British Journal of Mathematical and Statistical Psychology (1985), **38**, 152–160 *Printed in Great Britain*

The effect of skewed distributions on matrix permutation tests

Katherine Faust and **A. Kimball Romney**

This paper examines the effect of skewed data distributions on matrix permutation tests of association. An empirical example is presented in which incorrect results are obtained using the Mantel statistic. Alternative procedures are considered, and two different data transformations are suggested which provide efficient and effective solutions. Methods for analysing related correlations as a means of evaluating the relative goodness-of-fit of alternative representations of a set of data are employed.

1. Introduction

While permutation tests of association (Hubert & Schultz, 1976) have become widely applicable in the social sciences, there is relatively little written about the sensitivity of the test to the distributional properties of the data. One exception is Dietz (1983) who has investigated the effect of a rank order transformation on permutation tests involving matrices with different degrees of skewness. We recently encountered a situation in which badly skewed data led to incorrect results. This note demonstrates the effects of such skewed distributions on permutation tests of association, and presents the results of two alternative data transformations for correcting the problems which arise. Procedures for comparing the goodness-of-fit of various alternative representations of the raw data are illustrated.

2. Permutation tests of association

2.1. *Matrix permutation tests*

Matrix permutation tests, introduced by Mantel (1967) and elaborated by Hubert and his colleagues (Hubert & Schultz, 1976; Hubert & Baker, 1978) test the correspondence between proximity matrices. 'A permutation distribution and an associated significance test are developed for the specific hypothesis of "no conformity" reinterpreted as a random matching of the rows and (simultaneously) the columns of one . . . matrix to the rows and columns of a second' (Hubert & Baker, 1978, p. 31). Mantel's index, Γ, is the measure of correspondence. For matrices **X** and **Y** with values x_{ij} and y_{ij} the measure is

$$\Gamma = \sum_{i,j} (x_{ij})(y_{ij}). \tag{1}$$

A significance test for Γ is made in relation to the permutation distribution, either by Monte Carlo sampling from the distribution, or as an approximate test based on the first two moments of the distribution. Hubert & Schultz (1976) present formulae for the mean and variance of the permutation distribution. However, Mielke has found that in many common situations the permutation distribution is asymptotically non-normal (Mielke, 1978, 1979). Both Mielke and Hubert caution against use of the normal distribution for hypothesis testing, and suggest the use of alternative sampling distributions or Monte Carlo procedures (Mielke, 1978, 1979; Hubert & Golledge, 1981).

2.2. *Permutation tests for related correlations*

Recently procedures have been developed to use permutation tests for related correlation coefficients. Such tests provide a straightforward means of evaluating the relative strength of two alternative predictors of the same dependent variable, or the relative fit of two alternative representations of the same data (such as a multidimensional scaling solution and a hierarchical clustering, or multidimensional scaling solutions in different dimensions). Wolfe (1976) has demonstrated that the test of the hypothesis that two variables U and V are equally good predictors of the dependent variables X (H_0: $\rho_{XU} = \rho_{XV}$) reduces to a test of the hypothesis that X is uncorrelated with the difference between U and V, $(U-V)$, if $\mathrm{var}(U) = \mathrm{var}(V)$. He shows that the correlation between X and the difference between U and V is zero if and only if the correlation between X and U equals the correlation between X and V.

Hubert & Golledge (1981) have generalized Wolfe's finding to the matrix case. For standardized matrices **U** and **V** with mean zero and variance one, the sample correlation, expressed in terms of pairwise correlations, r_{ij}, is

$$r_{X,U-V} = \frac{r_{XU}-r_{XV}}{[2(1-r_{UV})]^{\frac{1}{2}}}, \quad r_{UV} \neq 1. \tag{2}$$

In the permutation context this strategy allows comparison of one matrix, **X**, with the difference between two matrices (**U** and **V**) as a test of whether **X** is equally correlated with **U** and **V**.

2.3. *Monotone invariant permutation tests*

The problem of the effect of distributional properties of the data on permutation tests has received relatively little attention. In his pathbreaking article introducing permutation tests for matrix comparisons, Mantel (1967) employed a reciprocal transformation on the distances separating instances of disease outbreaks. Given the distribution of distances this transformation had the desired effect of increasing dispersion among cases within one mile of each other and decreasing the dispersion among cases further than one mile apart. Thus one would infer that various appropriate transformations of the data to ranks, reciprocals or even dichotomous variables are permissible. However, matrix permutation tests have been generally treated as if they are relatively insensitive to the distribution of the data.

Dietz (1983) provides one of the first investigations of the effects of alternative data transformations on permutation tests. She employs both empirical and simulated data to test the relative power of rank order and reciprocal transformations under two different degrees of skewness of the data. In addition, as an alternative to Mantel's index, and expanding a measure proposed by Hubert (1978), she uses three measures of association related to Kendall's τ. Each measure produces a count of the preponderance of concordantly ordered over discordantly ordered comparisons of pairs of observations on two variables. For matrices **X** and **Y** with cell values x_{ij} and y_{ij}, the measures are:

$$K = \sum_{i,j,k,l} \mathrm{sign}\,[(x_{ij}-x_{kl})\,(y_{ij}-y_{kl})], \tag{3}$$

$$K_C = \sum_{\substack{j<k \\ j\neq i\neq k}} \mathrm{sign}\,[(x_{ij}-x_{ik})\,(y_{ij}-y_{ik})], \tag{4}$$

$$K_U = \sum_{\substack{i<j,\, i<k<1 \\ i\neq j\neq k}} \text{sign}\,[(x_{ij}-x_{kl})\,(y_{ij}-y_{kl})], \tag{5}$$

$$\text{where sign}\,(x) = \begin{cases} -1 \text{ for } x < 0 \\ \;\;\,0 \text{ for } x = 0 \\ +1 \text{ for } x > 0 \end{cases}$$

K includes all observations in the matrices and is similar to a measure proposed by Hubert (1978, p. 185). K_C consists of comparisons among linked cells (i.e. cells in the same row). K_U compares unlinked cells (i.e. cells in neither the same row nor the same column). It is clear that these measures are invariant under monotonic transformations of the data.

Dietz (1983) compared the relative power of these three K measures with the Mantel index computed on three different treatments of the data (the raw data, the data after a reciprocal transformation, and the data transformed to a complete rank order). She simulated 2000 comparisons, half with slightly skewed distribution and half with J-shaped distribution of data values, varying, for each, the number of items and the amount of error in the comparison matrix. Her results indicate that for slightly skewed distributions two K measures, K and K_C, or the Mantel statistic computed on the rank ordered data, result in only slight loss in power as compared to the Mantel statistic computed on the raw data. For strongly skewed (J-shaped) distributions the same three measures were more powerful than the Mantel statistic computed on the raw data. The Mantel statistic on the reciprocally transformed data was less powerful than the other measures for the slightly skewed distribution, and only slightly more powerful than the two K measures or the Mantel statistic on rank orders for the J-shaped distribution.

2.3. *Monotone invariant tests for related correlations*

The noticeable effect of skewed distributions on the power of permutation tests underscores the importance of alternative permutation test methods. The additional problem of providing a monotone invariant measure for related correlations remains an open issue. Wolfe (1977) has suggested a measure based on the Kendall τ coefficient. He shows that the Kendall τ between a variable, X, and a new variable, Z, defined as the difference between two variables, U and V, 'measures the relative degree of correlation or association' between U and V and the variable X (1977, p. 507). For pairs of observations on X and Z Kendall's τ is the probability (Pr) of concordant ordering minus the probability of discordant ordering for pairs of observations. Substituting the original two variables U and V for Z gives

$$\tau_{X,(U-V)} = 2\,\Pr[(u_i-u_j)\,(x_i-x_j) > (v_i-v_j)\,(x_i-x_j)] - 1. \tag{6}$$

This approach to measuring the relative ordinal association of two variables with a third is problematic in that it incorporates the magnitude of the rank differences into the measure, is not monotone invariant, and the resulting statistic is not interpretable in a proportional reduction in error sense. A preferable approach to this problem would provide a measure which could be interpreted as the proportional reduction in error using V versus using U as a predictor of X. Hildebrand *et al.* (1977) have proposed a measure along these lines. Their partial ∇

measure for variables with ordered categories has the form

$$\nabla = 1 - \frac{\text{observed errors predicting } X \text{ knowing } U \text{ and } V}{\text{observed errors predicting } X \text{ knowing only } U} \tag{7}$$

(Hildebrand *et al.*, 1977, p. 72).

Employing this statistic in an empirical situation requires specification of which combinations of values on variables X, U and V are to be considered prediction errors.

Along with Kendall's partial τ (Kendall, 1975) and Somers' partial d (Somers, 1974), this measure falls in the general category of partial rank order correlation measures, having proportional reduction in error interpretations. In addition, Hubert & Golledge (1981) have suggested a measure which captures the degree to which a matrix **X** is ordered concordantly with matrix **U** rather than with matrix **V**.

These approaches provide intuitively appealing solutions to the problem of a monotone invariant measure for related correlations, and are general enough to accommodate other specifications of error reduction. The drawbacks of such measures in the current context are generalization to complete rank orders on **X**, **U** and **V**, and the extensive computation required for permutation tests. An economical alternative is to apply an appropriate transformation to the raw data prior to calculating either the difference between matrices or the usual Mantel statistic. A well-chosen transformation mitigates the effects of skewed distributions which otherwise may lead to erroneous conclusions.

3. Empirical example

The empirical example that prompted the present inquiry on the effect of skewed distributions on permutation tests deals with the selection of the best representation of the structures of two sets of proximity data. Inconsistent results are presented which lead to choice of one representation on the basis of the monotone relationship between the original data and the resultant proximities in the representation, and the choice of a different representation on the basis of the linear association between the proximities and the data (as in the usual Mantel statistic or the Pearson correlation coefficient). Three alternative data representations, multidimensional scaling, single link hierarchical clustering and complete link hierachical clustering, are compared.

The data, from Bernard *et al.* (1979), are frequencies of observed communications among members of two relatively bounded groups. The first set of data is from a technical research group. Thirty-four members of the group were observed by a researcher walking through the group's office space every half hour for a week. All pairwise interactions among individuals were tallied. The second set of data is from a ham radio club. Radio communications among 44 members of the club were monitored for 27 days and simultaneous radio use by pairs of individuals was tallied.

For each group the original data consist of an $N \times N$ symmetric matrix in which off-diagonal entries (x_{ij}) are tallies of observed communications or interactions between pairs of individuals (i and j). Diagonal elements (self-interactions) are undefined and are set to zero. The data which were analysed are a subset of these original data, excluding individuals who were observed infrequently. The raw interaction data were normalized to equal marginal interactions for individuals ($x_{.i} = x_{i.} = x_{.j} = x_{j.}$ for all i and j) using iterative proportional fitting (Romney *et al.*, 1973; Bishop *et al.*, 1975). The resulting cell values in the normalized matrices represent individuals' preference for interacting, apart from their differential observed interaction frequencies. (See Romney & Faust, 1982 for a further discussion of this procedure in the same context.)

Data were represented using three alternative techniques, multidimensional scaling as implemented by KYST (Kruskal *et al.*, 1973), single link hierarchical clustering and complete link hierarchical clustering (Jardine & Sibson, 1971). The multidimensional scaling results are represented by the $N \times N$ matrix of interpoint distances from the 2-dimensional solution (stress is 0·29 for the hams group and 0·27 for the technical group). Single and complete link hierarchical clustering results are represented by the $N \times N$ matrix in which the cell values indicate the lowest level at which two objects are clustered.

The substantive question is: which data representation best captures the pattern of interactions in each group? Although simple measures of association between the input preferences and the resultant proximities provide a descriptive comparison of the strength of alternative representations, they do not allow a decision of whether an observed difference in fit between two alternative representations is significant. Since non-metric multidimensional scaling and both single and complete link hierarchical clustering seek to represent the rank order of the original similarities by the rank order of the proximities in the representation, any evaluation of the relative goodness-of-fit of alternative representations should reflect the relative ordinal association between the input data and the representations. A measure of monotone association, such as Goodman & Kruskal's γ is a logical and reasonable choice for such a measure.

Goodman & Kruskal's γ and Pearson's correlation coefficient provide an initial summary of the goodness-of-fit for each representation. Table 1 shows the association

Table 1. Correspondence between original similarity data and three representations

		Representational method		
Data set	Measure	Multidimensional scaling	Complete link clustering	Single link clustering
Technical group	γ	−0·47	−0·41	−0·22
	r	−0·49	−0·81	−0·74
Ham radio operators	γ	−0·42	−0·27	−0·05
	r	−0·51	−0·65	−0·57

between the original similarity data and the proximities from each representation, for each of the two data sets. If we focus on the γ coefficients these results indicate that the interpoint distances from the multidimensional scaling provide a better fit to the rank order of the original similarities than do either of the hierarchical clustering solutions. The single link clustering is least good for both sets of data. However, the Pearson correlation coefficient leads to a different ordering of the three alternative representations, namely, the two hierarchical clustering results appear to be superior to the interpoint distances from multidimensional scaling in both data sets.

Since permutation tests usually employ the Mantel statistic, which, like the Pearson product moment correlation coefficient and other interval level statistics is sensitive to skewed distributions, it is important to explore the conditions under which such tests give results inconsistent with the ordinal relationships between the matrices being compared. Since the distribution of cluster levels from a hierarchical clustering is strongly negatively skewed, comparison with the input data, which is positively skewed, accentuates the negative association between these matrices due to the few extreme pairs of observations with high original similarities and low cluster levels. The problem of skewed distributions is more acute for the hierarchical clustering than for the multidimensional scaling representation.

Table 2. Descriptive statistics for data matrices before and after logarithmic transformation

Measure	Behavioural interaction		Multidimensional scaling		Complete link clustering		Single link clustering	
	Raw	Log	Raw	Log	Raw	Log	Raw	Log
Technical group data								
Mean	3·70	−1·03	1·32	0·44	24·50	1·25	24·13	1·32
Variance	42·78	6·52	0·34	0·18	19·08	0·39	21·68	0·41
Skewness	2·94	0·38	0·18	−1·15	−2·96	1·35	−2·60	1·30
Kurtosis	12·78	−1·63	−0·76	1·45	9·60	1·27	6·94	0·95
Ham radio operator data								
Mean	4·75	−0·02	1·33	0·44	18·89	1·17	17·26	1·60
Variance	57·99	4·38	0·32	0·14	13·74	0·38	13·44	0·28
Skewness	2·80	−0·05	0·06	−0·68	−2·63	1·23	−2·00	0·54
Kurtosis	11·17	−1·31	−0·94	−0·34	7·42	0·66	4·46	0·12

Many data transformations are available that reduce the effect of outlying **data** values. Two transformations, the log transformation, and a rank order transformation are employed here. Since the distances from the multidimensional scaling are not heavily skewed they were not transformed. The original similarities and the cluster levels from each of the clustering procedures were transformed both to ranks and to logs. To make the log transformation appropriate for reducing (positive) skewness in the hierarchical cluster levels they were transformed to similarities prior to taking logs. (A value of 1 was assigned to pairs of objects which were clustered at the highest level of the hierarchy, and a value of $N-1$ was assigned to the pair of objects clustered at the lowest level.) Table 2 presents descriptive statistics for the matrices for each of the two data sets, before and after log transformation.

The relative ability of each data representation technique to capture the original similarities was examined using the heuristic strategy presented by Hubert & Golledge (1981). Interpretability of the resultant Z score and its sign requires that the subtracted matrices meet two conditions: first they must have equal mean and variance, and second they must be ordered in concordant direction. Equal mean and variance are obtained by standardization of each comparison matrix to mean of zero and standard deviation of one prior to subtraction. Concordant ordering is assured by expressing all comparison matrices as dissimilarities. (Recall that effective application of the log transformation to hierarchical cluster levels required recoding them as similarities. Logs on these were therefore discordant with the distances from the multidimensional scaling and were transformed to dissimilarities by multiplying each

Table 3. Correspondence between original similarity data and differences between three representations

	Comparison matrix		
Data matrix and measure	MDS minus complete link	MDS minus single link	Complete link minus single link
Technical group data			
Raw data			
γ	−0·16	−0·20	−0·18
r	0·31	0·22	−0·11
$p<$	0·002	0·002	0·022
Log transformed data			
γ	−0·10	−0·18	−0·17
r	−0·003	−0·12	−0·16
$p<$	0·49	0·01	0·002
Rank ordered data			
γ	−0·12	−0·21	−0·10
r	−0·17	−0·31	−0·18
$p<$	0·002	0·002	0·002
Ham radio operator data			
Raw data			
γ	−0·16	−0·27	−0·17
r	0·12	0·04	−0·08
$p<$	0·03	0·18	0·12
Log transformed data			
γ	−0·12	−0·27	−0·16
r	−0·20	−0·42	−0·23
$p<$	0·002	0·002	0·002
Rank ordered data			
γ	−0·14	−0·28	−0·12
r	−0·24	−0·42	−0·17
$p<$	0·002	0·002	0·018

value by −1.) The matrix of original similarities was compared to the difference between each pair of represented distances, as original values, after log transformation and after rank order transformation.

Table 3 presents results of the comparison of the original similarity data with the difference between each pair of matrices. Pearson correlations and γ coefficients describe the degree of association between the original data and the difference between matrices. Since the original data are similarities (larger values indicate stronger preference for interaction) and the representations are distances (larger values indicate greater distance between objects in the representation) negative values for r and γ reflect superiority of the first representation in the subtraction, while positive r and γ reflect superiority of the second matrix. Comparison of matrices was done using the Quadratic Assignment Program (Hubert & Schultz, 1976). Probability values for the observed index are based on a Monte Carlo sample of 499 from the permutation distribution.

The effect of heavily skewed distributions in the original input similarities and the levels from each of the hierarchical clustering procedures is apparent in the results of the permutation tests. Comparison of the original similarity data with differences between represented matrices, employing the Mantel statistic on the untransformed data, shows an apparent superiority of complete link clustering over either single link clustering or multidimensional scaling. These results are consistent with the Pearson correlations between the original similarities and the representations, but are contradicted by the coefficients reflecting the ordinal association. Either the log transformation or the rank order transformation of the original similarities and the levels from the hierarchical clustering procedures prior to subtraction leads to comparisons which produce an ordering of the representation techniques consistent with the rank order correlations between the orginal data and the representations.

4. Discussion

These results demonstrate that even though matrix permutation tests provide an appropriate comparison distribution for each observed index of correspondence between matrices, such permutation distributions are not immune to the effects of skewed distributions in the data matrices themselves. In the absence of additional information about the ordinal relationships in the data, researchers can easily be misled employing the standard comparison index based on the original data. While monotone invariant comparison indices, in the tradition of proportional reduction in prediction error, may provide the ultimate solution, wisely chosen normalizing data transformations are an effective and inexpensive alternative.

Acknowledgements

Portions of this work were funded by NSF Grant Number SES-8320173 to A. K. Romney and W. H. Batchelder.

References

Bernard, H. R., Killworth, P. D. & Sailer, L. (1979). Informant accuracy in social network data IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks*, **2**, 191–218.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975). *Discrete Multivariate Analysis*. Cambridge, MA: MIT Press.

Dietz, E. J. (1983). Permutation tests for association between two distance matrices. *Systematic Zoology*, **32** (1), 21–26.

Hildebrand, D. K., Laing, J. D. & Rosenthal, H. (1977). *Analysis of Ordinal Data*. Beverly Hills, CA: Sage.

Hubert, L. J. (1978). Generalized proximity function comparisons. *British Journal of Mathematical and Statistical Psychology*, **31**, 179–192.

Hubert, L. J. & Baker, F. B. (1978). Evaluating the conformity of sociometric measurements. *Psychometrika*, **43**, 31–41.

Hubert, L. J. & Golledge, R. G. (1981). A heuristic method for the comparison of related structures. *Journal of Mathematical Psychology*, **23**, 214–226.

Hubert, L. J. & Schultz, J. (1976). Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, **29**, 190–241.

Jardine, N. & Sibson, R. (1971). *Mathematical Taxonomy*. London: Wiley.

Kendall, M. G. (1975). *Rank Correlation Methods*, 4th ed. London: Griffin.

Kruskal, J. B., Young, F. W. & Seery, J. B. (1973). How to use KYST, a very flexible program to do multidimensional scaling and unfolding. Unpublished manuscript. Available from Bell Laboratories.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Mielke, P. W. (1978). Clarification and appropriate inferences for Mantel and Valand's nonparametric multivariate analysis technique. *Biometrics*, **34**, 277–282.

Mielke, P. W. (1979). On asymptotic non-normality of null distributions of MRPP statistics. *Communications in Statistics—Theory and Methods*, **15**, 1541–1550.
Romney, A. K. & Faust, K. (1982). Predicting the structure of a communications network from recalled data. *Social Networks*, **4**, 285–304.
Romney, A. K., Kieffer, M. & Klein, R. E. (1973). A normalization procedure for correcting biased response data. *Social Science Research*, **2**, 307–320.
Somers, R. H. (1974). Analysis of partial rank correlation measures based on the product-moment model: Part one. *Social Forces*, **53**, 229–246.
Wolfe, D. A. (1976). On testing equality of related correlation coefficients. *Biometrika*, **63**, 214–215.
Wolfe, D. A. (1977). A distribution-free test for related correlation coefficients. *Technometrics*, **19**, 507–509.

Received 31 July 1984; revised version received 17 April 1985.

Requests for reprints should be addressed to Katherine Faust, Department of Psychology, University of Illinois, 603 E. Daniel St., Champaign, Illinois 61820, USA.