardizing each variable to describe the common variance, so any linear transformation of all variables produce the same eigenvalues. This is born by the notable lack of difference between analyses conducted using the a variety of simulated distributional assumptions (Dinno, 2007). The central limit theorem would seem to make the selection of a distributional form for the random data moot with any sizeable number of iterations.

## References

Cota, A. A., Longman, R. S., Holden, R. R., and Fekken, G. C. (1993). Comparing Different Methods for Implementing Parallel Analysis: A Practical Index of Accuracy. *Educational and Psychological Measurement*, 53(4):865–876.

Dinno, A. 2008. Exploring the Sensitivity of Horn's Parallel Analysis to the Distributional Form of Simulated Data. *Multivariate Behavioral Research* Submitted.

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.

Guttman, L. (1954). Some necessary conditions for common-factor analysis. *Psychometrika*, 19(2):149–161.

Hayton, J. C., Allen, D. G., and Scarpello, V. (2004). Factor Retention Decisions in Exploratory Factor Analysis: a Tutorial on Parallel Analysis. *Organizational Research Methods*, 7(2):191–205.

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

Humphreys, L. G. and Montanelli, R. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A monte carlo study. *Psychometrika*, 41(3):341–348.

Jackson, D. A. (1993). Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology*, 74(8):2204–2214.

Kaiser, H. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, 20(1):141–151.

Lance, C. E., Butts, M. M., and Michels, L. C. (2006). The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say? *Organizational Research Methods*, 9(2):202–220.

Preacher, K. J. and MacCallum, R. C. (2003). Repairing Tom Swift's Electric Factor Analysis Machine. *Understanding Statistics*, 2(1):13–43.

Silverstein, A. B. (1977). Comparison of two criteria for determining the number of factors. *Psychological Reports*, 41(3):387–390.

Thompson, B. and Daniel, L. G. (1996). Factor Analytic Evidence for the Construct Validity of Scores: A Historical Overview and Some Guidelines. *Educational and Psychological Measurement*, 56(2):197.

Velicer, W. F., Eaton, C. A., and Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In Goffen, R. D. and Helms, E., editors, *Problems and Solutions in Human Assessment - Honoring Douglas N. Jackson at Seventy*, pages 41–71. Springer.

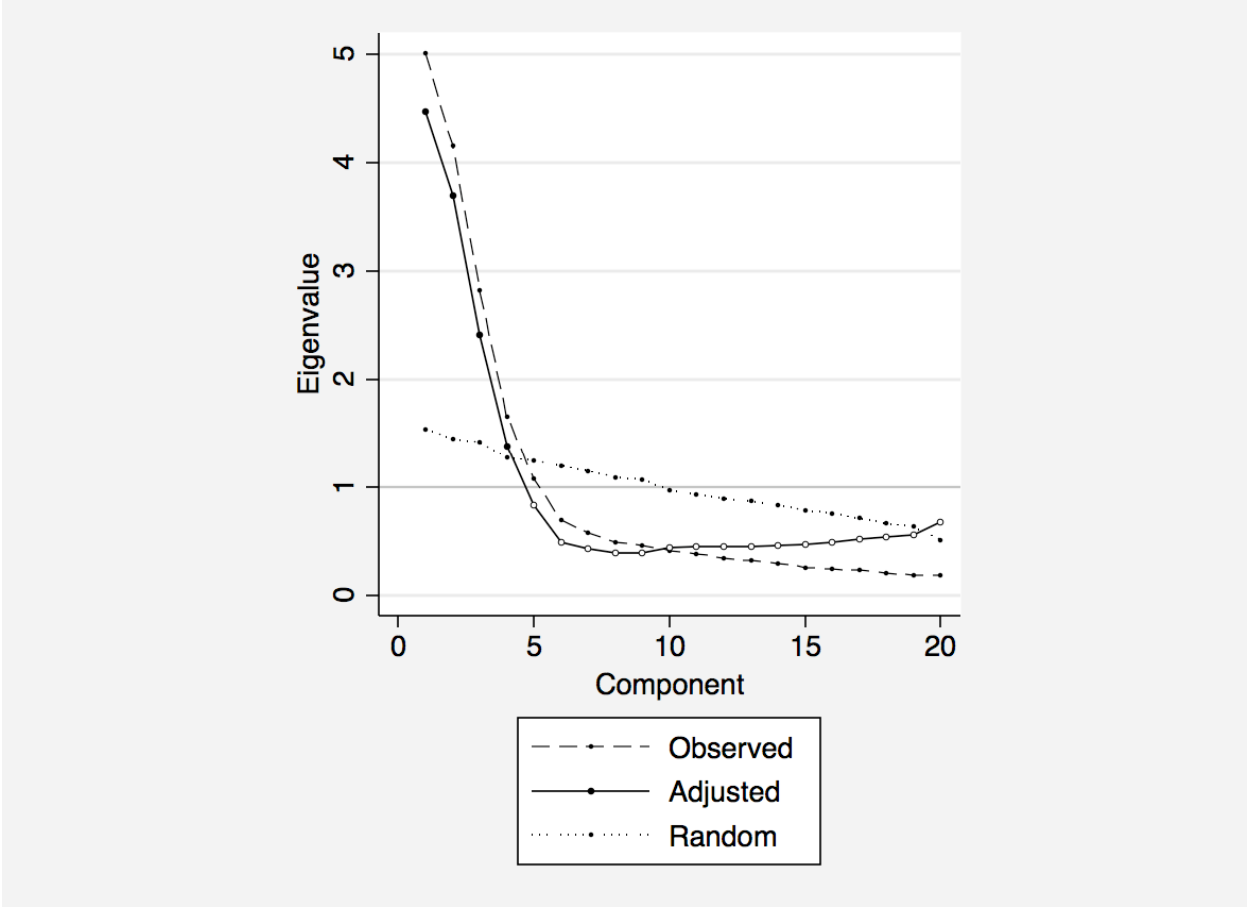
Velicer, W. F. and Jackson, D. N. (1990). Component Analysis versus Common Factor Analysis: Some Issues in Selecting an Appropriate Procedure. *Multivariate Behavioral Research*, 25(1):1–28.

Zwick, W. R. and Velicer, W. F. (1986). A comparison of five rules for determining the number of factors to retain. *Psychological Bulletin*, 99(3):432–442.

**About the Author**

Alexis Dinno is a social epidemiologist and social ecologist with a strong interest in applied quantitative methods. She is presently conducting research at the University of California at San Francisco, where she spends most of her time attending to the question of differential average effect and differential variance of state and county tobacco control policies according to a variety of individual social circumstances using multilevel models. She has an abiding interest in the links between applied research methods and theory.

**Figure 1**




**Caption for Figure 1**
A plot showing the results of the parallel analysis of a PCA on simulated data with four components underlying 20 variables. This example demonstrates the sometimes different results given by parallel analysis versus the eigenvalue greater than one rule for the number of components or factors to retain given. The former obtains the correct number of components, while the later overestimates the number of components.