Package 'paran'

July 2, 2014

Version 1.5.1

Date 2012-09-09

Title Horn's Test of Principal Components/Factors

Author Alexis Dinno <adinno@post.harvard.edu>

Maintainer Alexis Dinno <adinno@post.harvard.edu>

Depends R (>= 1.8.0), MASS

Description paran is an implementation of Horn's technique for numerically and graphically evaluating the components or factors retained in a principle components analysis (PCA) or common factor analysis (FA). Horn's method contrasts eigenvalues produced through a PCA or FA on a number of random data sets of uncorrelated variables with the same number of variables and observations as the experimental or observational data set to produce eigenvalues for components or factors that are adjusted for the sample error-induced inflation. Components with adjusted eigenvalues greater than one are retained. paran may also be used to conduct parallel analysis following Glorfeld's (1995) suggestions to reduce the likelihood of over-retention.

URL http://doyenne.com/Software/files/PA_for_PCA_vs_FA.pdf

License GPL-2

LazyLoad true

Encoding latin1

Repository CRAN

Date/Publication 2012-09-10 06:41:56

NeedsCompilation no

R topics documented:

---

paran *Horn's Parallel Analysis of Principal Components/Factors*

Description

paran performs Horn's parallel analysis for a principal component or common factor analysis, so as to adjust for finite sample bias in the retention of components.

Usage

```
paran(x, iterations=0, centile=0, quietly=FALSE,
      status=TRUE, all=FALSE, cfa=FALSE, graph=FALSE,
      color=TRUE, col=c("black","red","blue"),
      lty=c(1,2,3), lwd=1, legend=TRUE, file="",
      width=640, height=640, grdevice="png", seed=0, mat=NA, n=NA)
```

Arguments

x	a numeric matrix or data frame for principal component analysis, or common factor analysis.
iterations	sets the number of iterations with a user specified whole number representing the number of random data sets to be produced in the analysis. The default, indicated by iterations=0, is 30*P*, where *P* is the number of variables or columns in x.
centile	employs Monte Carlo estimates according to the user specified whole number between 1 and 99 indicating the centile used in estimating bias. The default is to use the mean. By selecting a conservative number, such as 95 or 99, and a large number of iterations, paran can be used to perform the modified version of parallel analysis suggested by Glorfeld (1995).
quietly	suppresses tabled output of the analysis, and only returns the vector of estimated biases.
status	indicates progress in the computation. Parallel analysis can take some time to complete given a large data set and/or a large number of iterations. The cfa option may noticably increase the computational requirements of paran.
all	report all eigenvalues (default reports only those components or factors that are retained).
cfa	performs a common factor analysis instead of a principal component analysis. This provides only the unrotated eigenvalues from the common factor model. As of version 1.4.0 paran performs parallel analysis for common factor analysis using a modified method. See Remarks for details.

graph	requests that a plot of the unadjusted, adjusted, and random eigenvalues in a format similar to that presented by Horn in his 1965 paper. Retained components or factors are indicated by the solid circular markers on the adjusted eigenvaue plot, and non-retained components or factors are indicated with hollow circular markers.
color	renders the graph in color with unadjusted eigenvalues drawn in red, adjusted eigenvalues drawn in black, and random eigenvalues drawn in blue if `color=TRUE`, and all lines drawn solid. If `color=FALSE`, 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.
col	sets the colors using a character vector with the color names of adjusted eigenvalues, unadjusted eigenvalues, and estimated random eigenvalues on the plot. These settings are used, only if `color=TRUE`.
lty	sets the line type using an integer vector of the line type codes for adjusted eigenvalues, unadjusted eigenvalues, and estimated random eigenvalues on the plot. These settings are used only if `color=FALSE`.
lwd	sets the line width. The default is `lwd=1`.
legend	draws a legend in the upper right corner of the plot. The default is `legend=TRUE`.
file	the png file in which to save the graph output if the analysis is graphed and `file` is given a character string representing a valid path. The default is not to save the graph.
width	the width in pixels of the png file. The default is `width=640`.
height	the height in pixels of the png file. The default is `height=640`.
grdevice	specifies which graphic `device` to format the graph as, if the user has used the `file` option. The default is `grdevice=png`.
seed	specifies that the random number is to be seeded with the supplied integer. Each random data set is seeded with the supplied value times the number of the iteration, so that entire parallel analyses may be exactly replicated, while each simulated data set maintains a pseudorandom distinction from each of the others. The default value of `seed=0` tells `paran` to use R's default timer based seed (see RNG).
mat	specifies that `paran` use the provided *correlation matrix* rather than supplying a data matrix through `x`. The `n` argument must also be supplied when `mat` is used.
n	the number of observations. Required when the correlation matrix is supplied with the `mat` option, rather than when the data matrix `x` is provided.

Details

`paran` is an implementation of Horn's (1965) technique for evaluating the components or factors retained in a principle component analysis (PCA) or common factor analysis (FA). According to Horn, a common interpretation of non-correlated data is that they are perfectly non-colinear, and one would expect therefore to see eigenvalues equal to 1 in a PCA (or 0 in an FA) of such random data. However, Horn notes that multi-colinearity occurs due to "sampling error and least-squares bias," even in uncorrelated data, and therefore actual PCAs of random data will reveal eigenvalues

of components greater than and less than 1, and FAs will reveal common factors greater than and less than 0. Horn's strategy is to contrast eigenvalues produced through a parallel PCA or FA on a number of random data sets (i.e. uncorrelated variables) with the same number of variables and observations as the experimental or observational dataset to produce eigenvalues for components or factors that are adjusted for the sample error-induced inflation. For PCA, values greater than 1 are retained in the adjustment given by:

$$\lambda_p - (\bar{\lambda}_p^r - 1)$$

and for FA, values greater than 0 are retained in the adjustment given by:

$$\lambda_p - \bar{\lambda}_p^r$$

where λ_p is the p^{th} eigenvalue of the observed data (for $p = 1$ to P), and $\bar{\lambda}_p^r$ is the corresponding mean eigenvalue of the `iterations` number of simulated random data sets.

`paran` performs a PCA or FA with no rotation and performs Horn's adjustment. The user may also specify how many times to make the contrast with a random dataset (default is 30 per variable). Values less than 1 will be ignored, and the default value assumed. Random datasets are generated using the `rnorm()` function. The program returns a vector of length *P* of the estimated bias for each eigenvector, where *P* = the number of variables in the analysis. By specifying a high `centile` users may employ `paran` to conduct parallel analysis following Glorfeld's suggestions to reduce the likelihood of over-retention. (Glorfeld, 1995)

Value

a list of objects relating to the parallel analysis:

`Retained components/factors`
: a scalar integer representing the number of components/factors retained

`Adjusted eigenvalues`
: a vector of the estimated eigenvalues adjusted for a finite sample size

`Unadjusted eigenvalues`
: a vector of the eigenvalues of the observed data from either an unrotated principal component analysis or an unrotated common factor analysis

`Random eigenvalues`
: a vector of the estimated (mean or centile) eigenvalues from `iterations` number of *N* by *P* random data sets

`Bias`
: a vector of the estimated bias of the unadjusted eigenvalues (i.e. the difference between the adjusted and unadjusted eigenvalues)

`Simulated eigenvalues`
: an `iterations` by *P* matrix with each row containing the eigenvalues from an equivalent principal component or common factor analysis on an *N* by *P* data set of uncorrelated random data

Remarks

Hayton, et al. (2004) urge 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 as I understand it is insensitive to standardizing transformations of each variable, and any linear transformation of all variables, and produces the same eigenvalues used in component or factor retention decisions. This is born by the notable lack of difference between analyses conducted using the a variety of simulated distributional assumptions (Dinno, 2009). 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. Former functionality implementing the recommendation by Hayton et al. (2004) has been removed, since parallel analysis is insensitive to it, and it only adds to the computation time required to conduct parallel analysis.

As of `paran` version 1.4.0 application of parallel analysis to common factor analysis has been revised. See the accompanying document Gently Clarifying the Application of Horn's Parallel Analysis to Principal Component Analysis Versus Factor Analysis.

Acknowledgement

A big thank you to Ulrich Keller of the University of Luxembourg for his thoughtful suggestions improving the interface for `paran`, especially the legend, and the `invisibly()` method for returning data, and `seed` option.

Author(s)

Alexis Dinno (alexis dot dinno at pdx dot edu)

References

Dinno A. 2009. Exploring the Sensitivity of Horn's Parallel Analysis to the Distributional Form of Simulated Data. *Multivariate Behavioral Research*. 44(3): 362–388

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

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

Zwick W. R., Velicer WF. 1986. Comparison of Five Rules for Determining the Number of Components to Retain. *Psychological Bulletin*. 99: 432–442

Examples

```
## perform a standard parallel analysis on the US Arrest data
paran(USArrests, iterations=5000)

## a conservative analysis with different result!
paran(USArrests, iterations=5000, centile=95)
```

Index