The Stata Journal (2009)
9, Number 2, pp. 291–298

Implementing Horn's parallel analysis for principal component analysis and factor analysis

Alexis Dinno
Department of Biological Sciences
California State University–East Bay
Hayward, CA
adinno@post.harvard.edu

Abstract. I present `paran`, an implementation of Horn's parallel analysis criteria for factor or component retention in common factor analysis or principal component analysis in Stata. The command 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: st0166, paran, parallel analysis, factor analysis, principal component analysis, factor retention, component retention, Horn's criterion

1 Introduction

A method for factor or component retention is implemented in the Stata command `paran`, based on classical parallel analysis (Horn 1965) and recent Monte Carlo extensions to it (Glorfeld 1995). A critical aspect of principal component analysis (PCA) or factor analysis (FA) is the researcher's decision of how many factors to retain. This decision can be motivated by a desire to reduce the analytic dimensionality of observed data, as when multiple scores are combined into one scale, or by a desire to unpack the structure underlying the covariance of observed data, as in exploratory FA (Velicer and Jackson 1990; Preacher and MacCallum 2003). While several criteria for retaining factors or components exist, a strong consensus has developed in the literature endorsing parallel analysis as among the most accurate methods (Montanelli and Humphreys 1976; Silverstein 1977; Zwick and Velicer 1986; Cota et al. 1993; Jackson 1993; Glorfeld 1995; Velicer, Eaton, and Fava 2000; Hayton, Allen, and Scarpello 2004; Lance, Butts, and Michels 2006). The choice of retention method is important, because 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 greater than 1 for principal components, or greater than 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 greater than 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 1 simply because of "sample bias". Indeed, the poor performance of the Kaiser rule has resulted in its criticism in the methodological literature (Silverstein 1977; Zwick and Velicer 1986; Jackson 1993; Glorfeld 1995; Velicer, Eaton, and Fava 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 1 for each component (minus 0 for common factors). By subtracting this bias estimate from the eigenvalues from a PCA or FA on the observed data, one retains those adjusted eigenvalues greater than 1 (Horn 1965), or greater than 0 for common factors. More recently, it has been suggested that a more conservative approach would be to generate many random datasets (e.g., 5,000), and use the 95th or 99th percentile rather than the mean (Glorfeld 1995).

The `paran` command implements parallel analysis and Glorfeld's extension to it. `paran` is a comprehensive command for parallel analysis, including the adaptation for FA, detailed reporting, graphing features including graphical representation of retained components, and Glorfeld's (1995) Monte Carlo extension to parallel analysis. Stata's user-written `fapara` command also implements a bare-bones parallel analysis computation after `pca` or `factor` has been run, with no options for textual or graphical output, and no implementation of the Monte Carlo variant. However, `fapara` makes its computations using Mata and may be faster.

2 Parallel analysis of data

2.1 Syntax

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

`paran` *varlist* [*if*] [*in*] [*weight*] [, *options*]

options	description
Model	
`iterations(#)`	specify the number of iterations
`centile(#)`	specify using a centile value instead of the mean
`factor(`*factor_type*`)`	use FA instead of PCA; default is `factor(pf)`
`citerate(#)`	communality reestimation iterations
Reporting	
`quietly`	suppress PCA or FA output
`nostatus`	suppress the status indicator
`all`	report all eigenvalues (default reports only those retained)
Graphing	
`graph`	graph unadjusted, adjusted, and random eigenvalues
`color`	render graph in color (default is black and white)
`lcolors(# # # # # # # # # #)`	specify colors using three RGB triples for observed, random, and adjusted eigenvalues (overrides the color option)
`saving(`*filename*`)`	save graph as a `.gph` file
`replace`	replace an existing file when `saving()`
Miscellaneous	
`protect(#)`	perform # optimizations and report the best solution
`seed(#)`	seed the random-number generator with the supplied integer
`copyleft`	display the general public license for `paran`

`fweight`s and `aweight`s are allowed; see [U] **11.1.6 weight**.

2.2 Options

Model options

`iterations(#)` sets 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 can be time consuming. The greater the number of iterations, the more accurate the estimates of sample bias.

`centile(#)` specifies that the supplied centile value is to be used instead of the mean (assumed median, because 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 the eigenvalues from a PCA of sample data. This option supercedes the older `pnf`, which was equivalent to `centile(95)`. Values of `centile()` must be greater than 0 and less than 100. Noninteger 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).

`factor(`*factor_type*`)` selects one of the FA estimation types: `pf`, `pcf`, `ipf`, or `ml` (principal factors, principal component 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. Note that conducting parallel analysis using factor methods other than `pf` is unorthodox. Interpret such results at your own risk. If `factor()` is not used, `paran` performs parallel analysis using PCA by default.

`citerate(#)` sets how many iterations will be used to reestimate communalities for the iterated principal factor type. `citerate()` can be used only with `factor(ipf)`.

Reporting options

`quietly` suppresses output for the PCA or FA. This option is used only if a varlist is specified in the `paran` command.

`nostatus` eliminates `paran`'s default behavior to indicate when every 10th percent of the computation is complete.

`all` reports all components or factors, not just those with unadjusted eigenvalues greater than 1 (or greater than 0 for FA). The default is to not report all components or factors.

Graphing options

`graph` draws a graph of the observed eigenvalues, the random eigenvalues, and the adjusted eigenvalues much like the graphs presented by Horn (1965).

`color` (use only with `graph`) renders the graph in color with unadjusted eigenvalues in red, adjusted eigenvalues in black, random eigenvalues in blue, and all lines solid. Without the `color` option, the graph is rendered in black and white, 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.

`lcolors(# # # # # # # # # #)` (use only with `graph`) specifies the colors of each line on the graph with three RGB triples. The first triple indicates the R, G, and B components of the observed eigenvalues, the second triple sets the values for the mean or centile random eigenvalues, and the third triple sets the values for the adjusted eigenvalues. These settings override the default (red, blue, and black) colors of the `color` option.

`saving(`*filename*`)` (use only with `graph`) outputs the graph to the specified filename as a `.gph` file.

`replace` overwrites an existing filename when the `saving()` option is used with `graph`.

Miscellaneous options

`protect(#)` sets the number of optimizations for the starting values option for the maximum-likelihood factor type. `protect()` can be used only with `factor(ml)`.

`seed(#)` specifies an integer seed for the random-number generator (see `help set seed`) so that the results of `paran` for a specific dataset can be exactly reproduced. The default behavior of `paran` is to not specify a seed.

`copyleft` displays the copying permission statement for `paran`. `paran` is free software, licensed under GNU General Public License. The full license can be obtained by typing

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

and following the on-screen directions.

3 Saved results

The results of `paran` are returned in the matrix `r(HornEv)`, which is a $1 \times P$ matrix of the adjusted eigenvalues.

4 Example

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

```
. use simdata
(Written by R.              )
. paran X1-X20, all graph quietly seed(1)
```

Which, after a moment, produces the following output and figure 1.

```
Computing: 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%

Results of Horn´s Parallel Analysis for principal components
600 iterations, using the mean estimate
-------------------------------------------------
Component   Adjusted    Unadjusted    Estimated
or Factor   Eigenvalue  Eigenvalue    Bias
-------------------------------------------------
 1          4.4965098   5.0115828     .51507294
 2          3.6971236   4.1506432     .45351958
 3          2.4540371   2.8268943     .37285721
 4          1.4124735   1.6603032     .24782968
 5          .86086266   1.0739862     .21312356
 6          .50264634   .69026757     .18762124
 7          .43161616   .57773755     .14612138
 8          .4171857    .48482887     .06764317
 9          .41289346   .46305685     .05016339
 10         .38351215   .41568018     .03216803
 11         .42688028   .38397364     -.04290664
 12         .39338348   .34320168     -.05018181
 13         .45403937   .31778023     -.13625914
 14         .45767627   .29589049     -.16178578
 15         .44264261   .25826351     -.1843791
 16         .49499729   .24047616     -.25452113
 17         .51687283   .23139579     -.28547704
 18         .54348581   .20336817     -.34011763
 19         .55823785   .18720304     -.3710348
 20         .64292367   .18346663     -.45945704
-------------------------------------------------
Criterion: retain adjusted components > 1
```

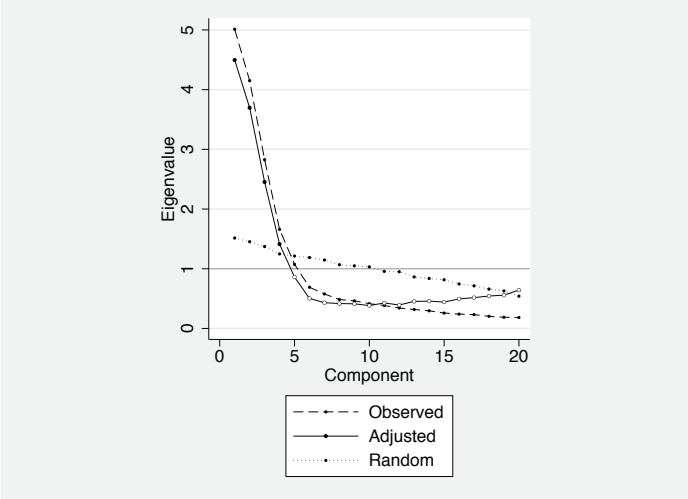

Figure 1. A plot showing the results of the parallel analysis of a PCA on simulated data with four true components underlying 20 variables. Nonretained components are marked with a hollow circle on the adjusted eigenvalues curve.

By specifying the seed for Stata's random-number generator, users can reproduce these results exactly. The default uses 30 times the number of variables iterations, or 600 here. We see in the results that four adjusted eigenvalues are greater than 1. A graph is produced, as shown in figure 1. The dashed line plots the unadjusted eigenvalues in decreasing order, in a "scree plot" as per Cattell (1966). The dotted line is the scree plot for the mean eigenvalues of random N by P data. The adjusted eigenvalues are plotted with a solid line. Horn's criterion corresponds to the point where the adjusted eigenvalues cross the horizontal line at $y = 1$, which is equivalent to the point where the unadjusted eigenvalues cross the curve of mean random eigenvalues. This example demonstrates the sometimes different results given by parallel analysis versus Kaiser's "eigenvalue greater than 1" rule for the number of components or factors to retain. The former obtained the correct number of components, while the latter overestimated the number of components.

5 Technical notes

Horn (1965) suggested that the simulated datasets be normal with means of 0 and unit variance. Thompson and Daniel (1996) asserted that data for the simulation be of the same "rank" as the observed data. More recently, Hayton, Allen, and Scarpello (2004) urged a parameterization of the random data to approximate the distribution of the observed data with respect to the middle (midpoint) and the observed minimum and maximum. However, PCA and FA standardize each variable to describe the total and common variance, respectively, so any linear transformation of all variables produce the same eigenvalues. This is born by the notable lack of difference between analyses conducted using a variety of simulated distributional assumptions (Dinno Forthcoming). 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.

6 References

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

Cota, A. A., R. S. Longman, R. R. Holden, and G. C. Fekken. 1993. Comparing different methods for implementing parallel analysis: A practical index of accuracy. *Educational and Psychological Measurement* 53: 865–876.

Dinno, A. Forthcoming. Exploring the sensitivity of Horn's parallel analysis to the distributional form of simulated data. *Multivariate Behavioral Research*.

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: 377–393.

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

Hayton, J. C., D. G. Allen, and V. Scarpello. 2004. Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods* 7: 191–205.

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

Jackson, D. A. 1993. Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology* 74: 2204–2214.

Kaiser, H. F. 1960. The application of electronic computers to factor analysis. *Educational and Psychological Measurement* 20: 141–151.

Lance, C. E., M. M. Butts, and L. C. Michels. 2006. The sources of four commonly reported cutoff criteria: What did they really say? *Organizational Research Methods* 9: 202–220.

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

Preacher, K. J., and R. C. MacCallum. 2003. Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics* 2: 13–43.

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

Thompson, B., and L. G. Daniel. 1996. Factor analytic evidence for the construct validity of scores: A historical overview and some guidelines. *Educational and Psychological Measurement* 56: 197–208.

Velicer, W. F., C. A. Eaton, and J. L. Fava. 2000. Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In *Problems and Solutions in Human Assessment: Honoring Douglas N. Jackson at Seventy*, ed. R. D. Goffin and E. Helmes, 41–71. Norwell, MA: Kluwer.

Velicer, W. F., and D. N. Jackson. 1990. Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research* 25: 1–28.

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

About the author

Alexis Dinno is a social epidemiologist and social ecologist with a strong interest in applied quantitative methods. She presently lectures part time at California State University–East Bay, and she conducts research in tobacco control policy and the links between urban neighborhood conditions and residents' depressive experiences. She has an abiding interest in the links between applied research methods and theory.