CLUSTERING RANK PREFERENCE DATA

James Beckett III, Bowling Green State University

INTRODUCTION

Situations for which rank preference data are appropriate are numerous. Problems involving N judges ranking k objects are common; the analysis of said problems being handled in straightforward fashion via the well-known Friedman χ_r^2 test. Large values of χ_r^2 (or equivalently Kendall's coefficient of concordance $W = \frac{\chi_r^2}{N(k-1)}$) indicate that the group of judges is basically in agreement on some consensus rank ordering of the objects. If χ_r^2 is not significant, we state that we have not found enough evidence to indicate that the ranks were not assigned randomly, i.e., no apparent difference in objects. After a significant χ_r^2, multiple comparisons [Miller (1966)] should be performed to find out which objects are judged different.

If judges can be a priori grouped into subgroups according to one or more classification factors, a more complete analysis is obtained through the use of ANACONDA (Analysis of Concordance) [Beckett & Schucany (1975)]. The concept of ANACONDA is based on partitioning the total agreement into the agreement (or disagreement) between and within the subgroups. The agreement between two subgroups of judges is measured by the statistic $\mathcal{L}$, which can be expressed as the inner product of the two rank sum vector $\underline{S}$ and $\underline{T}$, i.e., $\mathcal{L} = \underline{S}'\underline{T} = \sum_{j=1}^{k} S_j T_j$, where the elements of $\underline{S}$ and $\underline{T}$ are $S_j = \sum_{i=1}^{m} R_{ij}$, $j = 1,2,\ldots,k$ and $T_j = \sum_{i=1}^{n} R'_{ij}$, $j = 1,2,\ldots, k$ where $R_{ij}(R'_{ij})$ represents the rank given the j^{th} object by the i^{th} judge in group one (two). The small sample distribution has been tabulated [Schucany & Frawley (1973)] while the asymptotic distribution of $\mathcal{L}$ is normal for large m, n, and k. The linear scaling of $\mathcal{L}$, $\mathcal{W} = \frac{12\mathcal{L} - 3mnk(k+1)^2}{mn(k^3-k)}$ is often useful as a generalized coefficient of concordance such that $-1 \leq \mathcal{W} \leq 1$. Also it has shown that $\mathcal{W} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \rho_{ij}}{mn}$ where ρ_{ij} is Spearman's ρ between the i^{th} judge in group one and the j^{th} judge in group two, i.e., $\mathcal{W}$ is the average Spearman ρ between the two groups. Note that $\mathcal{W} = -1$ indicates disagreement between groups along with agreement within each group.

MOTIVATION

The underlying principle of cluster analysis is quite simple: identify subsets of individuals that tend to be relatively similar and group them together. There are two major steps common to the many methods used to cluster individuals: 1) computing quantitative indices of multivariate similarity between all pairs of individuals and 2) analyzing similarity matrices to identify homogeneous subgroups. Suppose N judges are ranking k objects and that we wish to cluster these judges on the basis of their preferences for the objects. We consider the N X N similarity matrix

$$\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & 1 & \cdots & \rho_{2N} \\ \vdots & \vdots & & \vdots \\ \rho_{N1} & \rho_{N2} & \cdots & 1 \end{pmatrix}$$

, where ρ_{ij} is the Spearman rank order correlation between the i^{th} and j^{th} judges. It is desirable to maximize the within cluster similarity and minimize the between cluster similarity. The minimization of $\mathcal{W}$ accomplishes both of these goals simultaneously.

The clustering procedures proposed herein should be considered as a logical third step in a comprehensive analysis of rank data following the Friedman χ^2 and multiple comparisons (if necessary). Clusters with interpretable or physical meaning might also indicate breakdowns of judges into subgroups such that an ANACONDA analysis might be illuminating for this data set or subsequent similar problems.

Suppose we have 6 judges ranking 3 objects A, B, and C in the following fashion.

	A	B	C
J1	1	2	3
J2	3	2	1
J3	1	2	3
J4	3	2	1
J5	1	2	3
J6	3	2	1

The obtained value of χ_r^2 is 0 indicating no agreement. However it is apparent the agreement of J2, J4, and J6 has been "cancelled" by the agreement (on the opposite ordering) of J1, J3, and J5. A conclusion of no agreement is clearly not appropriate if, for example, J1, J3, and J5 are women, while J2, J4, and J6 are men. In such a situation the subgroups should be considered separately; indeed, the value $\mathcal{W}$ for the male-female breakdown is -1 indicating agreement within each subgroup but on opposite orderings. We

examine the 6 x 6 similarity matrix as previously defined:

$$\begin{array}{c} \\ J1 \\ J2 \\ J3 \\ J4 \\ J5 \\ J6 \end{array} \begin{pmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \end{pmatrix}$$

Relabeling the judges 1,3,5,2,4,6 provides the clearer rearrangement of the similarity matrix below:

$$\begin{array}{c} \\ J1 \\ J3 \\ J5 \\ J2 \\ J4 \\ J6 \end{array} \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ \hline -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \end{array}\right)$$

The general idea herein is an extension of the above idea—a simple rearranging (relabeling) of the similarity matrix such that the elements in the upper right (lower left equivalently) corner of the matrix are small (ideally close to -1). The average of the elements in this block equals $\mathcal{W}$, the generalized coefficient of concordance between two groups of judges. $\mathcal{W}$ is (small) large if there is (dis)agreement between groups along with agreement within each group. Thus choosing members of clusters to minimize $\mathcal{W}$ will simultaneously maximize within cluster similarity and minimize betwen cluster similarity.

SPECIAL CONSIDERATIONS PARTICULAR TO RANK DATA

When one of the objects is clearly superior (or inferior) to the other k-1 objects, we must be wary of the high power of the Friedman test. Although it is proper that the Friedman test should reject, in considering the situation where we have k treatments of which one is a control (obviously inferior in e.g., agricultural or pharmaceutical studies) which has been added merely for reference, perhaps we should question our choice of objects if we are seeking a measure of agreement. The extreme high power of χ^2_r in this situation (one superior object) has been demonstrated by Beckett (1975). For example with a group of 6 judges ranking 5 objects with each judge recognizing the first object as clearly superior, the smallest value of χ^2_r attainable is 15 which itself is highly significant. Obviously clearly superior (or inferior) items are of no value in our clustering scheme and in fact their presence may mask some important inter-relationships between other objects and the potential subgroups. In such cases these "non-informative" objects should be ignored for purposes of clustering.

One could perform multiple comparisons (with small α) to separate or throw off objects clearly superior (or inferior) with significantly large or small rank totals. The remaining objects in the middle can be considered as the discriminating or "critical items". Regardless of the reduction (if any) of the objects to the critical subset, after the clusters are determined, for each cluster a cluster average rank profile should be presented based on all objects.

A usual problem in standard cluster analysis procedures is that larger problems quickly become too big for the computer. Here due to the data being in ranks, the data can be reduced to (k! + no. of unique tied rank orderings) since there are k! possible permutations of the ranks 1 to k. Disregarding ties, 1000 or more judges ranking 5 products can be reduced to at most 5! = 120 rank orderings each with a certain multiplicity.

PROCEDURES

With a divisive clustering algorithm we seek to divide the judges into two sub-groups or clusters. A feasible starting point might be to search for two pairs of judges who are as diametrically opposed as possible as measured by the smallest $\mathcal{W}$ (hopefully -1) obtained and use these pairs as the cluster nuclei to which judges will be added. Another approach which would be especially useful with a large number of judges would be to choose as the first cluster nucleus the observed consensus rank ordering of all N judges and to choose as the second cluster nucleus the conjugate rank ordering (opposite to the first cluster nucleus). Aside from being quicker, the latter approach would yield clusters representing the majority opinion (1st) and dissenting or minority opinion (2nd) as well as make the procedure less dependent on the order in which the data are read in.

After the two cluster nuclei are chosen, judges are added to clusters sequentially in such a way that $\mathcal{W}$ is minimized at each step. A stopping rule could be chosen (such as $\mathcal{W} \leq c$, for some chosen $c \leq 0$) or all judges could be forced into one of the two clusters. By stopping when $\mathcal{W}$ rises to some negative stopping value we would wind up with two clusters plus possibly some unclustered judges in the middle - these judges in the middle could be considered as making up a third cluster.

An agglomerative approach can be begun essentially by hand. The possible rank orderings can be grouped into classes. For example with k=4, Class 1 is chosen, say (1,2,3,4); then Class 2 contains those rank orders that can be obtained by one permutation of adjacent objects, i.e., {(2,1,3,4), (1,3,2,4), (1,2,4,3)}. Class 3 is obtained by two permutations of adjacent objects with reference to the Class 1 order or by trying one additional permutation referring to the rank orders in Class 2. For 4 objects we will have 7 classes; generally there are $\frac{k(k-1)}{2} + 1$ classes. The rank correlation between Class 1 and any one of rank orderings in Class 2 is .8 (generally $1 - \frac{12}{k(k^2-1)}$); the rank correlation between any two members within

class 2 is at least .40 $(1 - \frac{36}{n(n^2-1)})$. As long as we restrict our two clusters from having members from classes above and below the median class, W will remain below 0. This indicates that our cluster algorithm can be further streamlined by immediately adding to the cluster nuclei [Class 1 and Class $(\frac{k(k-1)}{2} + 1)$] those judges with rank preference orderings belonging to Class 2 and Class $(\frac{k(k-1)}{2})$, respectively.

EXAMPLES AND APPLICATIONS

Example 1 [Hollander & Wolfe, p. 140]. The data in Table 1 were obtained by Woodward (1970). Woodward, shortstop of the 1970 Cincinnati Reds National League baseball team, considered three methods of rounding first base. The best method is defined to be the one that, on the average, minimizes the time to reach second base.

TABLE 1. Rounding First Base Times

	Methods		
Players	Round Out	Narrow Angle	Wide Angle
1	5.40(1)	5.50(2)	5.55(3)
2	5.85(3)	5.70(1)	5.75(2)
3	5.20(1)	5.60(3)	5.50(2)
4	5.55(3)	5.50(2)	5.40(1)
5	5.90(3)	5.85(2)	5.70(1)
6	5.45(1)	5.55(2)	5.60(3)
7	5.40(2.5)	5.40(2.5)	5.35(1)
8	5.45(2)	5.50(3)	5.35(1)
9	5.25(3)	5.15(2)	5.00(1)
10	5.85(3)	5.80(2)	5.70(1)
11	5.25(3)	5.20(2)	5.10(1)
12	5.65(3)	5.55(2)	5.45(1)
13	5.60(3)	5.35(1)	5.45(2)
14	5.05(3)	5.00(2)	4.95(1)
15	5.50(2.5)	5.50(2.5)	5.40(1)
16	5.45(1)	5.55(3)	5.50(2)
17	5.55(2.5)	5.55(2.5)	5.35(1)
18	5.45(1)	5.50(2)	5.55(3)
19	5.50(3)	5.45(2)	5.25(1)
20	5.65(3)	5.60(2)	5.40(1)
21	5.70(3)	5.65(2)	5.55(1)
22	6.30(2.5)	6.30(2.5)	6.25(1)
	$R_1 = 53$	$R_2 = 47$	$R_3 = 32$

The value of χ_r^2 (adjusted for ties) is 11.1 which is significant at the .005 level. Hence we conclude the methods are not all the same with respect to speed. Multiple comparison of methods indicates Method 3 differs significantly from method 1 at the .01 experimentwise error rate. "Some" would continue and claim without statistical justification that method 3 is best. Regardless, the assumption of no block-treatment interaction (a fundamental assumption which is often overlooked) may be of greater concern here - Is one method best for all (types of) players? A quick perusal of the data shows players 1,6, and 18 performing opposite to the majority of the players. Perhaps method 1 really is best for these players due to some physical characteristics that they possess. Our cluster analysis provides the following result: Cluster 1: Players 2,4,5, 7-15,17,19-22; Cluster 2: Players 1,3,7,16,18 with $W = -.665$ highly significant and indicative of disagreement between the two groups. However, this disagreement has been manufactured and is meaningful only if the clusters are interpretable.

Example 2, [Gibbons, p. 353]. In a collaborative study of dry milk powders, six different types A to F are tested in each of seven different laboratories, and ranked in order of decreasing quality, that is, 1 = best, 6 = poorest. The results shown below are from Bliss (1967, p. 339).

TABLE 2.

	Rank for Powder					
Lab	A	B	C	D	E	F
1	2	3	6	1	5	4
2	2	1	3	4	5	6
3	1	2	3	5	4	6
4	2	3	1	5	6	4
5	4	1.5	1.5	6	3	5
6	1	3	4	5	2	6
7	2	4	1	5	6	3

Here W may be used as a check for an outlier (hospital 1). Employing hospital 1 as a singleton sub-group or cluster we obtain $W = -.1$ which is not significant. Had it been significant, an investigation of what makes hospital 1 significantly different from the others may have been profitable. However, not enough evidence is present to conclude all hospitals should not be considered as one group. In this situation W turns out to be a linear multiple of Page's L (1963).

SUMMARY AND COMMENTS

The informal procedures outlined herein should be useful in many of the problems for which a Friedman analysis is appropriate. Specifically, ANACONDA may be helpful in identifying agreement between and within subgroups of judges. Interpretable clusters may indicate future breakdowns or sub-groupings of judges as well as point out potential outliers and violations of the no block-treatment interaction assumption.

REFERENCES

Beckett, J. (1975). Some properties and applications of a statistic for analyzing concordance of rankings of groups of judges. Ph.D. dissertation, Southern Methodist University.

Beckett, J. and Schucany, W.R. (1975). ANACONDA: Analysis of concordance of g groups of judges. Proceedings of the Social Statistics Section of the American Statistical Assn., 311-313. (Presented at national ASA meeting in Atlanta, Aug. 75).

Bliss, C.I. (1967). Statistics in Biology. McGraw-Hill Book Co., New York.

Hollander and Wolfe (1973). Nonparametric Statistical Methods. John Wiley & Sons.

Gibbons, J.D. (1976). Nonparametric Methods for Quantitative Analysis. Holt, Rinehart, and Winston.

Miller, R.G. (1966). Simultaneous Statistical Inference. New York: McGraw-Hill Book Co.

Page, E.B. (1963). "Ordered hypothesis for multiple treatments: a significance test for linear ranks," Journal of the American Statistical Association, 58, 216-230.

Schucany, W.R. and Beckett, J. (1976). Analysis of multiple sets of incomplete rankings. Communications in Statistics, 5, 1327-1334. (Special issue: Recent theory and applications of nonparametric statistics.)

Schucany, W.R. and Frawley, W.H. (1973). "A rank test for two group concordance," Psychometrika, 38, 249-258.

Woodward, W.F. (1970). A comparison of base running methods in baseball. M.S. thesis, Florida State University.