Dimensionality of Hierarchical and Proximal Data Structures

David J. Krus and Patricia H. Krus
Arizona State University

The coefficient of correlation is a fairly general measure which subsumes other, more primitive relationships. At the fundamental classification level, similarities among objects and cladistic relationships were conceptualized as generic concepts underlying formation of proximal and hierarchical structures. Examples of these structures were isolated from data obtained by replicating Thurstone's classical study of nationality preferences and were subsequently interpreted.

The search for structure among the elements of data matrices is dependent on often tacitly assumed classifications of relationships to be analyzed. One of the more frequently assumed relationships is that of similarity and dissimilarity, as indexed by the (positive and negative) coefficient of correlation.

The classifactory principles inherent in the concept of correlation can be illustrated by considering Equation 1

$$r_{xy} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x \sigma_y} \qquad [1]$$

derived around the turn of the century by Pearson (1902, p. 292). Here, the product-moment coefficient of correlation is conceptualized in terms of added variances of the variables X and Y and their difference. Since the formation of a sum and a difference are two separate operations, it appears that correlation subsumes other, more primitive relationships.

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 3 Summer 1980 pp. 313–321
© Copyright 1980 West Publishing Co.

Consider a matrix **R** describing the correlation between variables X and Y as

$$\mathbf{R} = \begin{bmatrix} 1 & r \\ r & 1 \end{bmatrix} \qquad [2]$$

The characteristic equation of the matrix **R** can be written as

$$det(\mathbf{R} - \lambda \mathbf{I}) = 0 \qquad [3]$$

containing two eigenvalues λ_1 and λ_2. These eigenvalues can be computed as roots of the equation

$$\lambda^2 - 2\lambda + 1 - r^2 = 0 \qquad [4]$$

by the quadratic formula as

$$\lambda_1 = 1 + r \qquad [5]$$

and

$$\lambda_2 = 1 - r \qquad [6]$$

Variance, contributed by each principal component of the matrix **R** and indexed here as p and q can be computed by dividing the eigenvalues by the trace of **R** as

$$p = \lambda_1 / tr\,(\mathbf{R}) \qquad [7]$$

and

$$q = \lambda_2 / tr\,(\mathbf{R}) \qquad [8]$$

which for a case of bivariate relationship simplifies to

$$p = .5\,(1 + r) \qquad [9]$$

and

$$q = .5\,(1 - r) \qquad [10]$$

Equations 9 and 10 can be also written as

$$p = \frac{\Sigma(z_x + z_y)^2}{4N} \qquad [11]$$

and

$$q = \frac{\Sigma(z_x - z_y)^2}{4N} \qquad [12]$$

These expressions (Equations 11 and 12) are algebraically equivalent to Equations 9 and 10.

The coefficient of correlation is thus defined as a difference between *p* and *q* coefficients (i.e., $r = p - q$), which may account for its −1 +1 range. In Figure 1 both principal components of the matrix **R** were plotted for a binary data matrix, containing the equal number of *a*(1, 1) *b*(0, 0), *c*(1, 0) and *d*(0, 1) frequencies. In standard form these frequencies change to (1, 1), (−1, −1), (1, −1) and (−1, 1) response patterns, since standardization shifts the origin of the reference coordinates. The first principal component indexes the similarity (1, 1) and (0, 0) data patterns, and the second principal component indexes the dissimilarity (1, 0) and (0, 1) configurations of the data matrix elements.

Numerical Taxonomy

Properties of relationships of similarity and difference were primarily scrutinized within the framework of numerical taxonomy. With respect to biological classifications, two principal taxonomic relationships are usually hypothesized: (1) phenetic relationships based on the overall similarity of classified organisms and (2) cladistic relationships based on common lines of descent (Mayr, 1965; Sokal & Sneath, 1963).

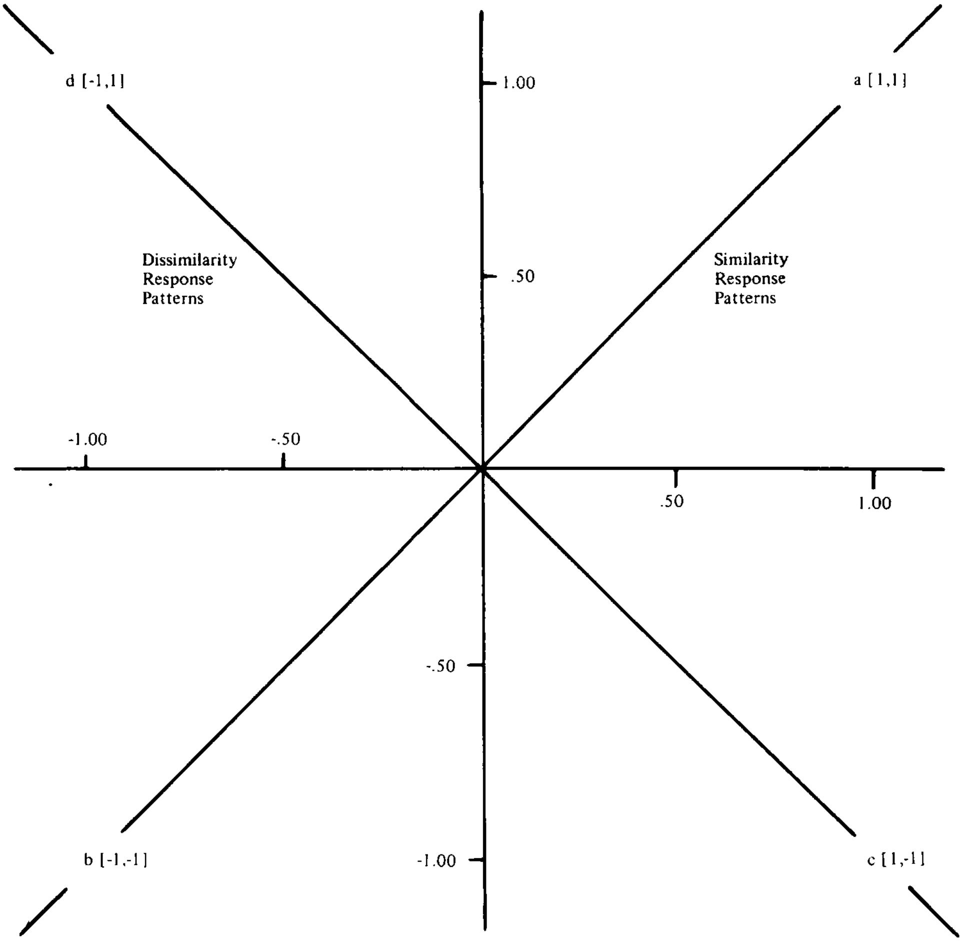
Pertaining to the above dichotomy, two elementary questions can be asked: (1) In what respect is A similar to B? and (2) Is A more similar to B than it is to C? Suggested answers to the first question underlie the logic behind various clustering algorithms; whereas the second question, establishing comparative similarities, is "fundamental to any attempt at clustering objects into homogenous groups" (Sokal, 1968, p. 178). The isolated phenetic relationships are frequently used for mapping the classified organisms into phenograms, a convenient two-dimensional representation of the results of a numerical classification. Cladistic relationships lead to construction of cladistic trees or cladograms, similar to dendrograms of ordinal test theory (Bart & Krus, 1973; Krus & Ceurvorst, 1979). Similar taxonomic relations are also used in statistical archaeology for the classification and chronological ordering of archaeological deposits (Robinson, 1951; Kendall, 1969, 1971).

Hierarchical and Proximal Relationships

At the binary data level the proximity question is related to the occurence of (1, 1) and (0, 0) data patterns. The measured objects or attributes either both possess or both lack a certain characteristic and are thus close, or similar, in this respect. The question of comparative similarities is related to the (1, 0) and (0, 1) response patterns. If these patterns can be ordered to form asymmetrical and transitive relationships, a hierarchical relationship of dominance is said to exist (Bart, 1976; Cliff, 1977, 1979; Krus & Bart, 1974). Obviously, other classificatory relations could be formed, for example, each of the four possible binary tuples could be considered separately.

Adopting the definition that (1, 1) and (0, 0) relations can be classified as proximal and (1, 0) and (0, 1) relations as hierarchical, the *p* and *q*

Figure 1
Schematic Representation of Similarity and Dissimilarity Relationships Subsumed by the Coefficient of Correlation



relationships that underlie the coefficient of correlation can be reconsidered. The *p* coefficient pertains to (1, 1) and (0, 0) relationships and, in this respect, is a measure of similarity among the data elements. The *q* coefficient, pertaining to (1, 0) and (0, 1) response patterns, however, is not a measure of dominance relationships, since the direction of dominance is cancelled by squaring the difference term in the numerator. The index *q* thus appears to measure the overall dissimilarity of data and not hierarchical dominance relations.

An index preserving the direction of the dominance was adapted from McNemar's (1947, p. 80) statistics for comparing two correlated frequencies and was applied to analysis of hierarchical relations as

$$\delta' = \frac{c - d}{(c + d)^{1/2}} \qquad [13]$$

where *c* and *d*, as previously, indicate frequencies of the (1, 0) and (0, 1) response patterns. An alternative index for continuous variables is

$$\delta_{ij} = \frac{(x_{ij} - x_{ji})/n}{(s_{ij}/n)^{1/2}} \qquad [14]$$

where X_{ij} and X_{ji} represent the cumulative difference between variables *i* and *j*, and S_{ij} stands for the standard deviation of a difference score between variables *i* and *j*.

These indices were used in both the binary and metric models of order analysis (Krus, 1977; Krus & Ceurvorst, 1979) and in various algorithms for tailored/adaptive testing programs as measures of hierarchical relations among the data elements (Cliff, Cudeck, & McCormick, 1979; McCormick & Cliff, 1977).

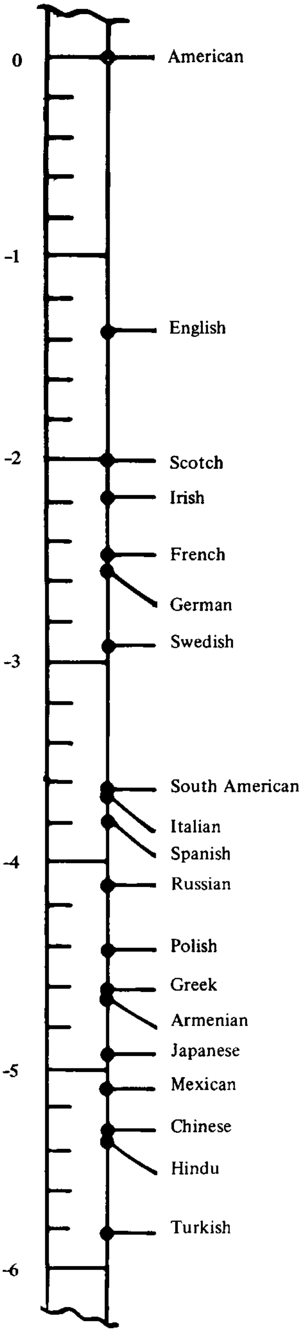
Hierarchical and Proximal Structures

How does the above taxonomy of data relationships pertain to analysis of real data structures? Before suggesting a tentative answer to this question, it should be kept in mind that assumptions about the nature of proximal and hierarchical structures are more specific than those made in the course of standard correlation-based multivariate analyses. Careful consideration should be given to the design of an experiment so that correct questions can be asked and an adequate structural type elected for the analysis.

Data

After some deliberation it was decided to illustrate the proposed taxonomy of relations by replicating Thurstone's (1928) classical study of nationality preferences. In the original study a single scale of nationality preferences was isolated, as shown in Figure 2; and yet, the presence of a proximal structure can be hypothesized in this type of data.

Figure 2
Scale of Nationality Preferences Adapted From Thurstone's 1928 Study



Following the administration of Thurstone's original questionnaire to 125 Arizona State University students in a pair comparison form, the number of times each nationality was selected in each pair comparison was counted and a standard subjects (n) × items (k) data matrix was obtained. This operation reduced the $n \times k(k-1)/2$ pair comparison data into a standard $n \times k$ form and avoided the problem of the ipsative nature of the pair comparison format. This matrix was factor analyzed using product-moment correlation coefficients and both p and q coefficients. Following factor analyses, the nationality preference data were analyzed by order analysis and by order analysis preceded by factor analysis of the proximity relations.

Factor Analyses

The factor analytic prediction from the proximal-hierarchical classification is that a solution based on product-moment coefficients should be approximated by a factorial structure based on p coefficients, whereas the structure based on q coefficients should contribute very little to the structural solution based on the "positive manifold" type of data and per se should be meaningless.

In the preliminary step, principal components for each solution were obtained. Cattell's (1966) scree test, Kaiser's (1960) factor extraction rule, and both geographical and sociopsychological meaningfulness were taken into consideration to

Table 1

Factor Loadings for the Nationality Preference Data Computed Using Pearson Product-Moment Coefficients of Correlation

Nationality	North	South	East	Orient
American	**.582**	.258	-.351	-.002
Swedish	**.679**	-.028	-.068	.134
Irish	**.674**	-.024	.086	-.137
English	**.591**	.210	.042	-.180
Scotch	**.547**	.024	-.066	.011
German	**.419**	-.070	.086	.042
French	**.360**	.321	-.227	.119
Italian	.113	**.292**	.072	.094
Spanish	-.098	**.713**	-.052	-.067
Mexican	-.044	**.656**	-.156	.153
S. American	.048	**.617**	.167	-.001
Greek	.257	**.456**	.368	.106
Russian	.114	.026	**.420**	.057
Armenian	-.087	.005	**.313**	-.021
Polish	.274	.013	**.463**	.066
Turkish	-.096	.256	**.576**	.051
Hindu	-.313	-.001	**.568**	.154
Chinese	-.094	.006	.178	**.643**
Japanese	.057	.186	.014	**.585**

determine the number of factors to be extracted. The first four principal components accounted for 74% of the variation in the matrix of *p* coefficients, and 49% of variance when the product-moment coefficients were analyzed. Four factors were extracted for both *p* and product-moment solutions. In the solution based on *q* coefficients the scree sharply dropped following the first eigenvalue; only the first factor was retained, accounting for 48% of variance.

Communality estimates from the squared multiple correlations were inserted in the diagonals of the matrices of *p, q,* and *r* coefficients and were iterated. The extracted principal factors were rotated orthogonally using Varimax. The resulting factor solutions are shown in Table 1 for the product-moment coefficients, in Table 2 for the *p* coefficients, and in the right column of Table 3 for the *q* coefficients.

For both the proximity and correlation matrices similar structures emerged, each extracted factor being readily identifiable in geographic terms, as in North, South, East, and West (Orient). No plausible hierarchical structure was implied in these solutions in either ascending or descending order of the factor loadings, nor did the addition of factors beyond the first one in the case of *q* coefficients suggest a structure.

Order Analyses

The order analytic prediction from the theory of proximal-hierarchical classifications is that the order analyses of the data based on pair comparison judgments should result in a hierarchical scale (cf. Krus & Krus, 1977) and that the order analysis preceded by factor analysis of proximal relations should reflect both the proximal and hierarchical data structures.

Table 2

Factor Loadings for the Nationality Preference Data Computed from p Coefficients

Nationality	North	South	East	Orient
American	**.734**	.423	.085	.190
Swedish	**.766**	.185	.248	.238
Irish	**.764**	.205	.332	.041
English	**.711**	.339	.306	.001
Scotch	**.700**	.242	.270	.171
German	**.615**	.180	.385	.180
French	**.577**	.468	.183	.264
Italian	.404	**.438**	.389	.221
Spanish	.278	**.747**	.325	.102
Mexican	.304	**.720**	.260	.271
S. American	.333	**.635**	.427	.082
Greek	.429	.463	**.507**	.115
Russian	.396	.192	**.619**	.161
Armenian	.304	.262	**.581**	.152
Polish	.477	.171	**.612**	.131
Turkish	.230	.368	**.708**	.117
Hindu	.127	.245	**.761**	.232
Chinese	.246	.244	.507	**.624**
Japanese	.346	.369	.347	**.497**

The order analysis started with the computation of the coefficients of dominance (as by Equation 14), followed by a search for connected, asymmetric, and transitive logical relations. At the $\alpha = .50$ level the order analytic solution converged into a single dimension, listed in the left column of Table 3. (For further discussion of the described procedure, see Krus, 1977). The scale did not discriminate between geographic proximities of the rated nations. It did, however, return an overall hierarchical structure congruent with a priori expectations.

The results of order analysis, preceded by factor analysis of p coefficients, are listed in Table 4. Both solutions were rotated by Varimax, and the final matrix of order loadings was rescaled with respect to the point of highest information density. Both the sociogeographic proximities and the hierarchical orders were reflected in the resulting structure.

Discussion

Analyses of both hypothetical and empirical data suggest the relevance of proximal and hierarchical classifications for data analysis. It appears that the explicit use of an implicit taxonomic relation may well clarify the interpretation of data. Further, the use of a new taxonomic relation may extend the scope of an analysis and may provide for an extended interpretative frame of reference.

Partial incongruencies in the joint proximal-hierarchical structure can be observed in Table 4, as for example, high scale values of Northern

Table 3

Hierarchical Structure Obtained by Order Analysis From the Nationality Preferences Data and the Factor Analytic Solution Based on q Coefficients

Nationality	Order Loadings	Factor Loadings
American	1.000	.667
Swedish	.837	.656
English	.742	.641
Irish	.693	.652
German	.645	.684
Scotch	.625	.670
French	.553	.657
Italian	.532	.669
Mexican	.463	.665
Spanish	.416	.681
Greek	.357	.600
Japanese	.315	.679
S. American	.272	.638
Russian	.212	.686
Chinese	.183	.708
Armenian	.134	.738
Polish	.083	.657
Turkish	.040	.669
Hindu	.000	.748

and Southern nationalities on the "Orient" dimension. These incongruencies were probably created by the sequential application of factor and order analyses and would possibly disappear if a coherent algorithm was elaborated to extract structures reflecting both of the discussed classificatory principles.

It is interesting to note that no proximity judgments were made directly by subjects. Since pair comparisons are a strictly hierarchical technique, the proximal structure, as reflected by the geographical polarization of the data, had to be introduced by juxtapositions of the hierarchical structures during the reduction of the data from the pair comparison form to the standard data matrix, as described above. The discrimination between proximal and hierarchial classifications may also contribute toward better understanding of the problem of comparability of factor and order analytic structures (cf. Bart, 1978; Krus, 1978).

The joint analysis and interpretation of proximal and hierarchical relations may result in new insights into the structure of the analyzed data. In the reanalysis of Thurstone's study of nationality preferences the authors attempted to demonstrate that neither the original interpretation of data as a hierarchical preference structure nor the factor analytic interpretation in terms of geographical proximities conveyed the full meaning of the integrated solution, containing the hierarchical classifications jointly with the geographic (proximity) classifications.

The preceding observations suggest that the complement of proximity structures by their hierarchical counterparts may be desirable if it is plausible to hypothesize the presence of the hierarchical structure. This may be achieved

Table 4

Joint Proximal and Hierarchical Structures of the Nationality Preference Data

Nationality	North	South	East	Orient
American	**1.000**	.861	.000	.751
Swedish	**.941**	.326	.290	.845
Irish	**.841**	.349	.547	.059
English	**.810**	.585	.469	.000
Scotch	**.782**	.436	.218	.557
German	**.703**	.250	.684	.581
French	.667	**.782**	.053	.643
Italian	.655	**.698**	.636	.663
Spanish	.235	**.997**	.290	.327
Mexican	.347	**.873**	.125	.680
S. American	.231	**.716**	.364	.138
Greek	.471	.602	**.763**	.353
Russian	.267	.082	**.712**	.399
Armenian	.149	.200	**.492**	.280
Polish	.369	.036	**.463**	.216
Turkish	.052	.127	**.189**	.089
Hindu	.000	.000	**.000**	.000
Chinese	.107	.172	.517	**.996**
Japanese	.338	.477	.271	**.915**

by separate analyses, or an extraction of joint proximal-hierarchical structure may be attempted.

It appears that prior to development of a formal algorithm for this type of analysis, the factor-order analytic concatenation may fill an interim need for this joint solution, especially when the presence of both structural types is strongly suspected. However, discussions of problems connected with the formulation of proximal and hierarchical classificatory principles are necessary to provide theoretical foundation for this future development.

References

Bart, W. M. Some results of ordering theory for Guttman scaling. *Educational and Psychological Measurement,* 1976, *36,* 141-148.

Bart, W. M. An empirical inquiry into the relationship between test factor structure and test hierarchical structure. *Applied Psychological Measurement,* 1978, *2,* 333-337.

Bart, W. M., & Krus, D.J. An ordering-theoretic method to determine hierarchies among items. *Educational and Psychological Measurement,* 1973, *33,* 291-300.

Cattell, R.B. The scree test for the number of factors. *Multivariate Behavioral Research,* 1966, *1,* 245-276.

Cliff, N. A theory of consistency of ordering generalizable to tailored testing. *Psychometrika,* 1977, *42,* 375-399.

Cliff, N. Test theory without true scores? *Psychometrika,* 1979, *44,* 373-393.

Cliff, N., Cudeck, R., & McCormick, D.J. Evaluation of implied orders as a basis for tailored testing with simulation data. *Applied Psychological Measurement,* 1979, *3,* 495-514.

Kaiser, H. F. The application of electronic computers to factor analysis. *Educational and Psychological Measurement,* 1960, *20,* 141-151.

Kendall, D. G. Some problems and methods in statistical archaeology. *World Archaeology,* 1969, *1,* 68-76.

Kendall, D. G. Abundance matrices and seriation in archaeology. *Zeitschrift für Wahrscheinlichkeitstheorie,* 1971, *17,* 104-112.

Krus, D. J. Order analysis: An inferential model of dimensional analysis and scaling. *Educational and Psychological Measurement,* 1977, *37,* 587-601.

Krus, D. J. Logical basis of dimensionality. *Applied Psychological Measurement,* 1978, *2,* 323-331.

Krus, D. J., & Bart, W.M. An ordering-theoretic method of multidimensional scaling of items. *Educational and Psychological Measurement,* 1974, *34,* 525-535.

Krus, D. J., & Ceurvorst, R. W. Dominance, information, and hierarchical scaling of variance space. *Applied Psychological Measurement,* 1979, *3,* 515-527.

Krus, D. J., & Krus, P. H. Normal scaling of the unidimensional dominance matrices: The domain-referenced model. *Educational and Psychological Measurement,* 1977, *37,* 189-193.

Mayr, E. Numerical phenetics and taxonomic theory. *Systematic Zoology,* 1965, *14,* 237-243.

McCormick, D. J., & Cliff, N. TAILOR-APL: An interactive program for individual tailored testing. *Educational and Psychological Measurement,* 1977, *37,* 771-774.

McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika,* 1947, *12,* 153-157.

Pearson, K. On the mathematical theory of errors in judgment with special reference to the personal equation. *Philosophical Transactions,* 1902, *198,* 235-299.

Robinson, W. S. A method for chronologically ordering archaeological deposits. *American Antiquity,* 1951, *16,* 293-301.

Sokal, R. R. Numerical taxonomy. In D. M. Messick (Ed.), *Mathematical thinking in the behavioral sciences.* San Francisco: Freeman, 1968.

Sokal, R. R., & Sneath, P. H. *Principles of numerical taxonomy.* San Francisco: Freeman, 1963.

Thurstone, L. L. An experimental study of nationality preferences. *Journal of General Psychology,* 1928, *1,* 405-425.

Acknowledgments

The authors thank Lynne Diaz for her help with the collection of data.

Authors' Address

Send requests for reprints or further information to David J. Krus, Director, University Testing Services, 302 Payne Hall, Arizona State University, Tempe, AZ 85281.