Extending Gower's general coefficient of similarity to ordinal characters

János Podani[1]

Summary

Podani, J.: Extending Gower's general coefficient of similarity to ordinal characters. – Taxon 48: 331-340. 1999. – ISSN 0040-0262.

The possibilities of calculating similarity based on ordinal characters are evaluated by distinguishing subtypes of the ordinal scale. Multivariate analysis is most problematic when ordinal variables appear together with other scale types in the data. This difficulty is solved by extending Gower's general coefficient of similarity to ordinal data types, facilitating cluster analysis and multidimensional scaling. Two alternatives, a non-metric and a metric version, are offered. The modified formula implies that ordinal variables are equally weighted with the others, and that partially and fully ranked data are both applicable, due to the inherent standardisation procedure. A morphological data set derived for the moss genus *Tortula* illustrates the new approach.

Introduction

Data tables in taxonomy and other fields of biology very often comprise heterogeneous characters describing the objects. There are several sources of this heterogeneity, different measurement units and different distributional properties of characters being the most frequent. From the viewpoint of multivariate analysis, however, an even more critical source of heterogeneity is the simultaneous presence of several measurement scales. Appearance in the same data matrix of the nominal, ordinal, interval and ratio scale variables (Anderberg, 1973), plus special cases of any scale type such as the presence/absence (Boolean) variables, is not allowed by the majority of multivariate exploratory techniques. There are three potential solutions of this "mixed data" problem. The first is to bring the different scales into the same scale, either losing information (when going "down" from the ratio scale towards the nominal) or incorporating external knowledge (when going "upwards"; Anderberg, 1973). Scale conversion is not without problems, and most users decide to retain the variables in their original form. Gordon (1981) suggests a second possibility, analysing the data separately for each variable type and then synthesising the results, but he himself is sceptical about that approach. A reasonable third solution is the use of specific coefficients that allow mixed data types. It works for methods which start from a resemblance matrix calculated among objects (e.g., in hierarchical clustering, multidimensional scaling). A quick overview of the taxonomic literature (e.g., Pimentel, 1979; Dunn & Everitt, 1982; Stuessy, 1990; Reyment, 1991) suggests that the best known and most widely used measure of this type is Gower's (1971) general coefficient of similarity, which has the additional advantage that it tolerates missing scores in the data matrix. Gower's index, in its original form, does not apply to ordinal data, which seems to be the obstacle preventing its general use in taxonomic and ecological surveys, as many biological characters imply ordinal

[1] Department of Plant Taxonomy and Ecology, Eötvös University, Ludovika tér 2, H-1083 Budapest, Hungary (e-mail: podani@ludens.elte.hu).

Table 1. Sample artificial ordinal data. This information assists the discussion of ordinal measures of similarity (see text)

OTUs:	1	2	3	4	5	6	7	8
1. Fully ranked variable	0	4	7	5	3	10	12	13
2. Fully ranked variable converted to ranks	1	3	5	4	2	6	7	8
3. Partially ranked variable	1	2	1	4	1	2	2	1
4. Partially ranked variable converted to ranks	2.5	6	2.5	8	2.5	6	6	2.5
5. T value	4	3	4	1	4	3	3	4

information. (Note that Sneath & Sokal, 1973, proposed to treat ordinal variables as nominal when using Gower's index, but this choice, as mentioned earlier, implies information loss.) Alternative formulations have also been proposed (Goodall, 1966, 1993; Lance & Williams, 1967, 1968, 1971; Parks, 1969; Burnaby, 1970; Anderson, 1971; Podani, 1980; Wishart, 1988). Most of them share with Gower's formula the property that ordered characters can only be incorporated with information loss or with "scaling up" to interval variables. Goodall's probabilistic coefficient is an exception, but this function assumes independence of variables – a criterion rarely satisfied in real data sets. A further problem is that self-similarity of an operational taxonomic unit (OTU) is not 1, but depends on the frequency of the OTU's character state in the entire sample. Nevertheless, Goodall's idea, that the number of values which lie between the two values for the two objects being compared be counted, proves to be useful. Whereas there have been significant developments in the area of ordinal data analysis (e.g., Critchlow, 1985; Dale, 1989; Diday & al., 1996; Monjardet, 1997), these methods do not treat mixed data types. Approaches to analysing ordinal data and mixtures of variables appear to be isolated. Therefore, I attempt to extend the applicability of Gower's formula to ordinal variables so as to offer a simple option of integrating information carried by mixed data. This paper gives the mathematical details and provides simple artificial and actual examples to illustrate the new method.

Types of ordinal characters

Ordinal variables cause headache to the data analyst because differences, products, and ratios of the states are not interpretable, whereas the operations subtraction, multiplication, and division are essential parts of computations. The difficulties are exacerbated when one considers that ordinal variables have several subtypes affecting the choice of method.

The first typification relates to a fundamental property of orderings: for fully ranked data all realised values representing the given variable are different, so that the objects can be unambiguously ordered based on that variable (Table 1, rows 1-2). This is not so with partially ranked data (Critchlow, 1985; Dale, 1989), in which the number of realised states of each variable is less than the number of objects, so that in the rank order of objects ties cannot be avoided: objects having the same score will take the same position in the ordering (Table 1, rows 3-4). The fully ranked type usually arises from direct observations of temporal sequences of objects, or from scaling down ratio scale variables to the ordinal type. Ordinal taxonomic characters usually belong to the second category, especially those that express more

or less subjective judgements on features which cannot be measured directly. For instance, descriptions involving statements like 'very few', 'few', 'some', 'many', 'very many' lead to ordinal characters (see a more concrete example below). Well-known examples from ecology are water, humidity, and acidity requirements of species on an arbitrary, empirically established scale ("indicator values", Mueller-Dombois & Ellenberg, 1974: 315).

The second categorisation of ordinal variables is based on their commensurability. Ordinal characters are said to be commensurable if all the scores can be ordered object-wise as well. That is, a variable can be used to rank the objects and, simultaneously, a given object can also be used to obtain an ordering of all the variables. Ranking is meaningful in both directions in the data matrix, as a striking manifestation of the attribute duality principle (Williams & Dale, 1965). On the contrary, non-commensurable ordinal variables provide explicit ranking of objects only; comparison of scores over an object is invalid.

If all characters in a data set are fully ranked, then commensurability applies after the scores are replaced by ranks. That all variables are fully ranked is exceptional, however, and does not mean automatically that each object will also be fully ranked. Partially ranked data may also satisfy the commensurability criterion if all variables bear, in some sense, a common relationship. For example, Braun-Blanquet's abundace/dominance scale and its derivatives, commonly used in vegetation science (e.g., Mueller-Dombois & Ellenberg, 1974), result in partially ranked commensurable characters: the AD scores of species (usually coded by 0, +, 1, 2, 3, 4, and 5). Looking at such a table by any variable (a species) we find that the sampling units can be ordered meaningfully starting from those from which the species is absent and ending with units in which the species has maximum abundance. Ranking scores by sampling units is meaningful again, because it expresses an order of importance of species at each site. In many cases, however, partially ranked data are also non-commensurable. In a taxonomic data matrix, for example, one character may refer to the density of hairs on the surface of the leaves (very dense, dense, moderately dense, hairy, scarcely hairy, glabrous) and another may indicate intensity of colour of the flower (from dark to pale). No matter how the data are coded, comparison of "pubescence" and "colour" for a plant individual is meaningless; so the object-wise comparison of two orderings would be unwise.

Admissible operations with ordinal data

A few comments are needed on the applicability of several resemblance coefficients to data sets containing ordinal variables only, to see what ideas already known from the literature can be adopted to develop an extension of Gower's formula. As said before, subtraction, multiplication, and division are not admissible in the treatment of raw ordinal data, and one may ask what kinds of operations have been or could be involved. For calculating Spearman's rank correlation (see e.g. Legendre & Legendre, 1983), the original scores need to be replaced by their ranks, and the formula is best suited to fully ranked data and to the comparison of variables. Formal application to objects, if each object generates full ranking of all variables, is also conceivable. This formula does use the operation of subtraction for rank scores, because the difference between two rank orders is interpreted as "the number of elementary changes to be made in the first rank order to move an element into the

position it takes in the second ordering". Kendall's τ is another well-known formula expressing similarity of two rankings (Diday & Simon, 1976), also based on the same idea, but less sensitive to the presence of ties. There are other approaches that endeavour to find the minimum number of moves necessary to transform one ranking into the other (Levenshtein measures; see Critchlow, 1985; Dale, 1989), a problem usually solved by combinatorial optimisation. Goodman and Kruskal's (1954) γ function may also be used for comparing objects. It is essentially the ratio "variable pairs similarly ordered for the two objects being compared" by "number of variable pairs that are ordered at all" (tied pairs are excluded). All these measures – if applied to a pair of objects – require commensurability of all variables, a condition usually not satisfied by taxonomic characters. Even when commensurability holds true, these measures are concerned with the comparison of full orderings. In contrast, Gower's formula compares objects variable by variable, so a different approach is needed. In developing this, I adopt the concept of taking minimum number of moves, but use it for two objects within the rank order that each variable implies.

A proposed extension of Gower's formula

Let $\mathbf{X} = \{x_{ij}\}$ be the data matrix with n rows (characters, variables) and m columns (objects, OTUs). For measuring similarity between objects j and k based on mixed variable types, with potentially missing scores, Gower (1971) proposed the following formula:

$$G_{jk} = \frac{\sum_{i=1}^{n} w_{ijk} s_{ijk}}{\sum_{i=1}^{n} w_{ijk}} \qquad (1)$$

where $w_{ijk} = 0$ if objects j and k cannot be compared for variable i because x_{ij} or x_{ik} is unknown.

– In addition, Gower defined, for binary variables (a):

$w_{ijk} = 1$ and $s_{ijk} = 0$ if $x_{ij} \neq x_{ik}$;
$w_{ijk} = s_{ijk} = 1$ if $x_{ij} = x_{ik} = 1$ or if $x_{ij} = x_{ik} = 0$, and double zeros (mutual absences) are included;
$w_{ijk} = s_{ijk} = 0$ if $x_{ij} = x_{ik} = 0$, and the double zeros are excluded from the comparison;

– for nominal variables (b):

$w_{ijk} = 1$ if x_{ij} and x_{ik} are known; then let
$s_{ijk} = 0$ if $x_{ij} \neq x_{ik}$;
$s_{ijk} = 1$ if $x_{ij} = x_{ik}$;

– and for variables measured on the interval and ratio scale (c):

$w_{ijk} = 1$ if x_{ij} and x_{ik} are both known, and $s_{ijk} = 1 - \{ | x_{ij} - x_{ik} | / (\text{range of variable } i) \}$.

The complement of the above formula is a dissimilarity measure. Note that for the presence/absence case, with double zeros included, the index reduces to the simple matching coefficient, with double zeros excluded we obtain the Jaccard index (Sneath & Sokal, 1973). For nominal variables, Gower's formula implies extension

of the simple matching coefficient to nominal data, and for interval and ratio scale variables the dissimilarity form corresponds to the mean character difference calculated such that each variable is standardised by range beforehand.

Here, the following procedure is added to the above definition:

- for ordinal variables, w_{ijk} is the same as in (c); all x_{ij} are replaced by their ranks r_{ij} determined over all objects (such that ties are also considered, as in rank correlation; see Table 1, rows 3 and 4); and then

$$s_{ijk} = 1, \text{ if } r_{ij} = r_{ik}, \tag{2a}$$

otherwise

$$s_{ijk} = 1 - \frac{|r_{ij} - r_{ik}| - (T_{ij} - 1)/2 - (T_{ik} - 1)/2}{\max\{r_i\} - \min\{r_i\} - (T_{i,\max} - 1)/2 - (T_{i,\min} - 1)/2} \tag{2b}$$

In formula (2b), T_{ij} is the number of objects which have the same rank score for variable i as object j (this group including j as well; Table 1, row 5); $T_{i,\max}$ is the number of objects which have the maximum rank, i.e., max $\{r_i\}$; and $T_{i,\min}$ is the number of objects with the minimum rank, i.e., min $\{r_i\}$, for variable i in the ordering. If there are no missing values for this variable, the denominator of formula (2b) simplifies to $m - T_{i,\max} - T_{i,\min} + 1$.

Verbal explanation may enhance interpretability of this formulation: the numerator is the minimum number of elementary steps (interchanges of neighbouring values in the ordering) required to put an object with the same value as x_{ij} into the position of another object which has the same value as x_{ik}. Table 1 has an example: the third row is a partially ordered variable, the scores being transformed into ranks as shown in row 4. The fifth row contains the T values, which are sizes of groups of objects with tied scores. In the example, $T_{i,\max}$ is 1 whereas $T_{i,\min}$ equals 4. From OTU 1 to 2 we need only one move, whereas from OTU 1 to 4 four steps are required. These are divided by the possible maximum (range), in this case 4, to yield 0.25 and 1.0, respectively, for the above two pairs. As seen, this is a dissimilarity for the given variable, hence the subtraction from 1 to get a similarity value. Each variable has the same importance, and variables with constant values are not allowed. This is in complete agreement with the treatment of other scale types in Gower's formula.

The use of formula (2b) implies that the number of objects with the same original score as objects j and k does not influence the result. Thus, the similarity of this pair derives directly from the number of other objects between them in the rank order and the numerator is always 1 if the original scores for j and k are neighbours on the realised ordinal scale. In a sense, formula (2b) provides a "nearest neighbour" measure of similarity in a [partial] rank order. As a serious consequence, the axiom of triangle inequality is not generally satisfied and the dissimilarity version is non-metric. For example, we need 4 moves from OTU 1 to OTU 4 (row 4, Table 1), which is larger than the sum of moves needed from OTU 1 to OTU 2 and from there to OTU 4 (1+1). This property may be critical only if the subsequent analysis is of a metric nature; ordinal procedures which consider only the rank order of input (dis)similarities are not concerned (e.g., single and complete link clustering, non-metric multidimensional scaling, minimum spanning trees). Nevertheless, as an actual example will demonstrate, violation of the triangle inequality by the ordinal

variables does not necessarily mean that the overall dissimilarities cannot be embedded into a Euclidean space.

If one wishes to metrise the initial ordinal space in order to obtain a metric coefficient, the relative rank differences given by the complement

$$S_{ijk} = 1 - \frac{|r_{ij} - r_{ik}|}{\max\{r_i\} - \min\{r_i\}} \quad (3)$$

may be incorporated into Gower's formula. For the above three pairs of OTUs (1-4, 1-2, and 2-4), the rank differences are 5.5, 3.5, and 2.0, respectively, showing that this formula obeys the metric axioms.

Obviously, equations (2b) and (3) are simplified greatly if there are no ties for variable *i*:

$$S_{ijk} = 1 - \frac{|r_{ij} - r_{ik}|}{m - 1} \quad (4)$$

If we consider equations (3) and (4), the idea to involve differences in ranks for two items within the same rank order becomes obvious. Formula (4) was suggested by Anderberg (1973: 129) to express simple linear agreement of ordered data in mixed data sets, showing that the case with rank differences is treated in the same way as the interval or ratio type variables. Note again that, whereas subtraction is inadmissible for raw ordinal scores, this operation is allowed (and used in many rank statistics) after the scores are converted into ranks.

For a data set comprising ordinal variables only, application of formula (2a-b) to each variable and summation over all variables derive a new similarity measure:

$$S_{jk} = 1 - \frac{1}{n}\sum_{i=1}^{n} S_{ijk} \quad (5)$$

Its range is from 0 to 1, indicating minimum and maximum similarity, respectively. Subtraction from 1 may be omitted, if one wishes to get a non-metric dissimilarity version for clustering or multidimensional scaling. The comparison of the performance of formula (5) with that of other ordinal measures is beyond the scope of the present paper and will be the subject of a different study.

An actual example

The use of the modified coefficient is illustrated using a small data set extracted from an extensive biometrical study of *Tortula* sect. *Rurales* De Not. *(Pottiaceae, Musci)* in the Carpathian Basin (Tóth, 1996). Seven morphological characters and 14 specimens (OTUs) representing 10 different taxa of seven species of the genus were selected (Table 2). The characters are measured on different scales; the presence of hydroid cells is binary nominal (0 = absent or 1 = present), the outline of leaves is multistate nominal (1 = oval, 2 = oblong, 3 = ovate, 4 = obovate, 5 = fiddle-shaped), the denticulation of hairs (1 = smooth hair, 2 = weakly denticulate, 3 = medium denticulate, 4 = strongly toothed) and the shape of leaf apex (1 = pointed, 2 = blunt, 3 = rounded, 4 = notched) are ordinal, whereas leaf length, leaf cell diameter, and the number of papillae per cell are ratio scale (continuous) variables. The latter three

Table 2. An actual morphological data matrix representing 10 taxa of *Tortula* as described in terms of seven morphological characters (for character states, see text). Taxa examined: *T. calcicolens* W. Kramer (C), *T. caninervis* (Mitt.) Broth. subsp. *caninervis* (CC), *T. caninervis* subsp. *spuria* (Amann) W. Kramer (CSS), *T. intermedia* (Brid.) De Not. (I), *T. norvegica* (Weber) Lindb. (N), *T. ruraliformis* (Besch.) Ingham (Rf), *T. ruralis* (Hedw.) P. Gaertn. & al. var. *ruralis* (R), *T. ruralis* var. *hirsuta* (Vent.) Podp. (RH), *T. ruralis* var. *submamillosa* W. Kramer (RS), and *T. virescens* (De Not.) De Not. (V).

	C	CC	CCS	I	I	N	N	Rf	RH	RS	R	R	V	V
presence of hydroid cells	0	1	1	0	0	0	0	0	0	0	0	0	0	1
outline of leaves	1	3	1	3	3	1	2	3	3	2	2	3	4	5
denticulation of hairs	4	2	4	2	2	3	2	2	2	2	1	2	4	4
shape of leaf apex	4	2	4	4	3	1	3	3	4	2	4	1	3	2
leaf length (mm)	2.0	2.25	2.25	2.0	2.13	2.38	2.13	2.25	3.25	3.13	2.5	1.88	2.0	2.38
leaf cell diameter (μm)	10.0	6.4	12.4	11.0	12.8	17.1	18.7	12.0	14.2	15.1	14.6	15.5	16.4	12.2
number of papillae per cell	4.0	2.0	3.2	2.6	4.0	4.3	5.4	3.6	2.6	3.9	3.9	4.9	3.8	4.0

were obtained as averages from ten replicate samples from the same herbarium capsule. There is considerable infraspecific heterogeneity in the genus, which is obvious even from this small subset of data. For this reason, the results cannot be used to draw far-reaching conclusions on the relative taxonomic positions of the OTUs.

Gower's formula was used to convert the raw data into dissimilarities which were in turn subjected to principal coordinate analysis (PCoA, or metric multidimensional scaling) to generate ordinations of OTUs (program SYN-TAX; Podani, 1997). To allow comparisons between the new options and former "alternatives", four variants of the dissimilarity matrix were computed as follows and were applied to the two ordinal variables: (a) neighbour interchange (equations 2a-b), (b) relative rank difference (equation 3), (c) simplification to nominal, and (d) arbitrary "conversion" to continuous scale. The PCoA solutions for the first two most important dimensions are shown in Fig. 1a-d. Notwithstanding the differences in the input dissimilarity matrix, these axes explain pretty much the same amount of variation in the data in all cases (42-43.4 %). This agreement allows direct visual comparison of ordination scatter diagrams. Interpretation is enhanced by outlines drawn around OTUs that represent the same species, regardless of their infraspecific status.

There is a high overall resemblance among the results. The OTUs appear to be arranged in the same manner along a horseshoe in all cases, this well-known "effect" being less conspicuous in Fig. 1a. The use of neighbour interchanges for the ordinal characters did not destroy the metric structure, because all eigenvalues were non-negative. (A negative eigenvalue did arise, as expected, when the analysis was restricted to the two ordinal characters.) The points representing *Tortula ruralis* are the most scattered in this ordination (Fig. 1a), which is in accordance with the relatively high within-taxon heterogeneity of ordinal characters (Table 2). In all other ordinations, these OTUs are much closer to one another, especially in case (d). The separation of taxa is the weakest when ordinal variables were treated as being nominal (Fig. 1c), showing the undesirable effect of this scale conversion.

Concluding remarks

Ordinal variables no longer present problems in calculating Gower's similarity based on mixed data: ordering information is retained in the comparison, and there is no need to simplify the scale to the nominal. The new addition to Gower's index involves a standardisation analogous to the treatment of ratio scale and interval scale variables, the minimum number of steps (neighbour interchanges) necessary to move one object into the position of the other in the ordering is divided by the maximum for the variable. Alternatively, relative rank differences can be incorporated into the formula. Counting these moves or rank differences is in fact the only "legal" manipulation with ordinal data, leading to an expression in which the whole set of objects may influence pairwise relationships. An advantage of this approach lies in its generality: fully and partially ranked data, commensurable and non-commensurable

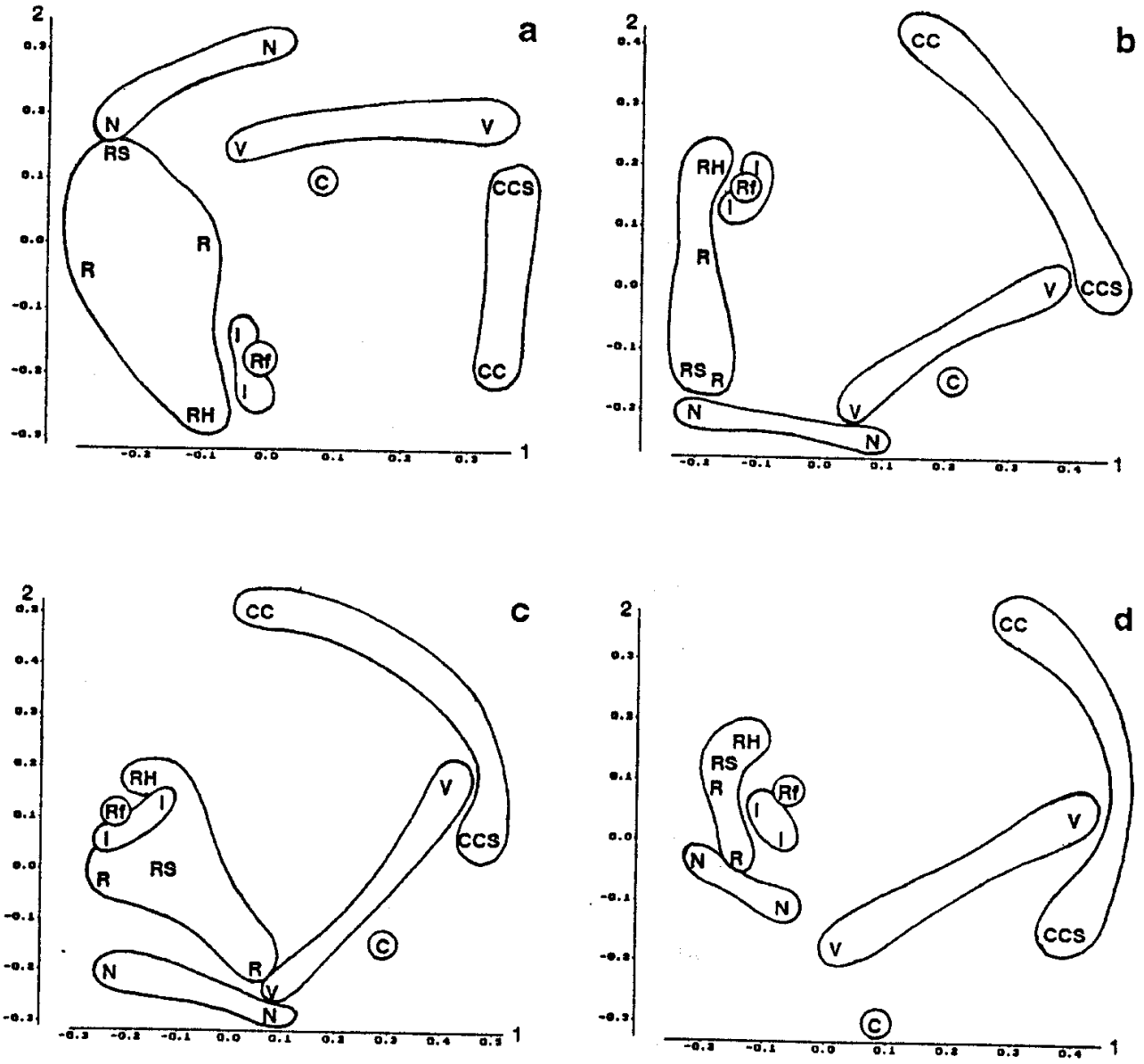

Fig. 1. Principal coordinates ordination, for dimensions 1 and 2, of *Tortula* taxa based on seven morphological characters. Partial dissimilarity for each ordinal variable is calculated using: (a), neighbour interchanges (equation 2a-b); (b), relative rank differences (equation 3); (c), conversion down to nominal scale, and (d), scores as if they were continuous. – See Table 2 for abbreviations of taxon names.

variables, as well as their mixtures, are all allowed, a feature not shared by former coefficients developed for ordinal characters. In the actual examples, the choice among four ways of treating ordinal characters did not prove to be critical, yet there were some differences suggesting that in more realistic studies with many more characters and OTUs the theoretically well-founded options should be preferred.

Computer applications

The SYN-TAX program package developed by the author (Podani, 1980, 1997) has been modified to accommodate the extended formula. The new option can be selected for dissimilarity-based analyses, such as agglomerative hierarchical clustering (metric and ordinal alike), partitioning, minimum spanning trees, neighbour joining, metric and non-metric multidimensional scaling. The dissimilarity matrix may also be saved for other types of analyses to be performed using different software. The input data must be arranged in a way such that variables are rows and objects are columns in an ASCII text file. First the binary variables are provided, then follow the multistate nominal, ordinal, and continuous variables, as exemplified by Table 2.

The choice between the ordinal (formulae 2a-b) and metric (formula 3) versions is facilitated as follows. If a variable is specified as being ordinal, formulae 2a-b are calculated automatically. If one wishes to use relative rank differences, then the ordinal scores must be replaced by their corresponding ranks in the input file, and the variable must be specified to be continuous.

Acknowledgements

The author is grateful to E. Feoli for discussions, to G. F. Estabrook and an anonymous referee for their constructive criticism of the manuscript. I thank Z. Tóth (Eötvös University, Budapest) for placing his *Tortula* data at my disposal. Financial support came from a Hungarian National Resarch Grant (OTKA, No. T19364).

Literature cited

Anderberg, M. R. 1973. *Cluster analysis for applications*. New York.

Anderson, A. J. B. 1971. Similarity measure for mixed attribute types. *Nature* 232: 416-417.

Burnaby, T. P. 1970. On a method for character weighting a similarity coefficient, employing the concept of information. *J. Int. Assoc. Math. Geol.* 2: 25-38.

Critchlow, D. E. 1985. *Metric methods for analyzing partially ranked data*. [Lecture notes in statistics, 34.] Berlin.

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

Diday, E., Lechevallier, Y. & Opitz, O. 1996. *Ordinal and symbolic data analysis*. New York.

– & Simon, J. C. 1979. Clustering analysis. Pp. 47-94 *in:* Fu, K. S. (ed.), *Digital pattern recognition*. New York.

Dunn, G. & Everitt, B. S. 1982. *An introduction to mathematical taxonomy*. Cambridge.

Goodall, D. W. 1966. A new similarity index based on probability. *Biometrics* 22: 882-907.

– 1993. Probabilistic indices for classification – some extensions. *Abstr. Bot.* 17: 125-132.

Goodman, L. A. & Kruskal, W. H. 1954. Measures of association for cross-classifications. *J. Amer. Statist. Assoc.* 49: 732-764.

Gordon, A. D. 1981. *Classification*. London.

Gower, J. C. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27: 857-871.

Lance, G. N. & Williams, W. T. 1967. Mixed data classificatory programs I. Agglomerative systems. *Austral. Computer J.* 1: 15-20.

– & – 1968. Mixed data classificatory programs II. Divisive systems. *Austral. Computer J.* 1: 82-85.

– & – 1971. A note on a new divisive classificatory program for mixed data. *Computer J.* 14: 154-155.

Legendre, L. & Legendre, P. 1983. *Numerical ecology*. Amsterdam.

Monjardet, B. 1997. Concordance between two linear orders: the Spearman and Kendall coefficients revisited. *J. Classific.* 14: 269-295.

Mueller-Dombois, D. & Ellenberg, H. 1974. *Aims and methods of vegetation ecology*. New York.

Parks, J. M. 1969. Classification of mixed mode data by R-mode factor analysis and Q-mode cluster analysis on distance function. Pp. 216-220 *in:* Cole, A. J. (ed.), *Numerical taxonomy*. London.

Pimentel, R. A. 1979. *Morphometrics. The multivariate analysis of biological data*. Dubuque.

Podani, J. 1980. SYN-TAX. Számítógépes programcsomag ökológiai, cönológiai és taxonómiai osztályozások végrehajtására [Computer program package for ecological, coenological and taxonomical classifications.] *Abstr. Bot.* 6: 1-158.

– 1997. SYN-TAX 5.1: a new version for PC and Macintosh computers. *Coenoses* 12: 149-152.

Reyment, R. A. 1991. *Multidimensional palaeobiology*. Oxford.

Sneath, P. H. A. & Sokal, R. R. 1973. *Numerical taxonomy*. San Francisco.

Stuessy, T. F. 1990. *Plant taxonomy*. New York.

Tóth, Z. 1996. *Rendszertani revízió és morfológiai bélyegek értékének vizsgálata a klasszikus taxonómia és a statisztika módszereivel a* Tortula *Hedw. sect.* Rurales *De Not.* (Pottiaceae, Musci, Bryophyta) *kárpát-medencei taxonjainál.* [*A taxonomic and statistical revision and reevaluation of morphological characters for* Tortula *Hedw. sect.* Rurales *De Not.* (Pottiaceae, Musci, Bryophyta) *taxa occurring in the Carpathian Basin.*] PhD Thesis, Eötvös University, Budapest.

Williams, W. T. & Dale, M. B. 1965. Fundamental problems in numerical taxonomy. *Advances Bot. Res.* 2: 35-68.

Wishart, D. 1988. Cluster analysis in information retrieval and diagnosis. Pp. 717-724 *in:* Bock, H. H. (ed.), *Classification and related methods of data analysis*. Amsterdam.