Appl. Statist. (1980),
29, *No.* 3, *pp.* 238–245

A Note on the Graphical Representation of Multivariate Binary Data

By C. F. Banfield† and J. C. Gower

Rothamsted Experimental Station, Harpenden, Herts, UK

[Received November 1978. Revised December 1979]

Summary

Various ordination methods for mapping n units characterized by v binary variables are in common use in which the distance between points P_i and P_j, representing units i and j, approximates some function (a similarity coefficient) of $(a_{ij}, b_{ij}, c_{ij}, d_{ij})$, the usual cell-counts in a 2×2 table. Ordination generally requires $(n-1)$ dimensions to represent the distances exactly, but the quantities $b_{ij}-c_{ij}$ can always be represented in one dimension. This leads to a simple graphical extension of ordination that helps with interpretation, reveals discrepancies, screens clustering possibilities and permits the recovery of approximations to all the (a, b, c, d)-values. Two examples illustrate the technique.

Keywords: MULTIVARIATE BINARY DATA; GRAPHICAL REPRESENTATION; ORDINATION; SIMILARITY

1. Introduction

In this note, we are concerned with the graphical representation of a set of n sample points, or units, each of which is described by V binary (0, 1) variables. The general approach to be followed is that of ordination, or multidimensional scaling, in which one attempts to represent the data points by a plot in a small number of dimensions. A good account of non-metric multidimensional scaling is that of Kruskal and Wish (1978). A general text for metric scaling does not seem available and several sources have to be cited, see Torgerson (1958); Gower (1966); Shepard and Carroll (1966); Sammon (1969); Blackith and Reyment (1971); Sibson, Bowyer and Osmond (paper not yet published). The number of matches and mismatches between the binary variables for any two given units can be represented as a 2×2 table, see Table 1. Thus there are a variables with value 1 for both units, and so on for b, c and d, with $a+b+c+d=v$. When we are concerned with units i and j, these quantities will be given suffices—thus a_{ij} represents the number of times unit i and unit j both have value 1. By modifying the terminology, the graphical displays discussed below apply equally to conventional 2×2 contingency tables classified by variables rather than by units, but this aspect is not pursued here.

Table 1
2×2 table giving the number of matches and mis-matches between two units

		First Unit	
		1	0
Second unit	1	*a*	*b*
	0	*c*	*d*

The resemblance between pairs of units may be quantified by calculating some similarity coefficient S that is a function of a, b, c, d. A variety of similarity coefficients in common use are

† Logica Ltd, Technical Group, 68 Newman St., London W1A 4SF.

discussed by Sneath and Sokal (1973). Simple examples of similarity coefficients that occur in the following are $S_{ij}=a_{ij}/v$ and $S_{ij}=(a_{ij}+d_{ij})/v$. Having selected a suitable coefficient, with n units, there are $\frac{1}{2}n(n-1)$ values S_{ij} $(i=2,\ldots,\mathrm{n};\ \mathrm{j}=1,\ldots,i-1)$ which can be formed into an $n\times n$ symmetric similarity matrix, usually with unit diagonal. Graphical representations of similarity matrices are used to help assimilate the complex net of relationships between the units. In ordination, the ith unit is represented as a point P_i on a map in two, or more, dimensions in such a way that the distance between two points P_i and P_j approximates the dissimilarity $1-S_{ij}$ or some function of $1-S_{ij}$. In ordinations based on similarity coefficients three stages of approximation are involved:

(i) The calculation of a, b, c, d loses information about which particular variables match or mis-match. Only the number of matches, etc. is preserved.

(ii) Whatever choice of similarity coefficient is made, information on all, or at least some, of the individual values of a, b, c, d is lost; only some function of them being known.

(iii) When only a few dimensions are used in an ordination, some information is lost on each similarity coefficient.

For example, with a simple matching similarity coefficient, $S_{ij}=(a_{ij}+d_{ij})/v$, we know only $(a_{ij}+d_{ij})$ and its complement $(b_{ij}+c_{ij})$ but not its component parts. When the attributes 0, 1 have a similar status and are labelled arbitrarily, this loss of information may not be serious. But when 0 implies lack of an attribute and 1 its presence, we would probably wish to take account of this difference in logical status and separate a_{ij} from d_{ij}. In general, all four components $(a_{ij}, b_{ij}, c_{ij}, d_{ij})$ may be required.

In this note we give a simple method for retaining all the information currently lost at stage (ii) of approximation and which, when used in conjunction with an ordination, recovers much of the information lost at stage (iii). The method relies on the result that to represent all $\binom{n}{2}$ values of $b_{ij}-c_{ij}$ exactly, only one dimension is needed. This follows easily from noting that $x_i = a_{ij}+c_{ij}$ is the number of 1-attributes (i.e. presences) for the ith unit, for all j, and that $x_j = a_{ij}+b_{ij}$ is the number of 1-attributes for the jth unit, for all i. Consequently $x_j-x_i = b_{ij}-c_{ij}$, and all $\frac{1}{2}n(n-1)$ differences $b_{ij}-c_{ij}$ may be represented by the one-dimensional plot of the values x_i $(i = 1,2,\ldots,n)$ which are merely the row-sums of the $n\times v$ binary data matrix.

The n x-values together with the $\frac{1}{2}n(n-1)$ similarity coefficients S_{ij} and $v = a_{ij}+b_{ij}+c_{ij}+d_{ij}$, allow all $2n(n-1)$ values $a_{ij}, b_{ij}, c_{ij}, d_{ij}$ to be evaluated. The details of evaluation depend on the choice of similarity coefficient and are illustrated here for just one choice. Suppose $S_{ij} = a_{ij}/v$. That is similarity is expressed as the proportion of "positive" matches in each comparison. Then

$$vS_{ij} = a_{ij};\quad x_i = a_{ij}+c_{ij};\quad x_j = a_{ij}+b_{ij};\quad v = a_{ij}+b_{ij}+c_{ij}+d_{ij}. \tag{1}$$

Hence

$$a_{ij} = vS_{ij};\quad b_{ij} = x_j-vS_{ij};\quad c_{ij} = x_i-vS_{ij};\quad d_{ij} = v-x_i-x_j+vS_{ij}. \tag{2}$$

Similar trivial calculations relate the a, b, c, d-values to other choices of similarity coefficient (see, for example, equation (5), below).

There is a link between x-values and the general canonical analysis of asymmetry discussed by Gower (1977) and Constantine and Gower (1978). Suppose $\mathbf{T}$ is the square matrix whose elements t_{ij} are defined as the number of properties possessed by i but not by j. Thus in terms of Table 1, $t_{ij} = c_{ij} = x_i-a_{ij}$ and $t_{ji} = b_{ij} = x_j-a_{ij}$. Then the above has shown that the skew-symmetric matrix $\mathbf{T}-\mathbf{T}'$ has elements x_i-x_j, and may be written $\mathbf{x1}'-\mathbf{1x}'$ where $\mathbf{1}$ is a vector of units and $\mathbf{x}$ is a vector of x-values. This is the simple canonical form of the special skew-symmetric matrix $\mathbf{T}-\mathbf{T}'$.

Apart from the simple but useful properties of the x_i discussed above, it is informative to superimpose them on ordination diagrams, when this can be done without confusion. The interpretation of such diagrams is discussed in the next section.

2. Combining x-Values with an Ordination

It is our intention here not to discuss particular ordination methods but to illustrate general principles of how x-values can be combined usefully with a wide range of ordinations. To be useful the ordination method selected should approximate adequately the similarities S_{ij}, so that these values can be reliably determined by direct measurement on the resulting map. For illustrative purposes we assume throughout this section that a matrix of simple matching coefficients $(a_{ij}+d_{ij})/v$ is ordinated by a principal co-ordinate analysis (Gower, 1966). We recall that this is equivalent to a principal components analysis of the given binary data. The resulting ordination diagrams give points with distances δ_{ij} approximating $(\frac{1}{2}v)^{\frac{1}{2}}\Delta(P_i, P_j)$ where

$$\Delta(P_i, P_j)^2 = 2(b_{ij}+c_{ij})/v. \tag{3}$$

To represent all $\binom{n}{2}$ distances exactly, $n-1$ dimensions will normally be needed, but good approximations δ_{ij} to Δ_{ij} in many fewer dimensions can often be found. This contrasts remarkably with the one dimension required to represent the $\binom{n}{2}$ values of $b_{ij}-c_{ij}$. For two points P_i, P_j with measured distance δ_{ij} in the ordination, we have

$$z_{ij} = \delta_{ij}^2 \doteqdot b_{ij}+c_{ij} \tag{4}$$

which together with $x_i = a_{ij}+c_{ij}$, $x_j = a_{ij}+b_{ij}$ and $v = a_{ij}+b_{ij}+c_{ij}+d_{ij}$, allows all four components $(a_{ij}, b_{ij}, c_{ij}, d_{ij})$ to be recovered, depending only on the accuracy of (4). Explicitly:

$$\left.\begin{aligned} x_i+x_j-z_{ij} &= 2a_{ij}, \\ -x_i+x_j+z_{ij} &= 2b_{ij}, \\ x_i-x_j+z_{ij} &= 2c_{ij}, \\ 2v-x_i-x_j-z_{ij} &= 2d_{ij}. \end{aligned}\right\} \tag{5}$$

By replacing z_{ij} by $\frac{1}{2}v\Delta(P_i, P_j)^2$ equation (5) gives the exact formulae relating the a, b, c, d-values to the simple matching coefficient, and may be compared with (2).

Because a, b, c, d are non-negative, it follows that when P_i and P_j are close in an accurate ordination, then δ_{ij} is small and therefore b_{ij} and c_{ij} are both small; hence x_i and x_j will both approximate a_{ij}. An appreciable difference between x-values of neighbouring points means that the ordination is distorted in that locality. If a set of points form a compact cluster, then all points in the set should have approximately equal x-values. The converse is not true but if any pair of x-values differ appreciably then the corresponding points must be distant and cannot belong to the same compact cluster. This suggests a very simple informal test for assessing clustering potential: if the x-values themselves do not aggregate into one or more clusters then there is little point in proceeding further. However, if the x-values do cluster, no compact multidimensional clustering necessarily follows. The choice of simple matching coefficient to measure similarity is not essential to this argument, only the adjacency of a pair of points implies that b_{ij} and c_{ij} are small.

The association between adjacent points and equality of x-values also suggests that contours linking points of constant x-value could be informative. Pairs of points on the same contour correspond to 2 × 2 tables in which $b = c$ although the actual common value of b and c will, in general, differ for different pair of points on the same contour. Contours of x-values are illustrated in the second example of Section 3.

Suppose P_0 and P_1 are points corresponding to units all of whose attributes are zero and unity respectively. Then with the simple matching coefficient (3) $\Delta(P_0, P_i)^2 = 2x_i/v$ where the factor $2/v$ can be absorbed into the scaling of the plot. This shows that contours of constant x-values are hyperspheres, centre P_0, and therefore in an undistorted plane-ordination these contours should be sets of concentric circles or, when P_0 is distant from the other points, sets of parallel lines. When the approximate position of P_0 is itself included in an ordination, either by operating on a sample of size $(n+1)$, or by superimposing P_0 on the n-sample ordination

(Gower, 1968), the squares of the plotted lengths $P_0 P_i$ approximate the value x_i. The superposition technique allows the residual distance of P_0 from the ordination space to be calculated, which indicates whether or not adequate co-ordinates of P_0 can be placed on the same map as the other points.

Tables with zero cells are important in some applications. We have

$$\begin{aligned} &\text{if } a_{ij} = 0 \quad \text{then angle } P_i P_0 P_j = \pi/2, \\ &\text{if } b_{ij} = 0 \quad \text{then angle } P_0 P_i P_j = P_1 P_j P_i = \pi/2, \\ &\text{if } c_{ij} = 0 \quad \text{then angle } P_0 P_j P_i = P_1 P_i P_j = \pi/2, \\ &\text{if } d_{ij} = 0 \quad \text{then angle } P_i P_1 P_j = \pi/2. \end{aligned}$$

Thus 2 × 2 tables with zero cells have simple geometrical interpretations in the full dimensional space, but these may not survive approximation in fewer dimensions, unless the ordination is a good one. In particular if P_0 and P_1 are *any* pair of complementary units then angle $P_0 P_i P_1 = \pi/2$ for all i, showing that in a perfect plane-ordination all units must lie on a circle with diameter $P_0 P_1$. There seems to be a case here for studying ordination procedures that preserve angles, especially right angles, rather than distances.

2.1. *Remarks on Choice of Ordination and Similarity Coefficient*

It has already been noted that neither the choice of similarity coefficient nor the choice of method for representing the values of the similarity matrix as functions of distances on a map, that is the ordination method, are critical. The combination of a simple matching coefficient with a principal components analysis discussed above is just one pair of choices. When positive matches are of special interest, one possibility is to choose $\Delta^2(P_i, P_j) = 2(1 - a_{ij}/v)$, which gives a Euclidean representation (Gower, 1971) but even non-Euclidean distances may be useful, as is shown in the first example below. Any ordination method that gives good approximations to the chosen similarities is acceptable and this includes non-metric methods for which measured similarities S^*_{ij} are monotonically related to observed similarities S_{ij} via a known empirical monotonic transformation function that can be used as a calibration curve (see, for example, Kruskal and Wish, 1978). However, metric ordination methods are more likely to preserve useful geometrical properties, such as those which were readily determined in the previous section for the simple matching coefficient. Of course different choices of similarity coefficient will give rise to different geometrical properties. For example, choosing $S_{ij} = a_{ij}/v$ implies that for P_i and P_j close, not only are b_{ij} and c_{ij} small, as with the simple matching coefficient, but also d_{ij} is small. It is a straightforward matter to investigate the properties associated with different coefficients.

2. Applications

3.1. *Ordination of the* Phthiracaroidea (*Mites*)

Sheals (1969) observed 128 characters for 53 species of the super family *Phthiracaroidea* from which we selected 46 two-state characters with no missing information. An ordination with distance defined by equation (3) was obtained by direct principal components analysis of the presence–absence data-matrix but the two-dimensional fit was only fair, accounting for 68 per cent of the total square distance. Also using (3) is slightly inconvenient for our purposes because to recover approximations to $b_{ij} + c_{ij}$, fitted distances δ_{ij} have to be squared. Therefore we decided to consider using a Hamming distance

$$\Delta(P_i, P_j) = (b_{ij} + c_{ij})/v. \tag{6}$$

Equation (6) is a special case of the city-block metric and unlike (4) derives from no real $(n-1)$-dimensional Euclidean configuration (Gower, 1971). This non-Euclidean property can have surprising effects. However, with these data a principal co-ordinate analysis of the distance-matrix yields acceptably small negative roots and a good two-dimensional real Euclidean fit ($\lambda_1 = 2{\cdot}28$, $\lambda_2 = 0{\cdot}21$, sum of positive roots = 2·80, sum of negative roots = −0·13) which is exhibited in Fig. 1, together with the x-values.

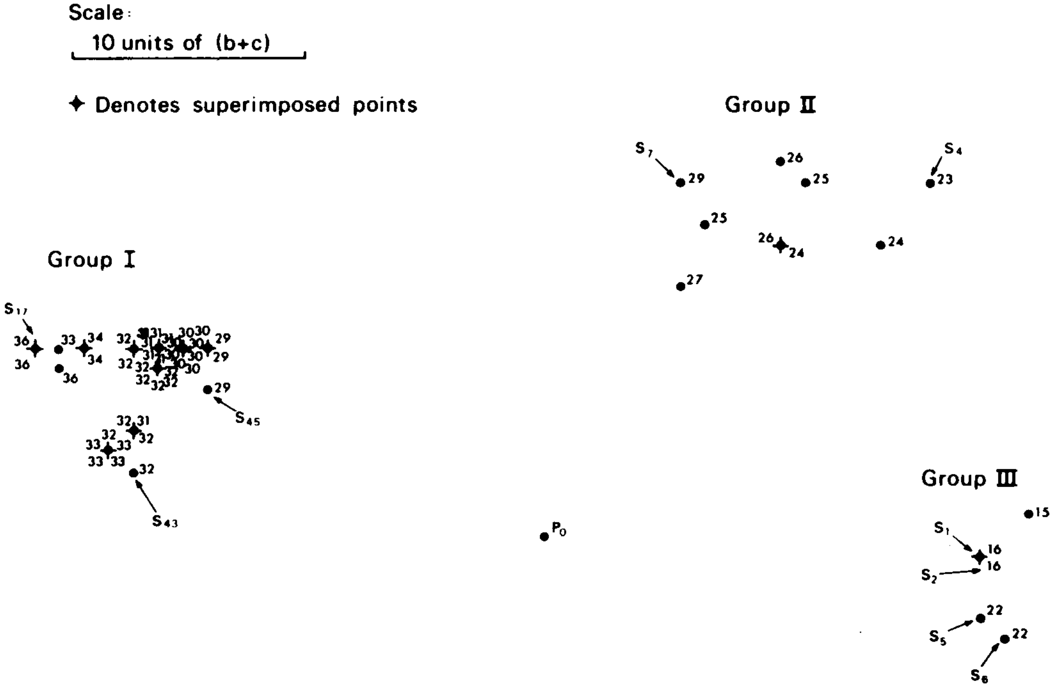

FIG. 1. Ordination of *Phthiracaroidea* using $\Delta(P_i, P_j) = b_{ij} + c_{ij}$. Each point represents one or more of 53 species with associated x-values. The notation S_i denotes species i, if it is referred to in the text.

As an additional check on the adequacy of fit we considered an alternative metric scaling method proposed by Sammon (1969) which determines the δ_{ij}, in a specified number of dimensions, by minimizing the criterion

$$C = \sum(\Delta_{ij} - \delta_{ij})^2.$$

This criterion does not require the Δ_{ij} to be a set of Euclidean distances. We did not find the configuration δ_{ij} that minimized C but used this criterion to compare the principal co-ordinates fits to distances defined by (3) and (6) respectively. The value of C when Δ_{ij} was defined by (6) and the principal co-ordinate distances δ_{ij} fitted in two dimensions was calculated to be $0{\cdot}010 \,\Sigma\, \Delta_{ij}^2$ which is even better than the value of $0{\cdot}020 \,\Sigma\, \Delta_{ij}^2$ found when (3) was used. This reinforces the acceptance of the principal co-ordinate fit to this data, using the non-Euclidean distance (6).

In Fig. 1 there are three distinct groups with a clear trend in x-values across all three groups from high values in group I to low values in group III, suggesting that the main group differences are in the number of characters possessed. The one exception is species number S_7 of group II which has as many characters (29) as those species of group I with fewest characters. This is an example of species with equal x-values belonging to different clusters. Nevertheless, the clustering of x-values closely follows the clustering of species, illustrating the informal clustering criterion discussed in Section 2.

Estimates of the original (a, b, c, d)-values can be found from Fig. 1 and estimates of the values both within and between groups are of interest.

Thus group I is compact, with adjacent members having b and c small and a and d both in the low thirties. The maximum distance within group I is that between species S_{17} and S_{45} for which we have

$$b+c = 8, \quad a+b = 36, \quad a+c = 29$$

giving estimates compared with true ( ) values: $a = 28\frac{1}{2}$ (29), $b = 7\frac{1}{2}$ (7), $c = \frac{1}{2}$ (0), $d = 27\frac{1}{2}$ (28) which may be used to characterize group I. Group II is more dispersed but covers a similar area to group I. The greatest diameter is between species S_4 and S_7 for which

$$b+c = 10, \quad a+b = 23, \quad a+c = 29$$

giving: $a = 21$ (20), $b = 2$ (3), $c = 8$ (9), $d = 33$ (32). In group III we find neighbouring points with rather discrepant x-values. This suggests that the ordination is distorted in this region so that the relationships between species S_1, S_2 and S_5, S_6 are probably not well represented in Fig. 1.

Coming now to investigate the differences between groups we find that the distance between the centres of groups I and II is about $b+c = 26$ character differences (with a minimum difference of 20 and a maximum difference of 37). At the centres of the two groups we have approximately $a+b = 31$ and $a+c = 25$ giving $a = 15, b = 16, c = 10, d = 23$ as average values for the numbers of matches and mis-matches between the two groups. When P_i is in group I and P_j is in group III, the area of triangle $P_0 P_i P_j$ as observed in Fig. 1 is small. This area is $\frac{1}{2}\{(a+b+c)abc\}^{\frac{1}{2}}$. Now $b_{ij}+c_{ij} \doteqdot 35$ (by measurement) and $b_{ij}-c_{ij} \doteqdot 14$ (by differencing x-values) so that $4bc = (b+c)^2-(b-c)^2 \doteqdot 1029$ is large and hence a_{ij} must be small for comparisons between these two groups, showing that they have few common characters. For example with species S_1 (group III), and S_{43} (group I) we have $a_{ij} = 6$, small as expected, contrasting with comparisons of group II with both other groups which have many common characters, as was seen above for groups II and I.

Only pairs of points with equal x-values have equal values of b_{ij} and c_{ij}, which can happen with these data only (except for S_7) when P_i and P_j are in the same group, implying that b_{ij} and c_{ij} are both small. The projection of P_0 onto the ordination plane is also shown in Fig. 1; the residual distance out of the plane of 20 character differences is large. This, coupled with a large average residual for group II relative to the other groups, accounts for the apparent failure of the lengths of the radii from P_0, to reproduce the x-values.

3.2. *Application to the Utilization of Compounds by Yeasts*

Barnett (1976) discusses the utilization of 23 compounds, mainly sugars, by 497 yeasts. He has made available to us more extensive data from which we have selected 29 compounds, 22 of which are included in the original 23. Barnett (1976) gives 2 × 2 tables for many of the different pairs of compounds showing the number of yeasts that use both (a_{ij}), one (b_{ij}, c_{ij}) or neither (d_{ij}) compound, and discusses why tables with low values of either or both of b_{ij} and c_{ij} give information on possible metabolic pathways. We illustrate the use of x-values when combined with ordination based on (3) in analysing these data; other aspects of analysis are discussed by Gower (1980). The plot of the 29 compounds relative to the first two principal axes is given together with their x-values in Fig. 2. This ordination accounts for only 45·1 per cent of the total similarity but does reveal interesting structure.

Examining next the x-values it is seen that Galactitol is utilized by only 44 yeasts and lies to the extreme right of the plot. The compounds ethanol and succinate are utilized by 400 and 390 of the 498 yeasts respectively, and are plotted to the extreme left. Between these two extremes there is an increasing (from right to left) utilization of compounds by the yeasts. This trend is so strong that contour bands of x-values can be superimposed on the plot. Because the ordination fit is poor it would be wise to treat measured distances on the plot as very rough approximations to $(b+c)^{\frac{1}{2}}$. Nevertheless, the evidence of real structure is ample enough to make it worthwhile, for exploratory purposes, to investigate Fig. 2 a little more closely. Somewhat surprisingly the x-values do not indicate any dramatic distortion in the ordination. The worst region seems to be that between the 200 and 300 x-value contours and especially the sudden change between melezitose ($x = 211$) and D-galactose ($x = 262$). Elsewhere citrate ($x = 315$) seems to be plotted further to the left than is consistent with good distance preservation.

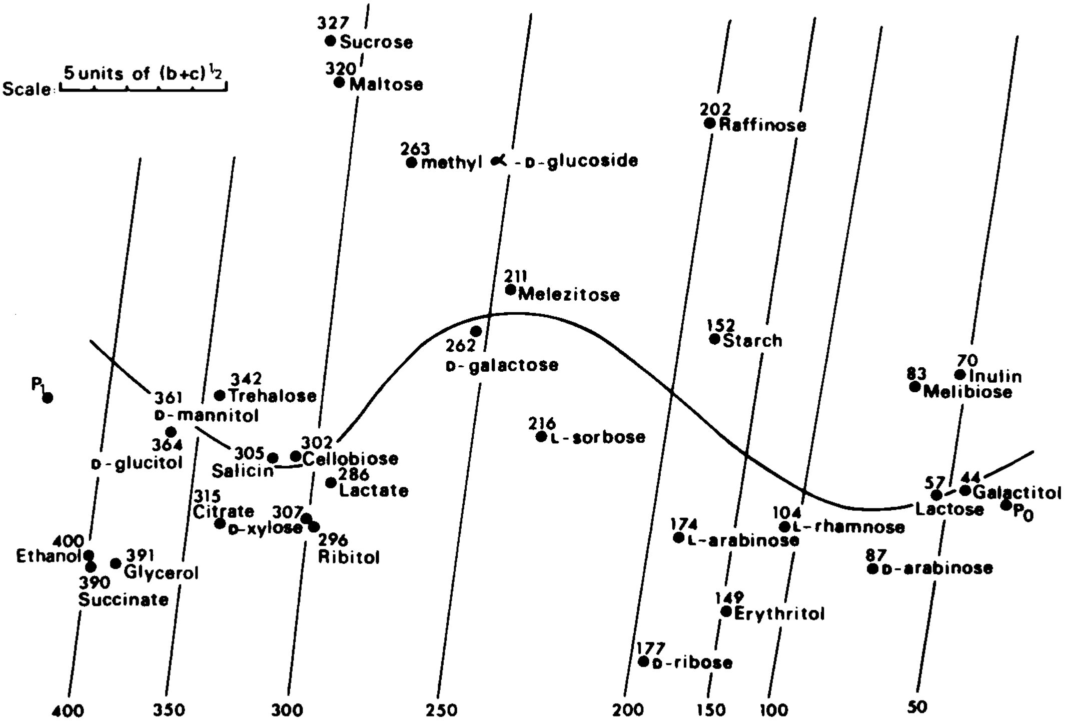

FIG. 2. Ordination of sugar-related compounds using $\Delta(P_i, P_j) = (b_{ij} + c_{ij})^{\frac{1}{2}}$. Each point represents one of 29 compounds with associated x-value·. The parallel lines are contours of constant x-value.

Neighbouring points indicate tables with small b and c. Examples consistent with Barnett's tables are salicin with cellobiose, sucrose with maltose and lactose with melibiose. Points close to P_0 will have small values of a and b as do lactose and D-arabinose. Points close to P_1 would have small d and b but there are no such substances in Fig. 2. When c is small and b large (or vice versa) we require $S_{ij}^2 = b + c$ and $|b - c|$ to be similar in size. The contours indicate the values of $b - c$ which are given exactly for any two substances by subtracting their x-values. The scale for $(b+c)^{\frac{1}{2}}$ allows $b + c$ to be calculated, but only approximately in view of the poor ordination. For given $b - c$ it is easy to evaluate $(b+c)^{\frac{1}{2}}$ in terms of the graph-scale and so pick up likely pairs of candidates for which b or $c = 0$. For example, for raffinose with sucrose or maltose, $b - c = 120$ and $b + c \div 125$. However, for maltose and L-arabinose, $b - c = 146$, which is similar to the two previous values, but $b + c \div 290$ (true value 215) corresponding to the much longer distance, and correctly ruling out the possibility of b or c being near zero.

4. CONCLUSION

1. Plotting x-values is a simple but effective method for checking the accuracy of ordinations of binary data and offers a useful screening test before cluster analysis.

2. x-values permit the recovery of approximate values of the entries of the 2×2 table from which the ordination was derived. In particular tables with special structure such as zero values can be picked out visually.
3. As illustrated by the two examples, x-values can be a useful interpretive device.

Acknowledgements

We thank Dr J. G. Sheals and Dr J. A. Barnett for making available to us their unpublished data.

References

Barnett, J. A. (1976). The utilisation of sugars by yeasts. *Advances in Carbohydrate Chemistry and Biochemistry*, **32**, 125–234.

Blackith, R. E. and Reyment, R. A. (1971). *Multivariate Morphometrics*, pp. 1–412. London and New York: Academic Press.

Constantine, A. G. and Gower, J. C. (1978). Graphical representation of asymmetric matrices. *Appl. Statist.*, **27**, 297–304.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325–338.

—— (1968). Adding a point to vector diagrams in multivariate analysis. *Biometrika*, **55**, 582–583.

—— (1971). A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 852–871.

—— (1977). The analysis of asymmetry and orthogonality. In *Recent Developments in Statistics* (J. Barra *et al.*, eds). Amsterdam: North-Holland.

—— (1980). Problems in interpreting asymmetrical chemical relationships. In *Chemosystematics: Principles and Practice* (F. Bisby, ed.). London and New York: Academic Press.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Sage University papers on Quantitative Applications in the Social Sciences. No. 07–011. Beverly Hills and London: Sage Publications.

Sammon, J. W. (1969). A non-linear mapping for data-structure analysis. *IEEE Trans. Computers*, C–**18**, 401–409.

Sheals, J. G. (1969). Computers in acarine taxonomy. *Acarologia*, **XI**, 376–396.

Shepard, R. N. and Carroll, J. D. (1966). Parametric representation of nonlinear data structures. In *Multivariate Analysis* (P. R. Krishnaiah, ed.), pp. 561–592. New York: Academic Press.

Sibson, R., Bowyer, A. and Osmond, C. (1980). Studies in the robustness of multidimensional scaling: Euclidean models and simulation studies. (Submitted for publication.)

Sneath, P. H. and Sokal, R. R. (1973). *Numerical Taxonomy*. San Francisco: W. H. Freeman & Co.

Torgerson, W. S. (1958). *Theory and Methods of Scaling*. New York: Wiley.