

# Gently Clarifying the Application of Horn's Parallel Analysis to Principal Component Analysis Versus Factor Analysis

Alexis Dinno

May 24, 2009

## Introduction

Horn's parallel analysis (PA) is an empirical method to decide how many components in a principal component analysis (PCA) or factors in a common factor analysis (CFA) drive the variance observed in a data set of  $n$  observations on  $p$  variables (Horn, 1965). This decision of how many components or factors to retain is critical in applications of PCA or CFA to reducing the dimensionality of data in analysis (as when compositing multiple scale items into a single score), and also in exploratory factor analysis where the different contributions of each factor to each observed variable help generate theory (Preacher and MacCallum, 2003; Velicer and Jackson, 1990). As will be shown, the development of PA was predicated upon properties of PCA. However, some have been exponents of the use of PA for CFA (Velicer, Eaton, and Fava, 2000). The correct application of PA with CFA requires modification to the original PA procedure. This paper attempts to clarify PA with respect to both PCA and CFA.

## Concerning eigenvalues in PCA and CFA

PCA and CFA are two similar methods used to describe the multicollinearity in an  $n$  by  $p$  matrix  $\mathbf{R}$  of observed data. Both methods produce eigenvalues— $\lambda$ s ordered in magnitude from largest ( $\lambda_1$ ) to smallest ( $\lambda_p$ )—which apportion variance along  $p$  unobserved dimensions. One major interpretive difference

between PCA and CFA, is that in the former, each (unrotated) eigenvalue represents a portion of total standardized variance in  $\mathbf{R}$ , and in the later each (unrotated) eigenvalue represents a portion of common standardized variance shared among all  $p$  variables. This means that the eigenvalues of a principal component analysis sum to  $p$ , and that the eigenvalues of a CFA sum to less than  $p$  (and eigenvalues from a CFA can be negative).

For purposes of this paper, PCA is taken to be a function of observed  $n$  by  $p$  data set  $\mathbf{R}$  that returns a set of  $p$  eigenvalues. If  $e(\mathbf{A})$  is a function returning the eigenvalues of square matrix  $\mathbf{A}$ , and  $\text{cov}(\mathbf{R})$  is a function returning the covariance matrix of  $\mathbf{R}$ , then, leaving out the issue of eigenvectors, a PCA of  $\mathbf{R}$  returns the matrix  $\mathbf{\Lambda}$  of eigenvalues as in Equation 1.

$$\mathbf{\Lambda}_{\mathbf{R}} = e(\text{cov}(\mathbf{R})) \quad (1)$$

Where

$$\mathbf{\Lambda}_{\mathbf{R}} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix} \quad (2)$$

If  $\mathbf{U}$  is a matrix of  $n$  observations of  $p$  uncorrelated variables, then as  $n$  approaches  $\infty$ ,  $\mathbf{\Lambda}_{\mathbf{U}}$  approaches the identity matrix  $\mathbf{I}$ . (Equation 3) This gives with the substantive interpretation of PCA as apportioning total standardized variance: if  $p$  variables are perfectly uncorrelated, then in an infinite population they must each explain exactly the same amount of standardized variance, namely  $(1/p)p$ , or 1. One can easily demonstrate this property by running the series of commands in R listed in the Appendix which return the eigenvalues of  $\mathbf{U}$  for progressively larger values of  $n$  (the command returns a vector of the  $\lambda$ s for a PCA of  $\mathbf{U}$ , rather than the matrix  $\mathbf{\Lambda}_{\mathbf{U}}$ ):

$$\lim_{n \rightarrow \infty} \mathbf{\Lambda}_{\mathbf{U}} = \mathbf{I} \quad (3)$$

CFA relevant to PA can be treated with the same formality. If the function  $\text{diag}(\mathbf{A})$  of a square matrix returns a square matrix with the main diagonal elements ( $a_{ij}$  where  $i = j$ ) of  $\mathbf{A}$ , and zeros in all other elements, and if  $\mathbf{A}^+$  is the Moore-Penrose inverse (also ‘generalized inverse’, or ‘pseudoinverse’) of

the matrix  $\mathbf{A}$ , then a CFA of  $\mathbf{R}$  returns the matrix  $\mathbf{\Lambda}_{\mathbf{R}}$  of eigenvalues as in Equation 4.

$$\mathbf{\Lambda}_{\mathbf{R}} = e \left( \text{cov}(\mathbf{R}) - \text{diag}(\text{cov}(\mathbf{R})^+)^+ \right) \quad (4)$$

If  $\mathbf{U}$  is a matrix of  $n$  observations on  $p$  uncorrelated variables, then as  $n$  approaches  $\infty$ ,  $\mathbf{\Lambda}_{\mathbf{U}}$  approaches the zero matrix  $\mathbf{0}$ . (Equation 5) This gives with the substantive interpretation of common factor analysis as apportioning common standardized variance: if  $p$  variables are perfectly uncorrelated, then in an infinite population there can be no common standardized variance, so each factor ‘explains’ zero variance. One can easily demonstrate this property by running the series of commands in R listed in the Appendix (requires the MASS package) which return the eigenvalues of  $\mathbf{U}$  for progressively larger values of  $n$  (the command returns a vector of the  $\lambda$ s for a CFA of  $\mathbf{U}$ , rather than the matrix  $\mathbf{\Lambda}_{\mathbf{U}}$ ):

$$\lim_{n \rightarrow \infty} \mathbf{\Lambda}_{\mathbf{U}} = \mathbf{0} \quad (5)$$

The difference between Equation 3 and Equation 5 is critical to the correct application of PA to PCA versus CFA.

## Applying PA

Kaiser (1960) asserted that in application of PCA one would retain those components with eigenvalues greater than one. (Equation 6)

$$\lambda_q \begin{cases} > 1 & \text{retain} \\ \leq 1 & \text{do not retain} \end{cases} \quad (6)$$

Where  $q$  indexes the eigenvalues from 1 to  $p$ .

Horn (1965) elaborated upon this logic by pointing out that applied researchers do not have an infinite number of observations. According to Horn, in order to account for “sampling error and least squares bias” due to finite  $n$ , one would want to:

1. conduct a parallel PCA on an  $n$  by  $p$  matrix of random values;
2. repeat this  $k$  times;

3. average the eigenvalues  $\lambda_q^r$  over  $k$ , to produce  $\bar{\lambda}_q^r$ ; and
4. adjust  $\lambda_q$  by subtracting from it  $(\bar{\lambda}_q^r - 1)$  to produce  $\lambda_q^{adj}$ .

The retention criterion of PA is to retain those first components with adjusted eigenvalues greater than one (technically, all components following the first component with an adjusted eigenvalue less than one were rejected; the adjustment to subsequent components often increases their eigenvalues, sometimes above the value of one). (Equation 7) This retention criterion can be stated in a mathematically equivalent way as retain those first components with unadjusted eigenvalues greater than the corresponding mean random eigenvalue. (Equation 8)

$$\lambda_q^{adj} \begin{cases} > 1 & \text{retain} \\ \leq 1 & \text{do not retain} \end{cases} \quad (7)$$

$$\lambda_q \begin{cases} > \bar{\lambda}_q^r & \text{retain} \\ \leq \bar{\lambda}_q^r & \text{do not retain} \end{cases} \quad (8)$$

PA must be amended for use with CFA by calculating the adjusted eigenvalue  $\lambda_q^{adj}$  as  $\lambda_q - \bar{\lambda}_q^r$ . The retention criteria must likewise be changed to retain those first adjusted eigenvalues greater than zero. (Equation 9) And as with PA for PCA, this criterion can be restated in an equivalent form as retain those unadjusted eigenvalues greater than the corresponding mean random eigenvalue. (Equation 8)

$$\lambda_q^{adj} \begin{cases} > 0 & \text{retain} \\ \leq 0 & \text{do not retain} \end{cases} \quad (9)$$

## References

- Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30, 179-185.
- Kaiser, H. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, 20, 141-151.
- Preacher, K. J. and MacCallum, R. C. (2003). Repairing Tom Swift's Electric Factor Analysis Machine. *Understanding Statistics*, 2, 13-43.

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, 1-28.

## Appendix

### The limiting case of $\lambda$ s in PCAs of uncorrelated data

```
n <- 100
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U), only.values = TRUE)[[1]]
n <- 10000
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U), only.values = TRUE)[[1]]
n <- 1000000
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U), only.values = TRUE)[[1]]
```

### The limiting case of $\lambda$ s in CFAs of uncorrelated data

```
library(MASS)
n <- 100
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U)-ginv(diag(diag(ginv(cor(U))))), only.values = TRUE)[[1]]
n <- 10000
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U)-ginv(diag(diag(ginv(cor(U))))), only.values = TRUE)[[1]]
n <- 1000000
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U)-ginv(diag(diag(ginv(cor(U))))), only.values = TRUE)[[1]]
```