COENOSES 12(2-3):127-130, 1997

A MEASURE OF DISCORDANCE FOR PARTIALLY RANKED DATA WHEN PRESENCE/ABSENCE IS ALSO MEANINGFUL

J. Podani

Department of Plant Taxonomy and Ecology, Eötvös University, H-1083 Ludovika tér 2, Budapest, Hungary.
E-mail: podani@ludens.elte.hu. Fax: 36 1 333 8764

Keywords: Braun-Blanquet scale, Goodman-Kruskal gamma, Kendall tau, Ordinal scale, Rank correlation, Ties.

Abstract: A dissimilarity measure for comparing objects described by ordinal variables with few states is derived from Kendall's tau and Goodman-Kruskal's gamma. The coefficient is calculated by examining all possible pairs of variables such that whenever ties appear, the presence/absence relationships become decisive. The measure is recommended for cases when many values in the data matrix are tied, and presence/absence relations are also of interest, e.g., in phytosociological classifications.

Introduction

It is generally acknowledged that we can distinguish among four measurement scales of variables: the nominal, ordinal, interval and ratio scales (Anderberg 1973). These imply in that order an increase of "information" conveyed by the data. The only property of the nominal scale is the distinguishability of states; the ordinal scale facilitates the interpretation of the sequence of states as well; on the interval scale the difference between states is also meaningful; whereas the ratio scale has a mathematical zero point allowing the operation of division. These properties impose strong constraints as to the selection of data analytical procedures. The limitations appear first when we wish to express the resemblance of objects, because the choice of the coefficient must be consistent with the scale of measurement. In this regard, the ordinal data type is the most difficult to handle. Unfortunately, it is a common practice to simplify it to the nominal (risking considerable loss of information) or to analyze ordinal data as if they were expressed on the interval scale. The second "solution" is even less admissible, since differences between the states of the ordinal variable are undefined. Notwithstanding its popularity in the biological sciences (e.g., the Braun-Blanquet scale and its derivatives widely used in vegetation studies, cf. Mueller-Dombois & Ellenberg 1974, Maarel 1979), the problems of the ordinal scale in determining the resemblance of objects have received relatively little attention (a noted exception is Dale 1989).

The ordinal type may be manifested in actual data in two different ways. In *fully ranked* data sets all scores pertaining to an object are different and can thus be ranked without ties. In this case the formal application of Spearman's rho or Kendall's tau to pairs of objects appears the best choice. However, full ranking is exceptional for objects: most of the ordinal scales have only a few states, so that in a data vector pertaining to an object there are inevitable ties. The number of tied scores in such *partially ranked* data can be extremely large, as in phytosociological tables with the above mentioned scale types.

To partially ranked data the Spearman rank correlation does not apply, since it is very sensitive to the presence of ties. However, there are other possibilities. Critchlow (1985) and Dale (1989) mention the Ulam distance, for example, which is defined as the minimum number of elementary changes to be applied to one (partial) rank order necessary to obtain the other. Notwithstanding their mathematical elegance, the calculation of such measures is relatively time-consuming. Goodall's probabilistic coefficient (see Goodall et al. 1991) can also be applied to ranked data; it is known for being always data set dependent: addition of new objects can substantially change the similarity structure, so the values are valid in the given context only. Diday & Simon (1976) proposed the use of the tie-adjusted Kendall's tau, which can be expressed as follows:

$$\tau_{jk} = 2(a-b)/\sqrt{([n(n-1)-2T_j][n(n-1)-2T_k])} \quad (1)$$

where n is the number of variables, a is the number of pairs of variables ordered for objects j and k identically, b is the number of pairs of variables that are reversely ordered in j and k. T_j and T_k are the numbers of tied variable pairs in objects j and k, respectively. A closely related formula is Goodman & Kruskal's gamma given by,

$$\gamma_{jk} = (a-b)/(a+b) \quad (2)$$

Both functions have the range of [-1,1], but for convencience the complements of (1)-(2) or their squared forms, can be used as a dissimilarity with the range of [0,2]. However, they are undefined for cases when all variables have a constant value in either or both objects. On the other hand, for fully ranked data they yield identical result. For example, with three variables as rows and the two objects as columns,

```
3 3
2 1                                    (3)
1 2
```

we get that a=2 (pairs 1-2 and 1-3) and b=1 (pair 2-3), yielding $\tau=\gamma=1/3$.

The denominator in equations 1-2 varies over object pairs, depending on the number of ties observed. As a consequence, the comparison of objects may be based on very different numbers of pairs of variables, which is undesirable. Both τ and γ disregard the information in the data for the tied pairs of variables. For example, the situations shown below,

```
0 0 and 2 1, also 0 1 and 1 2
1 0     1 1       0 1     1 2          (4a-d)
```

are treated equally. γ is unaffected by such pairs, whereas τ gets closer to zero as the number of such instances increases. As far as ordinality is concerned in the data, this may be acceptable, but one has the feeling that ties should also affect the result in some way. In many ordinal scale types (such as those mentioned above for phytosociological data) the value of 0 conveys a meaning essentially different from the others: it refers to the absence of a variable (e.g., species) whereas all other scores refer to species presence. With this in mind, cases 4a and 4b should not be weighted equally, because the pair of 4a contributes more to the dissimilarity of the objects than 4b. Also, the pair of 4c implies much higher dissimilarity than 4d, even though the sequential information is identical.

In this paper I will propose a formula for partially ranked data with many ties, which gives priority to ordering, but also considers presence/absence when no ordering appears. Its performance is illustrated by examples via comparisons with Kendall's tau and Goodman-Kruskal's gamma.

A New Measure of Discordance

The proposed formula is given by

$$DC_{jk} = 1 - 2(a-b+c-d)/[n(n-1)] \qquad (5)$$

where n, a and b are defined as above. c is the number of pairs of variables tied in both j and k, corresponding to joint presence or joint absence, as in the examples given below

```
1 1   or   1 2   or   0 0          (6a-c)
1 1        1 2        0 0
```

That is, such pairs of variables increase the similarity (decrease the dissimilarity) of the objects. d is the number of all pairs of variables that are tied at least for one of the objects being compared such that either one, two or three scores are zero. The following examples will clarify this:

```
1 0   or   1 1   or   1 0   or   0 1     (7a-d)
0 0        1 0        1 0        0 3
```

These pairs of variables indicate contradiction of the objects at least in presence/absence relations and will contribute to increased dissimilarity.

Note that $a+b+c+d=n$. Situations when the given pair of variables is tied for one of the objects only, and there are no absences, will be considered neutral, so their number does not appear in the numerator of (5). For example,

```
1 2    or    4 5                       (8a-b)
1 1          4 3
```

Although these pairs could be taken as indicators of joint presence, they are not counted because in one of the objects the values are ranked. (Otherwise 8a-b could not be distinguished from cases 6a-b for which no ranking appears at all.) The number of such neutral cases is implicitly considered, however, since the denominator of (5) is constant (see example 8 below). For fully ranked data, $c=d=0$ and $2(a+b)=n(n-1)$ so that DC becomes identical with the other two coefficients.

Examples

Artificial object pairs

The performance of the three measures discussed in this paper is evaluated first on the basis of simple artificial examples (Table 1). Object pairs 1-4 illustrate cases when all the three coefficients reach their extreme values, as expected. Pairs 5-7 illustrate situations when both τ and γ are undefined, whereas DC still produces meaningful results: pair 5 is the most similar (complete agreement in p/a relationships, no ranking in either objects), pair 6 has the intermediate value of 1 (full ranking in one of the objects, complete tie in the other), whereas pair 7 is the most dissimilar because of the disagreement in presence/absences. These three pairs illustrate the behavior of DC for completely tied scores.

Object pair 8 corresponds to a case when τ and γ produce maximum dissimilarity, because a=0 and b>0. DC, however, makes a correction for the three "neutral" pairs of variables, thus yielding a lower dissimilarity value. For object pair 9 γ remains to produce maximum dissimilarity (based merely on variables 3 and 4, yielding b=1), whereas the values of τ and DC are lower than the maximum. For τ it is because there is a correction for ties in the denominator, for DC it is because there is a correction in the numerator. Example 10 shows how these values are changed when a single value in the data

Table 1. Simple examples used in demonstrating the performance of three rank statistics (bottom rows). Asterisks indicate cases for which the function is undefined. The values of *a, b, c* and *d* are also given for your information.

		Object pairs										
		1	2	3	4	5	6	7	8	9	10	11
V	1	45	00	41	10	23	44	10	21	00	00	40
a	2	33	00	32	10	23	43	20	21	00	00	31
r.	3	22	22	23	01	23	42	10	12	01	01	01
	4	11	00	14	10	23	41	10	21	20	22	10
a		6	3	.	.	.	.	.	.	.	3	1
b		.	.	6	3	.	.	.	3	1	.	3
c		.	3	.	.	6	.	.	.	1	1	.
d		.	.	.	3	.	.	6	.	4	2	2

is modified (from 0 to 2): γ now takes maximum similarity, whereas the other two functions show a less abrupt change. Example 11 shows the behavior of functions in a general case. For *DC* this pair is as dissimilar as pair 9, for τ it is more dissimilar than pair 9, whereas for γ it is less dissimilar than pair 9, illustrating the contrasting behaviour of the coefficients.

Comparison of coefficients using field data

A large phytosociological data set containing Braun-Blanquet scores of 144 species in 127 sites from *Quercus ilex* forests in central Italy has been compiled from the literature and from own field work by P. Di Marzio (Univ. La Sapienza, Roma). This data matrix serves as a basis for the comparison of Kendall's tau, Goodman-Kruskal's gamma and *DC*. The coefficients were computed for all pairs of sites, yielding 8001 dissimilarity values for each coefficient. The three dissimilarity matrices were then compared numerically by the formal application of the product-moment correlation coefficient (*matrix correlation*, Sneath & Sokal 1973) and graphically by the *matrix plot* technique (Rohlf 1993) using the SYN-TAX 5.02 program package (Podani 1993).

The results are summarized in Figures 1a-c. In general, there is a close relationship among the rank statistics. Nevertheless, *DC* has the most unique behaviour, since its correlation with the other two functions is lower than the correlation between τ and γ. The graphic display confirms this observation, because the points are more scattered when one of the coefficients is *DC*, whereas τ and γ seem to have a more equivocal functional relationship (Fig. 1c). Despite the obvious nonlinearity between the coefficients, the linear correlation between τ and γ is quite high. The joint scatter of *DC* and τ (Fig. 1a) merits special attention: *DC* seems to be "classified" into evenly spaced groups so that for a given value of *DC* the τ measure varies considerably, and vice versa, especially in the middle of the range of the coefficients. In most cases, τ yields a higher dissimilarity score than *DC*. These observations are supported by analyses of other data sets which are not reported here.

Discussion

The discordance coefficient proposed in this paper attempts to combine two types of information, ordering of values and presence/absence such that rank order has priority, and presence/absence relationships are considered only if there are ties. Application to partially ranked data is limited to situations when the state of 0 refers to the absence and all other states indicate the presence of a variable. Such variables commonly appear in community studies, for example, in phytosociological analysis. The coefficient treats absence in the same way as most other dissimilarity coefficients: the potential reasons (see e.g., Green 1971) behind the absence of a species from a site are not distinguished from one another. Whenever coding is arbitrary, i.e. 0 has no specific meaning in contrast with other states, the use of *DC* is not recommended at all. In these cases τ, γ and other coding-in-

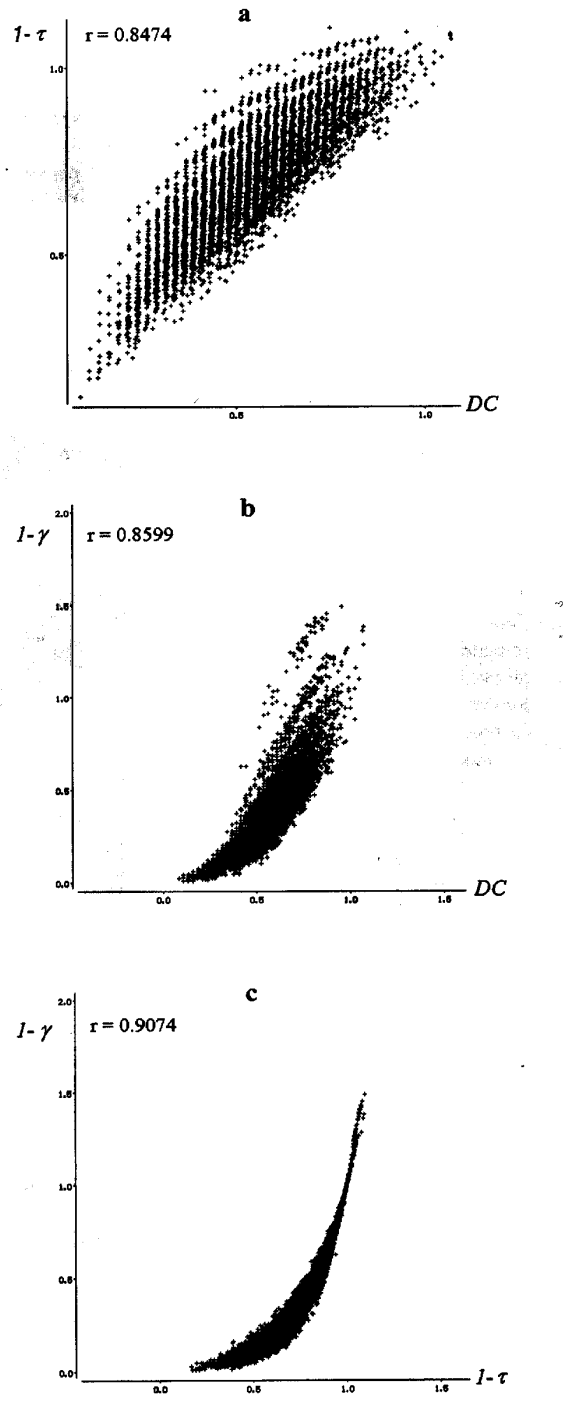

Figure 1. Graphical comparison of *DC*, 1-τ and 1-γ for the forest data. Matrix correlations are shown in top left of each diagram.

dependent measures, such as the Levenshtein formulae may be used.

Examples have demonstrated the following advantages of the new coefficient:

- There are no singularity problems, *DC* can be used in any circumstances, unlike τ and γ;
- *DC* has a constant denominator, so that all pairs of objects are compared on the same basis, unlike in τ and γ;
- In the numerator of *DC* the presence/absence relationships for tied scores are expressed in a very straightforward way. γ does not reflect these relationships at all, whereas τ implies a correction for ties in the denominator. This correction does not distinguish among the types of ties, whereas *DC* distinguishes three types: those that increase or decrease the dissimilarity and the neutral ones.

The metric properties of coefficients were not examined (although principal coordinates analyses of several *DC* matrices detected no negative eigenvalues, thus suggesting that *DC may* be a metric). Since the starting data are of ordinal type, the question whether or not the dissimilarity coefficient satisfies the metric axioms appears completely immaterial. If the user does not bother with the metric properties of the data, then why should he force requirements for metric axioms upon the subsequent data analytical steps? A more logical choice is that the matrix of rank statistics is examined by methods which consider only the rank order of the coefficients, i.e., ordinal methods are used throughout. Non-metric multidimensional scaling and ordinal cluster analysis procedures are obvious candidates for this purpose. It is not to say, of course, that matrices of rank statistics cannot be used by metric methods, but this choice is apparently less consistent with the startup situation. These problems will be examined in more detail in forthcoming papers.

Acknowledgements: This paper was written during my stay at the Department of Botany, University of Roma "La Sapienza". I am grateful to my host for the invitation and for the research facilities. Financial support also came from the Hungarian National Research Fund Grant (OTKA), No. T19364. I am also grateful to P. Di Marzio for placing the phytosociological data set to my disposal. I thank the referee for his comments on the manuscript.

References

Anderberg, M. R. 1973. Cluster Analysis for Applications. Academic, New York.

Critchlow, D. E. 1985. Metric methods for analyzing partially ranked data. Lecture Notes in Statistics 34. Springer, Berlin.

Dale, M. B. 1989. Dissimilarity for partially ranked data and its application to cover-abundance data. Vegetatio 82:1-12.

Diday, E. & J. C. Simon. 1976. Clustering methods. In: K. S. Fu (ed.): Digital Pattern Recognition. Springer, New York. pp. 47-94.

Goodall, D. W., P. Ganis & E. Feoli. 1991. Probabilistic methods in classification: a manual for seven computer programs. In: E. Feoli & L. Orlóci (eds.), Computer Assisted Vegetation Survey. Kluwer, Dordrecht. pp. 453-467.

Green, R. H. 1971. A multivariate statistical approach to the Hutchinsonian niche: Bivalve molluscs of central Canada. Ecology 52:543-556.

Maarel, E. van der. 1979. Multivariate methods in phytosociology, with reference to the Netherlands. In: M. J. A. Werger (ed.), The Study of Vegetation. Junk, The Hague. pp. 163-225.

Mueller-Dombois, D. & H. Ellenberg. 1974. Aims and Methods of Vegetation Ecology. Wiley, New York.

Podani, J. 1993. SYN-TAX. Version 5. User's Guide. Scientia, Budapest.

Rohlf, F. J. 1993. NT-SYS User's Manual. Exeter Software, Setauket, NY, USA.

Sneath, P. H. A. & R. R. Sokal. 1973. Numerical Taxonomy. Freeman, San Francisco.

Appendix: Presence/absence forms

Since presence/absence relations are also emphasized, for the three functions discussed in this paper I derived forms for the presence/absence case, i.e. for a data matrix with 0-s and 1-s. If the four values in the conventional 2x2 contingency table set up for two objects are labeled by *A, B, C* and *D*, that is $A+B+C+D=n$, we have the following relationships with *a, b, c* and *d*:

$a=AD$

$b=BC$

$c=0.5\ [\ A(A\text{-}1)+B(B\text{-}1)+C(C\text{-}1)+D(D\text{-}1)\ ]$

$d=AB+AC+BD+CD$

Then, after substitutions and rearrangements we get

$\gamma = (\ AD - BC\ ) / (\ AD + BC\ )$

which corresponds to Yule's second coefficient (cf. Anderberg 1973, p. 87). For the other two rank coefficients we have the following results:

$DC = 1 - [\ 2\ (AD\text{-}BC\text{-}AB\text{-}AC\text{-}BD\text{-}CD) + A^2+B^2+C^2+D^2 - n\ ] / [n(n\text{-}1)]$

and

$$\tau = \frac{2(AD-BC)}{\sqrt{(n^2-A^2-B^2-C^2-D^2+2AB+2CD)(n^2-A^2-B^2-C^2-D^2+2AC+2BD)}}$$

These formulae have no published counterparts in the family of presence/absence coefficients.