Chance Baselines for INDSCAL's Goodness-of-Fit Index

Hei-Ki Dong
The Ball Foundation

For purposes of evaluating INDSCAL's goodness-of-fit index (R_1), random data were generated for various combinations of N (the number of individuals), ranging from 3 to 30, and n (the number of stimuli), ranging from 10 to 40. Results from INDSCAL analyses of random data were then used to provide an easy-to-use equation for estimating the expected value of R_1, as a function of N, n, and r (the number of dimensions). The equation was highly successful in predicting expected values of R_1 for random data, as indicated by a multiple correlation of .9965.

A very popular multidimensional scaling (MDS) method for studying individual differences in perceptions of various stimulus domains is the INDSCAL method developed by Carroll and Chang (1970). A basic problem with INDSCAL, as with other scaling methods, is the problem of evaluating how well the model and method fit real data, that is, goodness-of-fit.

In the context of two-way nonmetric MDS, which involves the analysis of a symmetric matrix of proximities among n stimuli, several monte carlo studies (Klahr, 1969; Levine, 1978; Spence & Ogilvie, 1973; Stenson & Knoll, 1969; Wagenaar & Padmos, 1971) have been conducted to provide practitioners with chance or random baselines of Kruskal's (1964a, 1964b) fit index called stress.

In the context of three-way nonmetric MDS, which involves the analysis of an N symmetric matrix of proximities, MacCallum (1981) provided stress values for random data using ALSCAL (Takane, Young, & de Leeuw, 1977). Since ALSCAL optimizes a different fit statistic than INDSCAL, and INDSCAL is a metric, not a nonmetric, three-way MDS method, results from the MacCallum study cannot be used to evaluate the goodness-of-fit of INDSCAL solutions.

INDSCAL's goodness-of-fit index, R_1, is simply the product-moment correlation between elements of the lower or upper half (including the diagonal) of N observed and N solution scalar product matrices. Mathematically, it is defined as:

$$R_I = \frac{\sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{j} (s_{ijk} - s...)(\hat{s}_{ijk} - \hat{s}...)}{\left[\sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{j} (s_{ijk} - s...)^2 \sum_{i=1}^{N} \sum_{j=1}^{n} \sum_{k=1}^{j} (\hat{s}_{ijk} - \hat{s}...)^2 \right]^{1/2}}, \qquad (1)$$

where s_{ijk} is the scalar product for subject i and stimulus pair j and k,

$\hat{s}_{ijk}$ is the estimate of the scalar product s_{ijk}, and

$s...$ and $\hat{s}...$ are the grand means of the observed and estimated scalar products.

The primary purpose of this study was to provide users of INDSCAL with a quantitative basis for evaluating INDSCAL's fit statistic, R_I, by developing a simple and easy-to-use equation for determining expected values of R_I for random data.

Method

In order to determine expected values of R_I for random data a monte carlo study was carried out. This involved several steps. First, for each combination of N and n shown in Table 1, several lower-half matrices (replications) of order $N \times n \times n$ were constructed by randomly sampling elements from a uniform distribution in the interval 0 to 1. The uniform distribution was chosen for the sake of comparability with other studies (e.g., MacCallum, 1981); only a few data sets were generated for the largest values of N and n because obtained R_I values were very stable across replications. There were 141 random data sets representing 19 combinations of N and n, essentially covering sample sizes from 3 to 30 and numbers of stimuli from 10 to 40. Second, each data set generated was input to an INDSCAL program (Pruzansky, 1975) and scaled in two through five dimensions, with the maximum number of iterations set to 40 and data specified as a lower-half dissimilarity matrix without diagonals.

Next, for each solution, the value of R_I was recorded and means of R_I were computed for each combination of N, n, and r, the number of dimensions; there were 76 means of R_I representing various

Table 1
Number of Random Data Sets for Each Combination
of $\underline{N}$ and $\underline{n}$

Number of Individuals ($\underline{N}$)	Number of Stimuli ($\underline{n}$)	Number of Random Data Sets
3	10	10
3	20	10
5	10	10
5	20	10
10	10	10
10	20	10
10	30	5
15	10	10
15	20	10
15	30	5
20	10	10
20	20	10
20	30	5
20	40	5
25	20	5
25	30	5
25	40	3
30	20	5
30	30	3

combinations of N, n, and r. Then, nonlinear transformations of N, n, and r were entered into regression analyses in order to determine the best equation for predicting values of R_1 (MacCallum, 1981). Predictors, N, n, and r, were selected based on degrees of freedom rationales (Carroll & Chang, 1970); three parameters were involved in determining the number of degrees of freedom in the solution for an r-dimensional solution, which was $r(N + n - 2)$, and degrees of freedom for data, which was simply $Nn(n - 1)/2$.

Results and Discussion

The following regression equation, constructed using a logarithmic transformation, has proven to be very accurate in estimating the expected value of R_1 for random data:

$$E[\ln(R_1)] \cong .7961 - (.1918)\ln(N) - (.5896)\ln(n) + (.4285)\ln(r) \quad , \tag{2}$$

where E is the expected value and ln is the logarithm to the base e. The multiple correlation for this equation was .9965. Also, it should be mentioned that variances around the estimated expected values were fairly small, that is, standard deviations of R_1 within each of 76 conditions[1] investigated were all consistently around .01.

In practice, a user should compare the observed R_1 values from real data to the estimated expected values computed using the above equation. If an observed R_1 value is not substantially higher than the estimated expected value for random data, the data should be rejected as being random. As an illustration, take an example with $N = 20$, $n = 10$, and $r = 3$. The expected value of R_1 for random data with these parameters would be computed as follows:

$$E[\ln(R_1)] \cong .7961 - (.1918)2.9957 - (.5896)2.3026 + (.42851)1.0986 = -.6653 \quad , \tag{3}$$

so that $E(R_1) \cong .5141$. Thus, if an R_1 value obtained from real data was not substantially higher than .5141, the data should be rejected as having a random structure. Although ranges of N and n investigated in this study were typical of those encountered in empirical research, results reported here are, strictly speaking, limited to those specified in Table 1.

In short, the present study provides users of INDSCAL with a simple and easy-to-use equation for accurately estimating expected values of R_1 for random data, which can then be used to draw inferences regarding the fit of data by INDSCAL. However, results of this study cannot be used directly to determine the *dimensionality* of a given set of data. Although it is tempting to compare the series of observed and expected R_1 values and look for the point at which the two sets of values cross to determine the true dimensionality of the data, analogous to the parallel analysis technique developed by Montanelli and Humphreys (1976) for factor analysis, this type of procedure could not work in MDS because R_1 is measuring *overall fit* in a given number of dimensions (MacCallum, 1981).

Results of this study may be used indirectly, however, to help decide the dimensionality by comparing observed increments in R_1 against expected increments in R_1. For example, taking parameters from the previous illustration, $N = 20$, $n = 10$, and $r = 3$, the expected R_1 is .5141. For $r = 4$, the expected R_1 is .5816; an expected increment of $.5816 - .5141 = .0675$. Therefore, if the observed increment in going from three dimensions to four dimensions is not greater than .0675, there is little justification for retaining the fourth dimension.

In order to study the dimensionality problem more completely (Levine, 1978) and to provide more confidence in drawing inferences regarding the *structure* in the data, more information is needed about the behavior of R_1 for data that fit the INDSCAL model (MacCallum, 1981).

---

[1]Means and standard deviations of R_1 for 76 conditions are available from the author upon request.

References

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35*, 238–319.

Klahr, D. (1969). A monte carlo investigation of the statistical significance of Kruskal's nonmetric scaling procedure. *Psychometrika, 34*, 319–330.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika, 29*, 28–42.

Levine, D. M. (1978). A monte carlo study of Kruskal's variance based measure on stress. *Psychometrika, 43*, 307–315.

MacCallum, R. (1981). Evaluating goodness-of-fit in nonmetric multidimensional scaling by ALSCAL. *Applied Psychological Measurement, 5*, 377–382.

Montanelli, R. G., Jr., & Humphreys, L. G. (1976). Latent roots of random data correlation matrices with squared multiple correlations on the diagonal: A monte carlo study. *Psychometrika, 41*, 341–348.

Pruzansky, S. (1975). *How to use SINDSCAL: A computer program for individual differences in multidimensional scaling.* Unpublished manuscript.

Spence, I., & Ogilvie, J. C. (1973). A table of expected stress values for random rankings in nonmetric multidimensional scaling. *Multivariate Behavioral Research, 8*, 511–517.

Stenson, H. H., & Knoll, R. L. (1969). Goodness of fit for random rankings in Kruskal's nonmetric scaling procedure. *Psychological Bulletin, 71*, 122–126.

Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika, 42*, 7–62.

Wagenaar, W. A., & Padmos, P. (1971). Quantitative interpretation of stress in Kruskal's multidimensional scaling technique. *British Journal of Mathematical and Statistical Psychology, 24*, 101–110.

Acknowledgment

Author's Address

Send requests for reprints or further information to Hei-Ki Dong, The Ball Foundation, 800 Roosevelt Road, Building C, Suite 206, Glen Ellyn IL 60137, U.S.A.