Vegetatio 54, 27-35 (1983).
© Dr W. Junk Publishers, The Hague. Printed in The Netherlands.

On choosing a resemblance measure for non-linear predictive ordination

P. H. Fewster & L. Orlóci*
Department of Plant Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B7

Keywords: Horseshoe, Linear, Multi-dimensional scaling, Non-linear, Ordination

Abstract

The development of non-linear ordination techniques has stemmed in part from work suggesting that species behave non-linearly to changing environmental factors or gradients. Developments in this area can be seen in two related phases: new algorithms, and the incorporation of new resemblance measures. Emphasis in this paper is placed on resemblance measures incorporated into a method of multi-dimensional scaling. The results show that a resemblance measure which reflects the non-linearities of the data can produce significant improvement in ordination, if the standardizations have not been too 'severe'.

Introduction

The investigation and ordering of vegetation units with respect to known or presumed underlying environmental gradients has long been a major objective of ecological studies. More recent ecological ordinations have evolved from the central idea that the manner in which species respond to environmental influences must be considered. This implies that optimality of the solutions is tied to the method's success in incorporating suitable response models. R. H. Whittaker's contributions were most influential in this area in that they established a theoretical framework incorporating the notions of gradient, response, and utility. It is largely a consequence of his influence that shortcomings in linear ordinations were revealed. This in turn lead to the development of methods which assume non-linear species responses.

The literature (reviewed by Orlóci, 1978) illustrates that early efforts were largely concerned with multi-dimensional configurations where individuals (vegetation plots) occupied positions and the species served as dimensions. An ordination, by contrast, is visualized as a configuration of individuals in a space where major physical (environmental) factors serve as the axes. Hence the problem involves finding the best way to transfer or map individuals in species space into factor space with minimum distortion, and to identify these factors with greatest certainty. In other words, non-linearities need to be unfolded as much as possible so as to obtain a linear ordering.

The complexity of vegetation data can result from a number of factors, such as random variation (noise) and indeterminacy in measurement. More important, however, is the type of species response. This has been demonstrated in both field data and simulation experiments (see van Groenewoud, 1965; Noy-Meir & Austin, 1970; Gauch & Whittaker, 1972). If the response is linear, complexity is not great and efficient ordination algorithms are readily available. In dealing with non-linear data, however, it becomes important that the technique used incorporates devices to handle this non-linearity.

* One of the authors (L. Orlóci) was a recipient of an N.S.E.R.C. grant during the tenure of this project. The authors wish to thank C. Brambilla and G. Salzano for the use of their computer program. A copy of a modified version used here may be obtained from the first author at no charge.

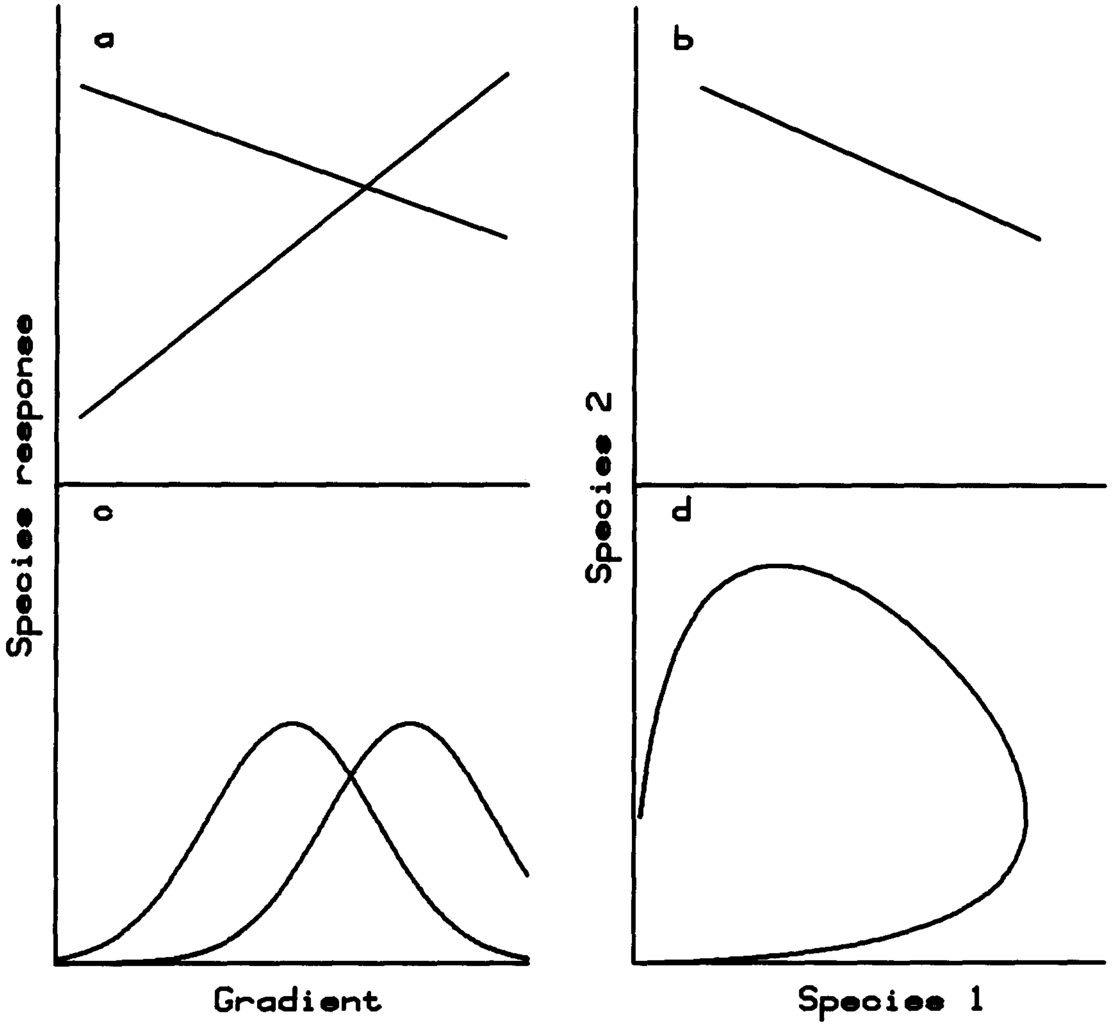

Fig. 1. Response trajectories for two species along a gradient with species response (a) linear and (c) Gaussian with their respective joint scatters (b) and (d) in species space.

Different cases are illustrated in Figure 1. Consider the simplest case where species respond linearly to an environmental gradient (such as changes in elevation up a mountainside), excluding noise. If individuals (vegetation plots) are placed at regular intervals along the gradient, and if two recorded species respond linearly to the gradient, a situation as in Figure 1a would be obtained. If the same information is graphed in species space, where the two species serve as coordinate axes, a straight line is obtained (Fig. 1b). If an ordination seeks to obtain an ordering of individuals, which is meaningful with respect to some environmental gradient (elevation in this case), a transformation is required. In the linear case this transformation is not complex since the ordering of individuals along the line of joint response in species space (Fig. 1b) is the same as the ordering along the abscissa of Figure 1a. A principal components analysis (PCA) using a Euclidean distance measure will return this line on the first axis. If the example is extended to more than two species, a straight line will still be obtained in species space. Clearly, since a linear species response produces a linear configuration in species space, a linear resemblance measure would be the most meaningful. In fact, since this is a Euclidean space, the Euclidean distance is an appropriate resemblance measure.

If a non-linear species response is assumed, a configuration such as shown in Figure 1c (Gaussian curves) might be obtained. The same information in two-dimensional species space is shown in Figure 1d. Again, the same basic shape (in multi-dimensional space) will be obtained no matter how many species there are. In this case the configuration as represented in species space has a horseshoe shape. Hence the transformation, which takes individual points on this horseshoe and maps them onto a straight line, is necessarily complex. This paper focuses on the development of resemblance measures which can be used in the transformation from a horseshoe to a straight line.

Method

The algorithm used in the analysis accomplishes multi-dimensional scaling (MDS). Lucid descriptions of the method are given by Fasham (1977) and Brambilla & Salzano (1981), after the original outline in Kruskal (1964a, b). Attention is drawn here to a few salient features of the algorithm before remarks concerning the choice of a resemblance measure are made.

MDS works iteratively toward a final solution by comparing distances obtained from the raw data with those from a 'proposed' solution. The choice of a distance measure for the 'proposed' solution therefore determines whether MDS is a linear or non-linear method. In this respect MDS differs from other methods which attempt to handle non-linearity in the data. These include methods which fit curved axes (e.g. Phillips, 1978) or specified response curves (e.g. Johnson, 1973; Gauch, Chase & Whittaker, 1974; Johnson & Goodall, 1979), and those which use regression analysis and scaling to reduce the curvature of an ordination configuration (e.g. Hill & Gauch, 1980).

The version of MDS used here begins either with a random initial point configuration, or one specified using the maximum variance criterion. By this criterion, the p most variable species are used to define an initial configuration, where p is the number of dimensions (D) for which a solution is sought. Kendall (1971) has suggested that the $(p + 1)$D solution of a pD data set is an appropriate strategy, since this reduces the chance of selecting a local minimum as a solution. Since the data sets tested in this paper all have a single underlying gradient (1D), two-dimensional solutions were sought in all cases.

The choice of an appropriate resemblance measure for data with non-linear species responses is difficult. Numerous possible resemblance measures are conceivable, each specific to a given species response type (cf. Austin, 1979). In any case, the familiar metric resemblance measures are non-optimal when non-linear species responses occur. As a simple example, consider PCA of the data in Figure 1d using Euclidean distance. The result would be a horseshoe, since the algorithm involves a simple geometric rotation in species space. Similarly, when a data set with Gaussian species responses is subjected to MDS analysis incorporating Euclidean distance, the result is again a horseshoe-shaped curve (Fig. 2). A straight line, representing the underlying gradient to which species are responding, is more desirable. In both cases a linear measure is inappropriate since it does not reflect the response structure of the data. For a curved configuration in species space, the problem amounts to developing an appropriate measure of the distance between two points A and B. A linear resemblance measure will give the distance of line AB, whereas a suitable non-linear measure will give a close approximation to the distance of arc AB.

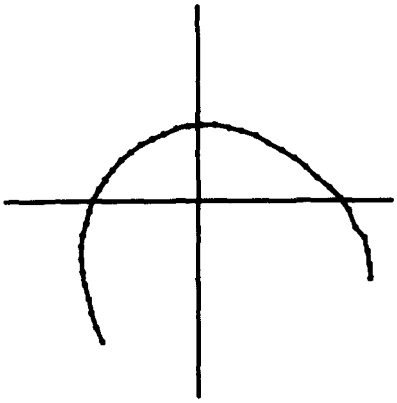

Fig. 2. Results from MDS in ordination space using a Gaussian data set with the linear distance option. Axis 1 is the abscissa and axis 2 the ordinate. Horseshoe-shaped, stress = 0.0079.

Orlóci (1978) and Hill & Gauch (1980) have noted that quantitative methods have specific uses and that the user must be careful in applying them. In addition, many authors (e.g. Austin, 1976, 1979, 1980; Werger et al., 1983) have found a variety of species responses in nature such as Gaussian, bimodal, skewed, plateau, and so on. The subject is still very much in the exploratory stage (Feoli & Feoli Chiapella, 1980). However, if a few simplifying assumptions are made, some progress may be made in developing a resemblance measure in which provision is made for the actual curvatures found in nature. An example of this type of approach has been described by Ihm & van Groenewoud (1975) who, assuming Gaussian species responses, applied appropriate transformations to the product moments prior to an eigenanalysis.

In developing appropriate distance measures, the first assumption we make (which will be relaxed later) is that of a single underlying gradient. It is further assumed that all species responses are of the same form, although the actual response type remains open to choice. The objective is that of predicting the ordering of individuals along an unknown gradient based on species scores. The measured responses Y for a set of p species are assumed to be a function of the levels of an environmental influence to be approximated by a set of ordination scores X (Fig. 3a). Hence $Y = f(X|m)$, where m represents a set of parameters of the response graphs. For any point X on the gradient (abscissa) in Figure 3b, a linear distance (Δ) to a second point $X + \Delta$ is defined. The figure shows that this distance is related to the species response distance $f(X) - f(X + \Delta)$. Although this new distance depends on the position of X along the gradient, a distance which is unique to the type of curve which $f(X|m)$ expresses can be derived. Furthermore, the restriction regarding a single gradient can be relaxed and a similar construction on each of t gradients can be produced, assuming the same type of species response.

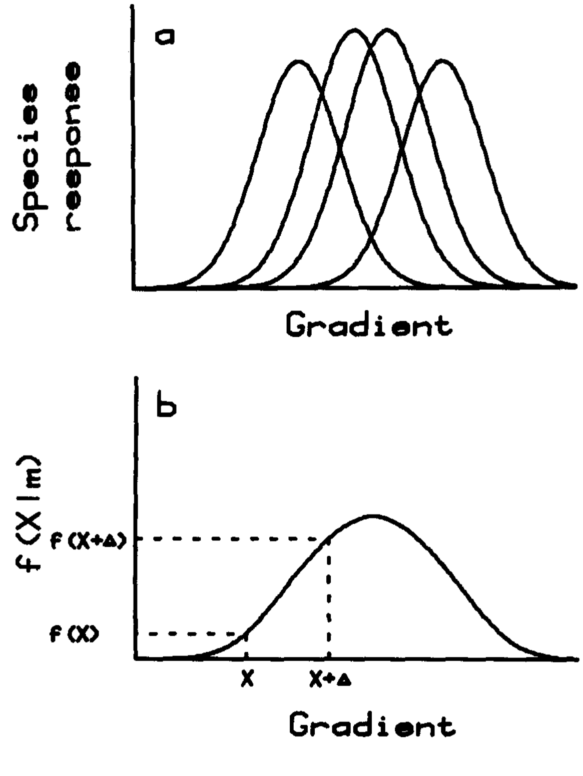

Fig. 3. Response trajectories for several species along a gradient (a) and a standard average trajectory (b).

Now, let $\Delta = |X_{ij} - X_{ik}|$ be the ith gradient distance between individuals j and k. Then the unique distance between individuals j and k on gradient i (the compositional distance of Orlóci, 1978, 1980) is

$$d^2(j,k|i) = \int_{-\infty}^{\infty} \{f(X|m) - f(X + \Delta|m)\}^2 \, \mathrm{d}X \qquad (1)$$

The power 2 was chosen because the integration is possible and because it leads to an interpretable formula. The composite compositional distance,

$$d^2(j,k) = \Sigma\, d^2(j,k|i), \quad i = 1,\ldots,t \qquad (2)$$

gives the distance between individuals j and k. What has been accomplished is a definition of the gradient distance in relation to a distance based on an assumed non-linear species response. Since this distance uses information about the actual species responses, it can be expected to have potential utility when linear species responses cannot be assumed.

Next, a few specific types of species responses are selected, and equation (1) solved to give actual distance (resemblance) measures.

The symmetric Gaussian curve was chosen since responses of this type (bell-shaped) have been reported many times in the literature and are thought to be common (Whittaker, 1956, 1967; van Groenewoud, 1965). The second choice was the skewed Gaussian, since this response type has also been noted (Austin, 1979). Finally, a parabolic curve, which has the basic bell shape but lacks the tails, was chosen. In nature such a response might be expected since a species might be out-competed or otherwise selected against at the extremes of its potential range (Forsythe & Loucks, 1972).

Derivation

The integral in equation (1) is now solved. It is noted that standardization of various parameters preceded integration in order to obtain a distance measure independent of these parameters. Note, however, that a substantial loss of general utility may be a consequence of standardizing too many parameters.

For the Gaussian function,

$$y = Be^{-(X-a)^2/2s} \tag{3}$$

where B & s are related to height and width respectively, only standardization to unit height and width ($m = (a,1)$) is needed since the constant a (the level of influence at which response is maximal) drops out following integration. After transformation, the function becomes,

$$y = e^{-(X-a)^2/2} \tag{4}$$

Thus

$$\begin{aligned} d^2(j,k|i) &= \int_{-\infty}^{\infty} \{e^{-(X-a_i)^2/2} - e^{-(X+\Delta-a_i)^2/2}\}^2 \, dX \qquad (5) \\ &\propto 2(1 - s_{ijk}) \end{aligned}$$

where $s_{ijk} = e^{-\Delta^2/4}$. A similar form was first reported by Gauch (1973).

For the skewed Gaussian function,

$$y = aX^b e^{-cX} \tag{6}$$

the standardization involves setting the parameters a, b and c to unity. Since the mode or abscissa of vertex $X_m = b/c$, an 'extra' or 'more severe' standardization is used compared to the Gaussian. The distance derived following integration is

$$d^2(j,k|i) \propto 1 + (2(\Delta^2 + \Delta - 1)e^{-2\Delta} - 2\Delta e^{-\Delta} \tag{7}$$

For the parabolic function

$$y = -aX^2 + bX + c \tag{8}$$

the standardization involves setting $a = b = 1$ and $c = 0$. This is 'more severe' than that of the skewed Gaussian, since $X_m = -b/2a$. The derived distance is

$$d^2(j,k|i) \propto 3\Delta^4 + \Delta^2 \tag{9}$$

Testing

The next phase is testing the technique (resemblance measure plus method). MDS can incorporate any one of the distance formulae as options for the ordination configuration. The original configuration distances vary, conforming to the ordination configuration distances. In this early stage of development it has been necessary to make some restrictions before generating the test data sets. They are: 1) a single gradient, 2) a few species (10) with the same type of response, and 3) random parameters for the response curves, within certain ranges. The gradient is conceived as being very broad, ranging between two extremes. A range of individual positions was defined between these two extremes where species optima would have an equal (random) chance of occurring. Ranges of constant probability were also chosen for parameters defining the height and width of the curves. The construct simulated the random appearance and disappearance of species along the gradient and implied that individuals (vegetation plots) had fewer species the further they were located from the middle of the gradient.

One data set was generated with Gaussian species responses whereas six sets each were generated for skewed Gaussian and parabolic responses. This was done in order to compare the results for consistency due to the 'extra' standardization. Beta diversity (Whittaker, 1972) ranged from 0.06 hc for the Gaussian set to 3.5 hc for the parabolic sets.

Results

The Gaussian data set produced 2D ordinations as illustrated in Figure 4. The first (Fig. 4a) resulted from the maximum variance option. Repetitions with the random option resulted in the same basic open shape as shown in Figure 4b. This was the only combination of distance measure and option which produced an ordination which is distinctly not a horseshoe. Results from the skewed Gaussian and

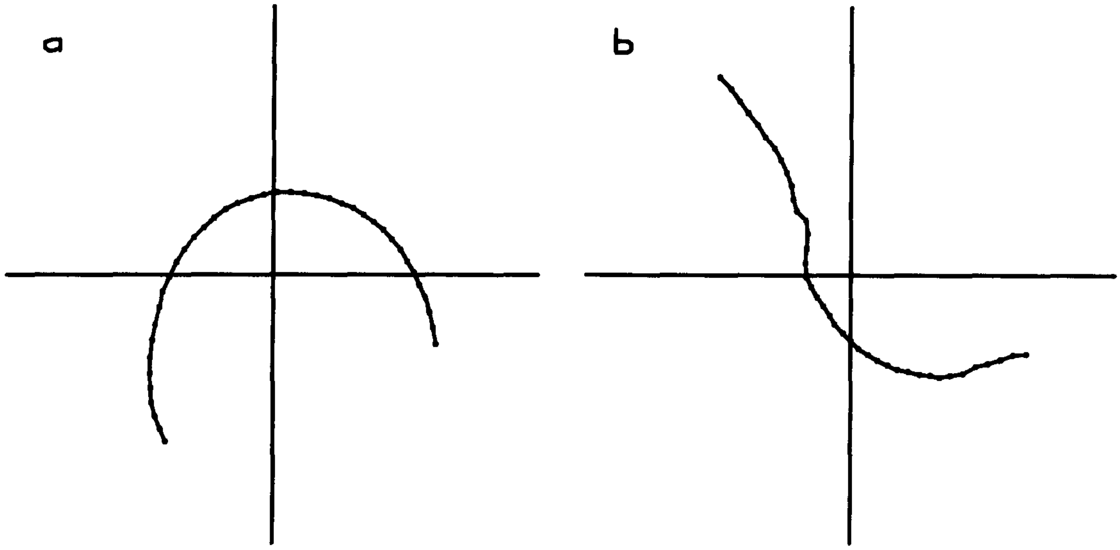

Fig. 4. Results from MDS using a Gaussian data set: (a) horseshoe-shaped, stress = 0.0089 and (b) open-shaped, stress = 0.013.

parabolic data sets, using the maximum variance option, are given respectively in Figures 5 and 6. The curves are asymmetric horseshoe-shaped in both distance options.

Discussion

In general, two aspects of ordination efficiency need to be considered. The first is the possibility that a solution may represent a scrambling of the true ordering, even though the algorithm and resemblance measure used are theoretically appropriate. Problems of this sort may arise, for example, from random variation (noise) in the data. In addition, a complex ordination algorithm like MDS has certain idiosyncrasies (particularly the problem of local minima) which may result in a misordering. The second aspect is the arch or horseshoe effect. Two examples are given (one in PCA and another

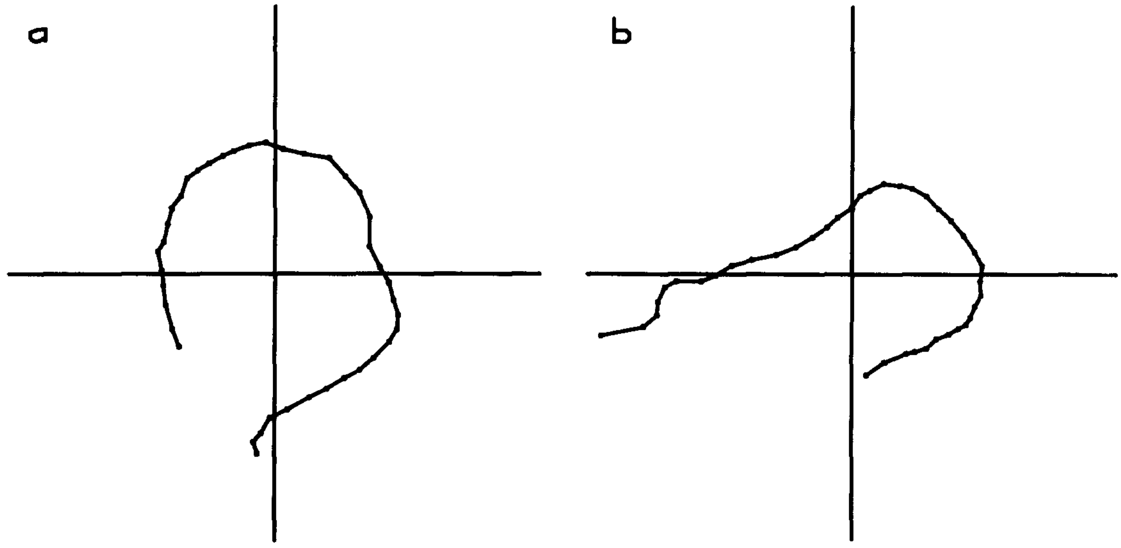

Fig. 5. Results from MDS using skewed Gaussian data sets: (a) involuted, asymmetric horseshoe-shaped, stress = 0.117 and (b) involuted, asymmetric horseshoe-shaped, stress = 0.118.

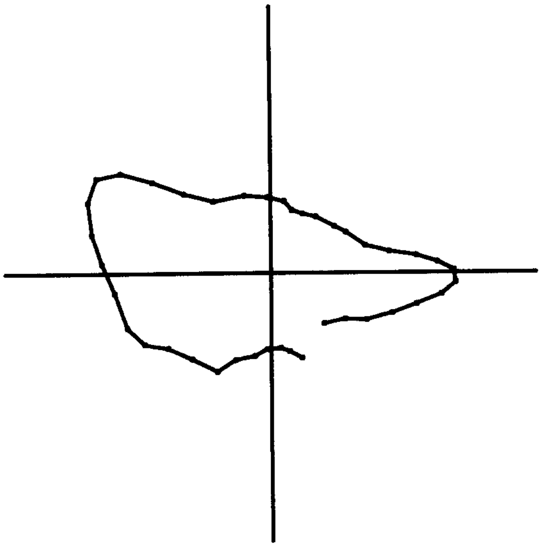

Fig. 6. Result from MDS using a parabolic data set. Asymmetric horseshoe-shaped with much involution, stress = 0.077.

in MDS, Figs. 1c, d and 2) showing that a linear resemblance measure used in ordinating non-linear data produces an involuted, horseshoe-shaped ordination configuration. Kendall (1971) points out that the involution of the horseshoe implies that, without knowing the number of gradients *a priori*, an ordering of individuals may be scrambled. This can be visualized by taking a 2D involuted horseshoe and projecting it onto a single dimension (line). The points at the involuted ends will be mixed in with the middle ones so that the ordering produced is very different from the true one. While some have felt that a linear ordering is necessary and have developed approaches to straightening the ordination configuration (Kendall, 1971; Hill & Gauch, 1980), others (Feoli & Feoli Chiapella, 1980) suggest that the horseshoe effect is revealing rather than detrimental to interpretation of the results.

Our approach in deriving an improved ordination has been to focus on the use of resemblance measures and to illustrate their importance in effectively mapping the data structure from species space into ordination space. Derived measures are combined with the MDS method to produce a non-linear ordination technique. The results using two distance measures derived for the skewed Gaussian and parabolic response are considered first. It has been noted that 'extra' standardization was needed in developing these measures in comparison with the Gaussian. We attribute the relative lack of success of these measures in handling the arc *AB* to these 'extra' standardizations in their derivation.

With respect to the Gaussian measure, MDS performed differently depending on which initialization option was used. Since the algorithm is one of minimization, different solutions may be obtained from different initial configurations: various local minima are conceivable, each returning a different solution. In general the maximum variance option produced less desirable results than the random option. This is because there is no bias in the random option as to the shape of the initial configuration. As such the final result is more a consequence of the distance measure and the method. Other approaches for initializing MDS include using the results of a linear ordination such as PCA. This also tends to give a 2D configuration like Figure 1d, since the initial PCA configuration is a horseshoe. From a developmental and exploratory perspective, the Gaussian measure derived here in conjunction with MDS and the random option has been shown to be workable within the given set of assumptions.

The Gaussian measure was derived by integrating a single response function. Why were not all response curves treated individually so as to obtain a more effective measure? Each curve could be integrated, providing the widths of response curves were first standardized. The results for each species would be the same (cf. equation (5)). Equation (5) could then be used as the collective function relating species to gradient position. Hence the only difference so far is the approach. Another question relates to the standardization of the widths (σ) of the response graphs. If the width for a single species (h) can be made more realistic by incorporating σ^2, integration will give the result $d^2(j,k|i,f(\sigma_h^2))$ which is related to equation (5) but differs by some function of the width. The compositional distance then becomes,

$$d^2(j,k|i) = d^2(j,k|i,f(\sigma_h^2)), \quad h = 1,\ldots,p$$

where p is the number of species. The problem is that since σ is a measure of X, until the gradient is determined, the σ_h^2 remain unknown. This suggests a possible feedback algorithm, in which the distance measure could be made more effective at each step.

Finally, since the perspective of ordination, and indeed that of data analysis, is often exploratory, concern is often directed towards obtaining insights. Attention in this paper has been focused on using resemblance measures which are in some way based on the same type of non-linearity as in the data. However, it is not always possible *a priori* to know much about the data structure. In reference to Figures 4b, 5 & 6, distinctive curves or 'signatures' are produced in the analysis, depending on the measure used, which reflect the underlying data structure. This would also be expected to happen for data with noise. In such a case, a cloud may result showing an overall trend much like one of the known signatures. Hence, it is possible, by trial, to obtain insight into the type of non-linear species response.

The MDS approach to ordination used here is one of many which may give an improvement in the handling of non-linearities in the data. The others include curve fitting, scaling, and *a posteriori* detrending. There are essential differences among these approaches in the conceptualization of the objective and the definition of optimality. With the methods tested here, a solution is regarded as being optimal if the ordination succeeds in unfolding a non-linear configuration. (In this respect, only the Gaussian response measure has been shown to be of utility, since the parabolic and skewed Gaussian measures both returned the horseshoe.) This implies that the type of species response assumed in the derivation of the distance measure is most likely correct. If a horseshoe type ordination configuration is obtained, the original assumption is deemed inappropriate. The actual shape of the ordination configuration may, however, suggest what type of response is depicted by the data. In other words, even if the solution of MDS is not optimal, the ordination still conveys information about properties of the data which relate to non-linearity.

Similar advantages can be seen with curve fitting, such as in the polynomial ordination of Phillips (1978). Here, however, the curvature anticipated by the model before fitting is not concerned directly with the type of response exhibited by the species. Thus as in MDS, non-linear trends are not removed before the user has a chance to detect their presence. By contrast, detrending (Hill & Gauch, 1980) removes trends from the data that the user of non-linear ordinations hopes to detect. This may be completely justifiable and may help greatly in scrutinizing the ordination results, but it cannot be condoned as a general strategy.

References

Austin, M. P., 1976. Performance of four ordination techniques assuming three different non-linear species response models. Vegetatio 33: 43–49.

Austin, M. P., 1979. Current approaches to the non-linearity problem in vegetation analysis. In: Patil, G. P. & Rosenzweig, M. L. (eds.), Contemporary Quantitative Ecology and Related Ecometrics, pp. 197–210. ICPH, Fairland, Maryland.

Austin, M. P., 1980. Searching for a model for use in vegetation analysis. Vegetatio 42: 11–21.

Brambilla, C. & Salzano, G, 1981. A non-metric multi-dimensional scaling method for non-linear dimension reduction. Istituto per le Applicazioni del Calcolo 'Mauro Picone'. Quaderni, Serie III - N. 121. 35 pp.

Fasham, M. J. R., 1977. A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines, and coenoplanes. Ecology 58: 551-561.

Feoli, D. & Feoli Chiapella, L., 1980. Evaluation of ordination methods through simulated coenoclines: some comments. Vegetatio 42: 35–41.

Forsythe, W. L. & Loucks, O. L., 1972. A transformation for species response to habitat factors. Ecology 53: 1112–1119.

Gauch, H. G., 1973. The relationship between sample similarity and ecological distance. Ecology 54: 618–622.

Gauch, H. G., Chase, G. B. & Whittaker, R. H., 1974. Ordination of vegetation samples by Gaussian species distributions. Ecology 55: 1382-1390.

Gauch, H. G. & Whittaker, R. H., 1972. Comparison of ordination techniques. Ecology 53: 868–875.

Groenewoud, H. van, 1965. Ordination and classification of Swiss and Canadian coniferous forests by various biometric and other methods. Ber. Geobot. Inst. ETH Stiftg. Rübel, Zürich 36: 28–102.

Hill, M. O. & Gauch, H. G., 1980. Detrended correspondence analysis: an improved ordination technique. Vegetatio 42: 47–58.

Ihm, P. & van Groenewoud, H., 1975. A multivariate ordering of vegetation data based on Gaussian type gradient response curves. J. Ecol. 63: 767–777.

Johnson, R., 1973. A study of some multivariate methods for the analysis of botanical data. Ph.D. dissertation, Utah State Univ., Logan, Utah.

Johnson, R. W. & Goodall, D. W., 1980. A maximum likelihood approach to non-linear ordination. Vegetatio 41: 133–142.

Kendall, D. G., 1971. Seriation from abundance matrices. In: Hodson, F. R., Kendall, D. G. & Tautu, P. (eds.), Mathematics in the Archaeological and Historical Sciences, pp. 215-252. Edinburgh Univ. Press, Edingburgh.

Kruskal, J. B., 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29: 1-27.

Kruskal, J. B., 1964b. Nonmetric multidimensional scaling: a numerical method. Psychometrika 29: 115-129.

Noy-Meir, I. & Austin, M. P., 1970. Principal component ordination and simulated vegetational data. Ecology 51: 551-552.

Orlóci, L., 1978. Multivariate Analysis in Vegetation Research. 2nd ed. Junk, The Hague. 451 pp.

Orlóci, L., 1980. An algorithm for predictive ordination. Vegetatio 42: 23-25.

Phillips, D. L., 1978. Polynomial ordination: field and computer simulation testing of a new method. Vegetatio 37: 129-140.

Werger, M. J. A., Louppen, J. M. W. & Eppink, J. H. M., 1983. Species performances and vegetation boundaries along an environmental gradient. Vegetatio (in press).

Whittaker, R. H., 1956. Vegetation of the Great Smoky Mountains. Ecol. Monog. 26: 1-80.

Whittaker, R. H., 1967. Gradient analysis of vegetation. Biol. Rev. 42: 207-264.

Whittaker, R. H., 1972. Evolution and measurement of species diversity. Taxon 21: 213-251.

Accepted 30. 3. 1983.