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

Alexis Dinno
Portland State University

May 20, 2013

Introduction

Horn's parallel analysis (PA) is an empirical method used 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 & MacCallum, 2003; Velicer & 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, & 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 $\mathbf{X}$ of observed data. Both methods produce eigenvalues—λs ordered in magnitude from largest (λ_1) to smallest (λ_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 $\mathbf{X}$, 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 $\mathbf{X}$ that returns a set of p eigenvalues. If $e(\mathbf{A})$ is a function returning the eigenvalues of square matrix $\mathbf{A}$, and $\mathrm{cor}(\mathbf{X})$ is the correlation matrix[1] of $\mathbf{X}$, then, leaving out the issue of eigenvectors, a PCA of $\mathbf{X}$ returns the matrix $\mathbf{\Lambda}$ of eigenvalues as in (1).

$$\mathbf{\Lambda_X} = e(\mathrm{cor}(\mathbf{X})) \tag{1}$$

Where

$$\mathbf{\Lambda_X} = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_p \end{bmatrix}, \tag{2}$$

and $\lambda_1 > \lambda_2 > \cdots > \lambda_p$.

If $\mathbf{U}$ is a matrix of n observations of p uncorrelated variables, then as n approaches ∞, $\mathbf{\Lambda_U}$ approaches the p by p identity matrix $\mathbf{I}$ (3). This jibes 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 $(1/p) \times p$, or 1.

$$\lim_{n\to\infty} \mathbf{\Lambda_U} = \underset{p\times p}{\mathbf{I}} \tag{3}$$

One can easily demonstrate this limiting property by running the series of commands in R listed in Appendix A which return the eigenvalues of a PCA of $\mathbf{U}$ for progressively larger values of n (the commands return the diagonal of $\mathbf{\Lambda_U}$).

The behavior of CFA relevant to PA in the limit of n can be approached in the same fashion. If the function $\mathrm{diag}(\mathbf{A})$ 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

Alexis Dinno, School of Community Health, Portland State University.
Correspondence concerning this paper should be addressed to Alexis Dinno, School of Community Health, Portland State University, PO Box 751–SCH, Portland, Oregon 97207-0751 United States of America. Email alexis.dinno@pdx.edu
This paper should be cited as an unpublished manuscript using the URL: http://doyenne.com/Software/files/PA_for_PCA_vs_FA.pdf

[1] A previous version of this document mistakenly used the term 'covariance matrix.' While PCA and CFA can be performed using covariance matrices with specific constraining assumptions (Gorsuch, 1983), the arguments presented here were and are relevant to *correlation* matrices.

elements, and if $\mathbf{A}^{+}$ is the Moore-Penrose inverse (also 'generalized inverse', or 'pseudoinverse') of the matrix $\mathbf{A}$, then a CFA of $\mathbf{X}$ returns the matrix $\mathbf{\Lambda_X}$ of eigenvalues as in (4).

$$\mathbf{\Lambda_X} = e\left(\mathrm{cor}\,(\mathbf{X}) - \mathrm{diag}\,(\mathrm{cor}\,(\mathbf{X})^{+})^{+}\right) \tag{4}$$

If $\mathbf{U}$ is a matrix of n observations on p uncorrelated variables, then as n approaches ∞, $\mathbf{\Lambda_U}$ approaches the p by p zero matrix $\mathbf{0}$ (5). This jibes 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 common variance.

$$\lim_{n\to\infty} \mathbf{\Lambda_U} = \underset{p\times p}{\mathbf{0}} \tag{5}$$

One can easily demonstrate this property by running the series of commands in R listed in Appendix B (requires the MASS package from http://cran.r-project.org) which return the eigenvalues of $\mathbf{U}$ for progressively larger values of n (the commands return the diagonal of $\mathbf{\Lambda_U}$).

The difference between (3) and (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 (6).

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

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 uncorrelated random values;
2. repeat this k times;
3. average each of the eigenvalues λ_q^r over k, to produce $\bar{\lambda}_q^r$; and
4. adjust λ_q by subtracting from it $\left(\bar{\lambda}_q^r - 1\right)$ to produce λ_q^{adj}.

The retention criterion of PA is to retain those first components with adjusted eigenvalues greater than one (7). Technically, PA is a stopping rule in PCA, because the adjustment to subsequent components—especially the last few components—may sometimes increase their eigenvalues above the value of one. The retention criterion in (7) can be stated in a mathematically equivalent way as "retain those first components with unadjusted eigenvalues greater than the corresponding mean random eigenvalue" (8).

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

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

PA must be amended for use with CFA by calculating the adjusted eigenvalue λ_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 (9). Technically, PA is a stopping rule in CFA, because the adjustment to subsequent common factors—especially the last few factors—may sometimes increase their eigenvalues above the value of one. 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" (8).

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

References

Gorsuch, R. L. (1983). *Factor Analysis 2nd ed.* NJ: Lawrence Erlbaum Associates.

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

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

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

Velicer, W. F., Eaton, C. A., & 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 R. D. Goffen & E. Helms (Eds.), *Problems and Solutions in Human Assessment - Honoring Douglas N. Jackson at Seventy* (pp. 41–71). Springer.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some further observations. *Multivariate Behavioral Research*, *25*(1), 97–114.

Appendix A

*

The limiting case of $\mathbf{\Lambda}$ in PCAs of uncorrelated data

```
n <- 100
p <- 20
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U), only.values = TRUE)[[1]]
n <- 10000
p <- 20
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U), only.values = TRUE)[[1]]
n <- 1000000
p <- 20
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U), only.values = TRUE)[[1]]
```

Appendix B

*

The limiting case of $\mathbf{\Lambda}$ in CFAs of uncorrelated data

```
library(MASS)
n <- 100
p <- 20
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U)-ginv(diag(diag(ginv(cor(U))))), only.values = TRUE)[[1]]
n <- 10000
p <- 20
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U)-ginv(diag(diag(ginv(cor(U))))), only.values = TRUE)[[1]]
n <- 1000000
p <- 20
U <- matrix(rnorm(n*p),n,p)
eigen(cor(U)-ginv(diag(diag(ginv(cor(U))))), only.values = TRUE)[[1]]
```