Method of Making an Inconsistent Matrix Consistent*

Marion F. Shaycoft

American Institutes for Research

A. Problem.

When correlation matrices are used for multiple regression, canonical correlation, factor analysis, or any other form of multivariate analysis, it is of course highly desirable to have all the correlation coefficients based on exactly the same cases. Unfortunately this isn't always feasible. Particularly when the number of variables is large, data are likely to be missing on some variables for some of the individuals; and if the matrix were based only on cases that had complete data for all the variables, the N might be so small that the sampling errors would be unduly large. But when an n-variable correlation matrix is not based entirely on the same group of cases it may be internally inconsistent; this means that there is no set of N x n real numbers which, functioning as scores of N individuals on n variables, would produce that matrix.

Missing data are not the sole source of inconsistent matrices. For instance when one has two or more consistent matrices, each based on the same n variables but on

*This paper was presented at the Annual Convention of the American Psychological Association, 3 September 1967, in Washington, D.C.

different groups of cases, one may wish to combine these matrices, via Fisher's z, to obtain a single matrix representing the correlations for the "average group." This average matrix can be inconsistent.

Inconsistent matrices can also be produced by correcting some or all variables for attenuation--in other words by dividing by the square roots of the reliability coefficients of the variables. This procedure will produce an inconsistent matrix only if some of the values used as reliability coefficients produce an overcorrection because they are under-estimates of the true reliability of the variable <u>for the group on which the correlation matrix is based</u>.

B. <u>Determining whether a matrix is inconsistent</u>.

Some matrices proclaim their own inconsistency loudly, by containing one or more correlation coefficients outside the +1 to -1 range. This can happen, for instance, with the correction-for-attenuation procedure.

In other cases the zero-order correlations may look quite all right, so that it is not immediately apparent to the naked eye that the matrix is inconsistent, but the existence of inconsistency becomes unmistakable later on, when a partial correlation or multiple or canonical correlation turns out to be outside the +1 to -1 range. (As a matter of fact it may even turn out to be an imaginary number, though this couldn't happen unless some of the zero-order r's were greater than +1 or less than -1.)

Table 1 summarizes some situations in which these obvious symptoms of matrix inconsistency can occur.

But even if none of the individual zero-order correlation coefficients and none of the higher-order correlations one happens to have computed from them turn out to be outside the +1 to -1 range, it is a simple matter to determine whether the matrix is consistent. All that is necessary is to compute the eigenvalues. If none of the eigenvalues is negative, the matrix is consistent--in other words Gramian. If one or more of the eigenvalues is negative, the matrix is inconsistent (non-Gramian).

C. Principle underlying procedure for adjusting an inconsistent matrix to make it consistent.

Because the use of inconsistent correlation matrices in multivariate analysis leads to so many difficulties and is so undesirable on theoretical grounds, and because situations in which such matrices can arise are so prevalent, the author has developed

TABLE 1. Inconsistent correlation matrices: some sources and obvious manifestions of inconsistency

	Situation	Procedure for getting matrix	Can this procedure produce an inconsistent matrix?	Which of the obvious manifestations of inconsistency indicated below in col. 4 can occur when the indicated procedure is used?			
				Manifestation of inconsistency	Can it occur in: Zero-order r?	Partial r?	Multiple R?
1.	Missing data	Base each r on available cases	Yes	$r > 1$ $r < -1$ r imaginary	No No No	Yes Yes No	Yes No No
2.	Fallible measures	Correction for attenuation	Yes	$r > 1$ $r < -1$ r imaginary	Yes Yes No	Yes Yes Yes	Yes No Yes
3.	A single matrix, representing the average of several groups, is required	Average via Fisher z	Yes	$r > 1$ $r < -1$ r imaginary	No No No	Yes Yes No	Yes No No
	(1)	(2)	(3)	(4)	(5)	(6)	(7)

a procedure for adjusting an inconsistent matrix, by means of very small modifications in some of the correlations, to make it into a consistent matrix.

Basically the procedure is just the reverse of factor analysis. In factor analysis one breaks a set of correlated variables down into a set of factor loadings which could explain the correlations. In the reverse procedure one takes a set of factor loadings and from it builds up the correlation matrix compatible with it. In other words instead of analyzing a correlation matrix in terms of loadings on constituent factors, one synthesizes a correlation matrix from the factor loadings.

Thus the procedure for making an inconsistent matrix consistent might reasonably be called "correlation synthesis."

To recapitulate: In factor analysis one normally starts with a mathematically reasonable correlation matrix and finds a factor pattern compatible with it. In what we are calling correlation synthesis, on the other hand, one starts with a reasonable factor pattern and develops a correlation matrix compatible with it. The key word here is the word "reasonable." In the present context, a correlation matrix is considered reasonable if it contains no values inconsistent with the definition of the correlation coefficient (i.e., no values outside the ± 1 to -1 range) and no internal inconsistencies; in other words if it could be produced by real data; a factor pattern is considered reasonable if it does not involve such illogicalities as imaginary loadings and imaginary factor scores.

The entire correlation synthesis procedure is outlined in the next section.

D. Procedure for checking and correcting correlation matrix for inconsistency.

1. Test for consistency

   a. Put 1's in the diagonal of n-variable matrix to be tested. Call this matrix M.

   b. Compute all eigenvalues.
      m = number of eigenvalues ≥ 0
      m' = number of eigenvalues > 0

   c. If $m = n$, matrix M is consistent.
      If $m < n$, matrix M is inconsistent and it is then necessary to apply the subsequent steps.

2. Test for applicability of correction procedure

   a. Do complete principal components analysis.

      A = the complete n x n matrix of factor loadings, for n principal components. (The last n - m components are imaginary.)

      $M - AA' = 0$

      B = the n x m matrix of factor loadings for first m principal components only.

b. Compute $M - BB'$, and inspect all terms in it.

1) If these residuals are small (preferable small enough to be regarded as due to sampling error) the correction procedure (i.e., Step 3) is clearly applicable.

2) If, on the other hand, there are some <u>large</u> residuals, the researcher should probably give serious consideration to the question of whether he really wants to analyze that set of data, or whether the departures from inconsistency are so serious that Matrix M is not worth analyzing, even after adjustment.

3. Correction procedure ("Correlation synthesis")

a. Let D = n x n diagonal matrix of terms from the diagonal of BB'.

b. Compute $R = D^{-1/2} BB'D^{-1/2}$

R is the corrected correlation matrix. It is a consistent matrix, with 1's in the diagonal. All off-diagonal terms are 1 or less, and the matrix is of rank m', with no negative eigenvalues.

4. Check on magnitude of adjustments.

As a final step, compute R - M and inspect all terms in it. These terms, which usually turn out to be very small, represent the changes in the correlation coefficients to make the matrix consistent. If, however, some of the differences are large, the researcher should again seriously consider whether his analysis of that particular matrix is defensible. It may be, or it may not, depending on the reasons for the discrepancies. The researcher, through familiarity with his own data, may have some insight into what those reasons are.

E. Example of application of procedure

The procedure described has been used operationally just once, thus far. In analyzing some data obtained by retesting over 7,000 Project TALENT students three years after their initial testing in 1960 as ninth-graders,* a mammoth missing-data problem was faced. Because of limitations on testing time each student could be given only half of the original battery. To insure that the correlation between any pair of test variables could be determined on some group, six overlapping batteries were set up, each requiring only one day of testing time and thus only half as long as the original battery, and each consisting essentially of the tests in two of the four original half-days of testing. This meant that some pairs of tests (mostly pairs where both variables were from the original testing) were taken by virtually the entire group. For other pairs, only about half the cases were available; in this category were pairs for which both tests were in the same half-day of retesting and also pairs where one variable was from the original testing and the other from the retest. For still

*A complete report of this study is present in: Shaycoft, Marion F. The High School Years: Growth in Cognitive Skills. (Interim report 3 to the U.S. Office of Education, Cooperative Research Project No. 3051.) Pittsburgh: Project TALENT Office, American Institutes for Research and University of Pittsburgh, 1967.

other pairs of tests only about one-sixth of the cases were available. These were tests in two different half-days of retesting; the only students having scores for both these retest variables were those taking the particular retest battery combining these two half-days.

Table 2 summarizes the results of applying the "correlation synthesis" correction procedure to two 99-variable matrices, one based on boys in this TALENT retest, and one on girls. Because of the magnitude of the missing data problem, a special procedure for correcting each initial correlation matrix for missing data* was applied to the initial correlation matrix to produce an intermediate matrix. The correlation synthesis procedure to correct for inconsistency was then applied to the intermediate matrices to produce the final matrices. It can be seen from Table 2 that of the 99 eigenvalues for each intermediate matrix, two small negative ones occurred in the boys' matrix and one in the girls'. All three were of course eliminated in the final matrix. The largest difference between a correlation in the intermediate matrix (Matrix M) for boys and the corresponding final matrix (Matrix R) was only .015. This occurred

*This procedure, the details of which are irrelevant to the present paper, is discussed briefly elsewhere. (See Shaycoft, op.cit., Chapter 3.) A report now in preparation will fully document it.

TABLE 2. Basic data concerning the development of consistent correlation Matrices 1A and 1B

Matrix	Sex	No. of cases on which individual r's are based**	No. of variables	Matrices with 1's in diagonal	No. of eigenvalues $\lambda<0$	No. of eigenvalues $\lambda=0$	No. of eigenvalues $\lambda>0$	Values of negative eigenvalues
1A	M	483-3441	99	Initial matrix (pseudo-matrix)*	1	0	98	-.31
				Intermediate matrix	2	0	97	-.01,-.29
				Final (consistent) matrix*	0	2	97	--
1B	F	496-3676	99	Initial matrix (pseudo-matrix)*	1	0	98	-.10
				Intermediate matrix	1	0	98	-.13
				Final (consistent) matrix*	0	1	98	--

*These matrices are presented in full in:
Shaycoft, Marion F. The High School Years: Growth in Cognitive Skills, (Interim report 3 to the U.S. Office of Education, Cooperative Research Project No. 3051.) Pittsburgh: Project TALENT Office, American Institutes for Research and University of Pittsburgh, 1967.

Tables I-1 and I-2 contain initial matrices 1A and 1B respectively.
Tables 6-1a and 6-1b contain final matrices 1A and 1B respectively.

**N's vary widely because of extensive missing data.

for a correlation in the vicinity of .36. The fact that .015 was the largest change and that most of the other changes were far smaller helps to demonstrate the effectiveness of the correlation synthesis procedure for making an inconsistent matrix consistent.