DOCUMENT RESUME

ED 074 106 TM 002 469

AUTHCR Halperin, Silas
TITLE The Incorrect Measurement of Components.
PUB DATE Feb 73
NOTE 10p.; Paper presented at the annual meeting of the American Educational Research Association, New Orleans, Louisiana, February, 1973

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Correlation; *Factor Analysis; *Orthogonal Rotation; Scoring Formulas; Speeches; Statistical Analysis; Technical Reports; Test Interpretation; *Weighted Scores

ABSTRACT

Factor loadings, used directly or as the basis of binary values, are not appropriate as weights to produce component scores from a rotated solution. A series of examples showing the results of an incorrect measurement of components is given. Several correlation matrices were taken from books on factor analysis and multivariate analysis. Each matrix was submitted to a principal components decomposition, and those vectors whose roots were greater than 1.0 were then rotated according to the normalized varimax criterion. Following this, several matrices were calculated. A clear pattern resulted. Components supposedly orthogonal to each other often come out highly correlated. In addition, factor loadings that are small in the rotated principal components structure are often moderate to large in R/zh and R/zk. Thus, the researcher who interprets and names components finds that his scores have a completely different meaning than he anticipates. Any use of these scores in further analysis will result in a serious distortion of conclusions. (Author/KM)

THE INCORRECT MEASUREMENT OF COMPONENTS

By

Silas Halperin

Syracuse University

Paper presented at the annual meeting of the

American Educational Research Association

New Orleans

February 28, 1973

Introduction

Factor analysis has proved to be a popular mode of data analysis in the behavioral sciences. However, what is often referred to as factor analysis is actually based on the component model rather than the factor model and should be called component analysis. Certainly, of all the methods currently available, the most widely used is the principal components analysis followed by the varimax rotation. It remains popular not because it is based on a method or model which is universally appropriate, but because it is simple, because computer programs are readily available and because component scores are easily calculated. Yet, in spite of the fact that component scores are easily found, there is a history of them being found incorrectly.

Good discussions of the measurement of components from the rotated principal components solution have been written by such authors as Harman (1967), Kaiser (1962), and Glass and Maguire (1966). Still, many users of the model have their own ideas of how these scores should be obtained. Their methods have great intuitive appeal and appear reasonable and obvious. One of the most popular of these methods uses factor loadings as weights of the standardized variables and the resulting linear combinations are called "factor scores." Another method, claimed to yield approximate factor scores, sums those standardized variables with "high" loadings on a particular component. Thus, weights of zero and one are used in the linear composite.

Both methods are technically incorrect. However, a detailed discussion of the mathematical reasons why the methods are incorrect may not be persuasive to the user who finds them intuitively appealing. Yet these users need to be convinced. For this reason, the theoretical development of this study is supplemented by a series of examples showing the results of an incorrect measurement of components.

Theoretical development

If n variables are observed for each of N subjects, the component model, stated in matrix form, is given by

(1) $$Z = A_T\, F_T \ ,$$

where Z is the nxN matrix of standardized scores, A_T is the complete nxn pattern matrix and F_T is a nxN matrix of component scores. If A_T represents a complete principal components solution, then

(2) $$F_T = A_T^{\ -1}\, Z \ .$$

However, the user of the component model usually wishes to retain only the largest components, say the first m. Thus we work with only the first m columns of A_T and the first m rows of F_T, which we might denote as A and F respectively. Since A is an nxm matrix and has no inverse, we cannot use equation (2) to solve for component scores. Still, a simple solution for F exists:

(3) $$F = (A'A)^{-1}\, A'Z = D^{-2}\, A'Z \ ,$$

where $A'A = D^2$ is a diagonal matrix of the m largest eigenvalues of the correlation matrix. It is readily proved that the covariance matrix, S_F, of the components equals the identity matrix and the matrix of variable by component correlations, R_{ZF}, equals the pattern matrix A. Also, since the inverted eigenvalues scale the rows of A', the factor loadings become the effective weights in the linear combinations which produce the component scores. Thus, the first method mentioned in the introduction is correct but for a scaling constant and the second method, using weights of zero and one, becomes an approximation which Schweiker (1967) has argued possesses some desirable properties.

However, if an orthogonal rotation is performed which results in a new matrix of loadings, say B, we now have a new set of component scores, G, which can be shown to satisfy the following equation:

$$G = (B'B)^{-1} B'Z \ . \tag{4}$$

As before, the covariance matrix of these component scores, denoted S_G, equals the identity matrix and the matrix of variable by component correlations, R_{ZG}, equals the rotated matrix of loadings B.

A comparison of equations (3) and (4) indicates a great deal of similarity and one important difference. It will be observed that, while A'A is a diagonal matrix, the product B'B is not. Hence, the weights required to calculate component scores are no longer simply scalings of the loadings in the rows of B'. If we persist in the use of the rows of B' as weighting vectors, we will find scores which do not reflect the properties of the solution.

We can now formalize method one relative to an orthogonally rotated principal components solution. Using loadings as weights, we find a set of composite scores, denoted H, given by

$$H = B' Z. \tag{5}$$

The matrix H will not equal the correct scores G, nor will it possess properties of the solution given by B.

Two matrices were utilized to determine the extent to which the incorrect component scores do not reflect the solution. The first matrix, R_{ZH}, contains the variable by component correlations. If H were correct, R_{ZH} and B would be equal. The second is the correlation matrix, R_{HH} of the incorrect component scores. The components should be uncorrelated, so R_{HH} should equal the identity matrix.

Starting with equation (5) we can easily derive S_H, the matrix of component covariances:

$$S_H = \frac{1}{N} \; H \, H'$$

$$= \frac{1}{N} \; B'ZZ'B$$

But

$$\frac{1}{N} ZZ' = R$$

so

(6) $$S_H = B' \; R \; B.$$

It can further be derived that this expression will not reduce down to a diagonal matrix.

To find R_{ZH}, we need the variances of H, which reside in the diagonal of S_H. Denoting a diagonal matrix of these variances as D_H, we find

$$R_{ZH} = \frac{1}{N} \; Z \; H'D_H^{-1/2}$$

$$= \frac{1}{N} \; ZZ'B \; D_H^{-1/2}$$

(7) $$R_{ZH} = R \; BD_H^{-1/2} \quad ,$$

which will not in general equal the matrix B of loadings.

Finally,

(8) $$R_{HH} = D_H^{-1/2} \; S_H \; D_H^{-1/2} \quad .$$

Again, R_{HH} should be an identity matrix, but can be shown not to be.

Before studying the three matrices developed above, it is useful to consider incorrect component scores of the second variety. If, rather than using elements of B as weights, we simply sum the standarized scores of those variables with "large" loadings, we are effectively using a binary matrix of weights. It was arbitrarily decided to assign a weight of ±1.0 to all loadings whose absolute value was greater than .5 and a zero to all loadings of .5 or less in absolute value. Denote by C this binary analogue to B. Then component scores

(9) $$K = C' Z$$

have a covariance matrix

(10) $$S_K = C' RC,$$

and a correlation matrix

(11) $$R_{KK} = D_K^{-1/2} S_K D_K^{-1/2}$$

where D_K is a diagonal matrix of component score variances. Also, the matrix of variable by component correlations is given by

(12) $$R_{ZK} = R C D_K^{-1/2} .$$

With the mathematics developed, it remains to study these matrices for each procedure. The matrices were calculated for each of a variety of examples and compared to the form which they would take if they were an adequate reflection of their respective solutions.

Illustrations and Results

Several correlation matrices were taken from books on factor analysis and multivariate analysis. Each matrix was submitted to a principal components decomposition and those vectors whose roots were greater than 1.0 were then rotated according to the normalized varimax criterion. Following this, matrices R_{ZH}, R_{ZK}, R_{HH} and R_{KK}, among others, were calculated. A clear pattern resulted and five examples best illustrating this pattern were chosen and are reported in Tables 1a-1e below. Table 1f contains the only example in which the pattern was not obvious.

The pattern exhibited is clear. Components supposedly orthogonal to each other often come out highly correlated (e.g. $r_{H_2H_3}$ = .86 in Table 1d).

TABLE 1

Illustrative Examples of Variable-by-Component and Component-by-Component Correlation Matrices, Using Correct Components (G), Components Using Loadings as Weights (H) and Components Using Binary Weights (K)

1a: Eight Physical Variables from Harman (1967, p.80)

$$R_{ZH} = \begin{bmatrix} .93 & .63 \\ .93 & .58 \\ .91 & .55 \\ .92 & .60 \\ .57 & .91 \\ .48 & .84 \\ .42 & .80 \\ .52 & .78 \end{bmatrix} \qquad R_{ZG} = B = \begin{bmatrix} .90 & .26 \\ .93 & .20 \\ .92 & .16 \\ .90 & .23 \\ .25 & .89 \\ .18 & .84 \\ .11 & .84 \\ .25 & .75 \end{bmatrix} \qquad R_{ZK} = \begin{bmatrix} .94 & .46 \\ .95 & .41 \\ .93 & .38 \\ .93 & .43 \\ .44 & .91 \\ .37 & .86 \\ .30 & .84 \\ .40 & .80 \end{bmatrix}$$

$$R_{HH} = \begin{bmatrix} 1.00 & 0.74 \\ 0.74 & 1.00 \end{bmatrix} \qquad R_{GG} = I = \begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix} \qquad R_{KK} = \begin{bmatrix} 1.00 & 0.44 \\ 0.44 & 1.00 \end{bmatrix}$$

1b: Eight Variables on 100 Rectangles from Cooley and Lohnes (1971, p.134)

$$R_{ZH} = \begin{bmatrix} .95 & .55 \\ .45 & .90 \\ .96 & .56 \\ .46 & .89 \\ .99 & .82 \\ .95 & .93 \\ .82 & .99 \\ .99 & .70 \end{bmatrix} \qquad R_{ZG} = B = \begin{bmatrix} .99 & .06 \\ .09 & .99 \\ .99 & .56 \\ .10 & .89 \\ .91 & .82 \\ .77 & .93 \\ .55 & .99 \\ .97 & .70 \end{bmatrix} \qquad R_{ZK} = \begin{bmatrix} .94 & .46 \\ .48 & .94 \\ .94 & .47 \\ .48 & .92 \\ .99 & .75 \\ .96 & .89 \\ .84 & .98 \\ .99 & .62 \end{bmatrix}$$

$$R_{HH} = \begin{bmatrix} 1.00 & 0.79 \\ 0.79 & 1.00 \end{bmatrix} \qquad R_{GG} = I = \begin{bmatrix} 1.00 & 0.00 \\ 0.00 & 1.00 \end{bmatrix} \qquad R_{KK} = \begin{bmatrix} 1.00 & 0.74 \\ 0.74 & 1.00 \end{bmatrix}$$

1c: A 9 Variable Example from Horst (1965, p.122)

$$R_{ZH} = \begin{bmatrix} .91 & .21 & .53 \\ .92 & .23 & .56 \\ .90 & .39 & .53 \\ .27 & .85 & .40 \\ .18 & .87 & .26 \\ .25 & .87 & .31 \\ .45 & .33 & .84 \\ .45 & .25 & .84 \\ .53 & .49 & .80 \end{bmatrix} \qquad R_{ZG} = B = \begin{bmatrix} .92 & .00 & .18 \\ .91 & .02 & .22 \\ .89 & .21 & .15 \\ .07 & .83 & .19 \\ .03 & .89 & .04 \\ .09 & .88 & .07 \\ .15 & .11 & .85 \\ .16 & .01 & .88 \\ .26 & .29 & .71 \end{bmatrix} \qquad R_{ZK} = \begin{bmatrix} .93 & .09 & .38 \\ .94 & .11 & .41 \\ .91 & .27 & .38 \\ .18 & .86 & .31 \\ .11 & .89 & .19 \\ .17 & .89 & .23 \\ .33 & .22 & .86 \\ .33 & .14 & .86 \\ .40 & .35 & .81 \end{bmatrix}$$

$$R_{HH} = \begin{bmatrix} 1.00 & 0.40 & 0.71 \\ 0.40 & 1.00 & 0.51 \\ 0.71 & 0.51 & 1.00 \end{bmatrix} \qquad R_{GG} = I = \begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 0.00 \end{bmatrix} \qquad R_{KK} = \begin{bmatrix} 1.00 & 0.17 & 0.42 \\ 0.17 & 1.00 & 0.28 \\ 0.42 & 0.28 & 1.00 \end{bmatrix}$$

1d: A Contrived 10 Variable Example from Mulaik (1972, p.228)

$$\begin{array}{ccc}
R_{ZH} & R_{ZG} = B & R_{ZK} \\
\begin{bmatrix} .59 & .89 & .70 \\ .43 & .76 & .53 \\ .70 & .85 & .58 \\ .73 & .65 & .86 \\ .39 & .58 & .73 \\ .52 & .50 & .77 \\ .86 & .50 & .53 \\ .90 & .69 & .61 \\ .86 & .61 & .75 \\ .80 & .83 & .81 \end{bmatrix} &
\begin{bmatrix} .16 & .84 & .34 \\ .05 & .81 & .19 \\ .44 & .79 & .06 \\ .44 & .19 & .74 \\ -.03 & .35 & .75 \\ .20 & .11 & .81 \\ .94 & .08 & .10 \\ .87 & .34 & .10 \\ .76 & .12 & .45 \\ .51 & .54 & .45 \end{bmatrix} &
\begin{bmatrix} .50 & .90 & .56 \\ .37 & .79 & .40 \\ .63 & .86 & .39 \\ .66 & .55 & .86 \\ .34 & .49 & .80 \\ .45 & .43 & .83 \\ .89 & .41 & .36 \\ .91 & .61 & .42 \\ .88 & .51 & .61 \\ .80 & .82 & .64 \end{bmatrix} \\
R_{HH} & R_{GG} = I & R_{KK} \\
\begin{bmatrix} 1.00 & 0.83 & 0.84 \\ 0.83 & 1.00 & 0.86 \\ 0.84 & 0.86 & 1.00 \end{bmatrix} &
\begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{bmatrix} &
\begin{bmatrix} 1.00 & 0.68 & 0.58 \\ 0.68 & 1.00 & 0.59 \\ 0.58 & 0.59 & 1.00 \end{bmatrix}
\end{array}$$

1e: A Contrived 10 Variable Example from Fruchter (1954,p.36)

$$\begin{array}{ccc}
R_{ZH} & R_{ZG} = B & R_{ZK} \\
\begin{bmatrix} .48 & .91 & .78 \\ .48 & .80 & .89 \\ .33 & .34 & .88 \\ .91 & .79 & .47 \\ .81 & .90 & .46 \\ .36 & .88 & .31 \\ .90 & .46 & .81 \\ .79 & .47 & .92 \\ .87 & .32 & .36 \\ .96 & .84 & .85 \end{bmatrix} &
\begin{bmatrix} .03 & .81 & .58 \\ .02 & .62 & .78 \\ -.01 & .03 & .99 \\ .83 & .56 & -.01 \\ .63 & .77 & -.02 \\ .05 & .99 & -.03 \\ .79 & -.02 & .61 \\ .59 & -.01 & .81 \\ .99 & -.05 & .01 \\ .79 & .48 & .49 \end{bmatrix} &
\begin{bmatrix} .51 & .91 & .79 \\ .51 & .79 & .91 \\ .36 & .33 & .89 \\ .90 & .79 & .46 \\ .81 & .91 & .46 \\ .36 & .89 & .33 \\ .90 & .46 & .79 \\ .81 & .46 & .91 \\ .86 & .33 & .33 \\ .97 & .84 & .84 \end{bmatrix} \\
R_{HH} & R_{GG} = I & R_{KK} \\
\begin{bmatrix} 1.00 & 0.71 & 0.71 \\ 0.71 & 1.00 & 0.69 \\ 0.71 & -.69 & 1.00 \end{bmatrix} &
\begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{bmatrix} &
\begin{bmatrix} 1.00 & 0.72 & 0.72 \\ 0.72 & 1.00 & C.69 \\ 0.72 & 0.69 & 1.00 \end{bmatrix}
\end{array}$$

1f: Ten Variables Related to Educational and Occupational Aspirations of 17 Year-Old Boys from VandeGeer (1971, p.165)

$$\begin{array}{ccc}
R_{ZH} & R_{ZG} = B & R_{ZK} \\
\begin{bmatrix} .10 & .45 & -.47 \\ .46 & .65 & -.10 \\ .49 & .54 & .22 \\ .58 & .42 & .44 \\ .73 & .50 & -.15 \\ .34 & .15 & -.77 \\ .56 & .80 & .08 \\ .53 & .82 & .04 \\ .83 & .56 & -.13 \\ .81 & .56 & -.06 \end{bmatrix} &
\begin{bmatrix} -.16 & .58 & -.46 \\ .23 & .61 & -.08 \\ .35 & .43 & .24 \\ .55 & .19 & .47 \\ .72 & .19 & -.13 \\ .37 & -.03 & -.76 \\ .27 & .76 & .10 \\ .21 & .81 & .06 \\ .81 & .22 & -.10 \\ .78 & .23 & -.03 \end{bmatrix} &
\begin{bmatrix} .10 & .58 & -.11 \\ .35 & .70 & -.10 \\ .35 & .35 & -.09 \\ .67 & .26 & .04 \\ .75 & .35 & -.20 \\ .21 & .13 & -1.00 \\ .44 & .78 & -.08 \\ .40 & .80 & -.07 \\ .84 & .38 & -.28 \\ .81 & .38 & -.20 \end{bmatrix} \\
R_{HH} & R_{GG} = I & R_{KK} \\
\begin{bmatrix} 1.00 & 0.78 & -.05 \\ 0.78 & 1.00 & -.04 \\ -.05 & -.04 & 1.00 \end{bmatrix} &
\begin{bmatrix} 1.00 & 0.00 & 0.00 \\ 0.00 & 1.00 & 0.00 \\ 0.00 & 0.00 & 1.00 \end{bmatrix} &
\begin{bmatrix} 1.00 & 0.45 & -.21 \\ 0.45 & 1.00 & -.13 \\ -.21 & -.13 & 1.00 \end{bmatrix}
\end{array}$$

In addition, factor loadings which are small in matrix B, the rotated principal components structure, are often moderate to large in R_{ZH} and R_{ZK} (e.g., $b_{92} = .12$ while $r_{Z_9H_2} = .61$ and $r_{Z_9K_2} = .51$ in Table 1d). Thus, the researcher who interprets and names components finds that his scores, H or K, have a completely different meaning than he anticipates. Any use of these scores in further analysis will result in a serious distortion of conclusions. Even in Table 1e, where the distortion was minimal, we find $r_{H_1H_2}$ to be .78 and one factor loading of .23 resulting in r_{ZH} and r_{ZK} of .56 and .38 respectively.

Conclusions

The mathematics and illustrations lead us to the conclusion that factor loadings, used directly or as the basis of binary values are not appropriate as weights to produce component scores from a rotated solution. But intuitively they may still make sense. In an attempt to dispatch this idea, let us draw an anology. Linear composites are also used as the basis of prediction in multiple regression. However, no person at all familiar with regression analysis would use predictor-criterion correlations as regression weights. Why then should the use of correlations as weights be considered intuitively appealing in component analysis?

REFERENCES

Cooley, W. W. and Lohnes, P. R. Multivariate data analysis. New York: Wiley, 1971.

Fruchter, B. Introduction to factor analysis. Princeton: D. Van Nostrand Co., 1954.

Glass, G. V. and Maguire, T. O. Abuses of Factor Scores. American Educational Research Journal, 1966, 3, 297-304.

Harman, H. H. Modern factor analysis. Chicago: University of Chicago Press, 1967.

Horst, P. Factor analysis of data matrices. New York: Holt, Rinehart and Winston, 1965.

Kaiser, H. F. Formulas for component scores. Psychometrika, 1962, 27, 83-87.

Mulaik, S. A. The foundations of factor analysis. New York: McGraw-Hill Book Co., 1972.

Schweiker, R. F. Factor scores aren't sacred: Comments on "Abuses of factor scores." American Educational Research Journal, 1967, 4, 168-170.

Van de Geer, J. P. Introduction to multivariate analysis for the social sciences. San Francisco: W. H. Freeman and Co., 1971.