Ecology, 83(12), 2002, pp. 3331–3343

RESEMBLANCE COEFFICIENTS AND THE HORSESHOE EFFECT IN PRINCIPAL COORDINATES ANALYSIS

J. Podani[1] and I. Miklós

Department of Plant Taxonomy and Ecology, Eötvös University, Pázmány P. s. 1/c H-1117 Budapest, Hungary

Abstract. Although principal coordinates analysis is one of the most widely used ordination methods in ecology, no study had been undertaken as yet on the combined effect of gradient type and resemblance coefficient on the results. We examine the performance of principal coordinates analysis with different choices of the resemblance function and different types of a single underlying gradient. Whereas unimodal species response to long gradients always leads to horseshoe (or arch)-shaped configurations in the first two dimensions, the converse is not true; curvilinear arrangements cannot generally be explained by the Gaussian model. Several resemblance coefficients widely used in ecology produce paradoxical arches from perfectly linear data. Species richness changes alone may also lead to a horseshoe for even more distance functions, with the noted exception of Manhattan metric. The appearance of arches is a mathematical necessity in these cases; true artifacts are introduced only if distances are treated inappropriately before eigenanalysis. Examples illustrate that similar configurations (curves and even circles) may arise from very different data structures; therefore the shape of the point scatter is insufficient by itself to identify background ecological phenomena. The horseshoe effect may be diminished and eigenvalue extraction may be made more efficient if input measures are raised to high powers; but this operation is recommended only in combination with standard analyses, as part of a comparative approach. We derive a new distance function, implying standardization by species totals, from the chi-square distance. We found that this function improves gradient recovery when there is unimodal species response and some species have their optima outside the range of study.

Key words: arch effect; Bray-Curtis coefficient; chi-square distance; correspondence analysis; detrending; dissimilarity; gradients; linear scaling; Manhattan metric; ordination; principal coordinates analysis; resemblance coefficients and the horseshoe effect.

Introduction

Curved arrangements of points (the so-called ''horseshoe effect,'' ''arch effect,'' or ''Guttman effect'') have been commonly observed in ecological ordinations obtained by linear scaling methods (principal components analysis [PCA], correspondence analysis [CoA], principal coordinates analysis [PCoA]) and even by nonmetric multidimensional scaling (NMDS) from presumably nonlinear data structures (Goodall 1954, Kendall 1971, Gauch 1982, Pielou 1984, Kenkel and Orlóci 1986, Legendre and Legendre 1998, Podani 2000). These ''effects,'' also referred to as ''distortions,'' ''artifacts,'' or ''faults,'' have been generally considered as a disadvantage in numerical ecology. The major concern is that assumed one- or few-dimensional background gradients are not reproduced by linearly arranged points in the abstract, usually two-dimensional ordination spaces. A vast amount of work has been devoted to modify ordination procedures to resolve these problems (detrending, Gaussian ordination, unfolding, distance adjustments, principal curves, data standardizations, etc.) and no review of this subject is attempted here (see Williamson 1978, Bradfield and Kenkel 1987, Dale 1994, 2000, De'ath 1999*a*, *b*, and Legendre and Gallagher 2001). What we wish to emphasize now is that despite the familiarity of ecologists with the phenomenon, the understanding of its mathematical background is still unsatisfactory, which is a potential source of confusion in interpreting ordination results and in evaluating the relative merits of scaling procedures. In particular, even though PCoA has been one of the most widely used ordination methods, its performance with different resemblance coefficients as applied to artificial data remains uninvestigated, most attention has been paid to the detrimental effects of negative eigenvalues obtained from nonmetric input dissimilarities (Gower and Legendre 1986, Digby and Kempton 1987, Legendre and Legendre 1998, Legendre and Anderson 1999). Furthermore, it is striking that arches and horseshoes derived from other metric ordination methods (CoA, PCA, and their constrained forms, ter Braak 1986, ter Braak and Prentice 1988) are treated in much more detail in the literature than PCoA arches. Yet, there is a prevailing view that curved arrangements are always of the same mathematical nature in ecological ordinations. A recent example is Økland (1999) who deals with the relative amount of variation explained by metric ordination axes. Whereas Økland (1999) makes many good points questioning the utility of eigenvalue/total inertia ratios as an interpretative aid, he explicitly states that second and

Manuscript received 22 June 2001; revised 18 April 2002; accepted 25 April 2002.

[1] E-mail: podani@ludens.elte.hu

many subsequent ordination axes are ''polynomial distortions'' of the first axis, regardless of the ordination procedure used. This is an obvious simplification, however, even though a PCoA horseshoe may appear at first sight identical in shape to a CoA arch, especially for short gradients. In fact, Gauch (1982), Pielou (1984), ter Braak (1987) and Heiser (1987) already warned that there are at least two kinds of arched arrangements (Heiser mentioned two more cases for nonmetric ordinations). For the unimodal species response type, the basic observation is that the CoA parabola for axes 1 and 2 allows a unidirectional ordering of objects to be made (Greenacre 1984, Legendre and Legendre 1998; but see exception in Wartenberg et al. [1987: Fig. 3] for a short gradient). As demonstrated in the present paper as well, PCoA- (and PCA-) created curves often, but not always, exhibit involution (bending inwards) at both ends of the gradient (Austin and Noy-Meir [1971] were the first to recognize this), making projection of points to the first axis meaningless. Ter Braak (1987) proposed to distinguish between these two cases by restricting the term ''horseshoe'' to PCA/PCoA and the term ''arch'' to CoA ordinations (the names being valid obviously for the first two dimensions).

The present paper examines when and why arched arrangements of points emerge from PCoA of actual field data and artificial data sets with known properties. We raise the issue whether the power transformation of distances improves gradient recovery of ordinations and we emphasize the importance of the comparative approach to ecological ordination.

Material and Methods

Illustrative data

In order to visualize the effect of the resemblance measure and other operations on the resulting ordination, we shall restrict ourselves to one-dimensional underlying gradients. Considering two or more meaningful background dimensions would extremely complicate interpretations and explanations, and we believe that the one-dimensional case is sufficient to elaborate the topic. To evaluate efficiently the mathematical properties of curved arrangements, we must rely on artificial data representing ''idealized'' gradients even though these perfect forms may appear unlikely and unrealistic ecologically. Three kinds of one-dimensional background gradients are considered:

1) *Unimodal response* of species to a long gradient (Fig. 1). Each ''species'' appears in three objects (sites, sampling units, etc.), with values corresponding to a simple unimodal, roughly ''Gaussian'' response. One species disappears; another species appears, whereas for the third species the score changes for each new object along the gradient. As a result of this regular displacement, a narrow band of non-zero values appears diagonally in the data matrix (=Petrie matrix, Kendall 1971). Further regularity in the data is that all objects have the same number of species (3) so that some species have their optima outside the gradient. Gradient length (species turnover) is extremely high; the species composition completely changes 13 times along the series from object 1 to 40. Objects 41 and 42 are not part of the artificial ecological data set; they are only used to demonstrate an interesting feature of PCoA ordinations.

2) *Perfectly linear* response of species to a short ecological gradient (Fig. 2). This idealizes a hypothetical situation when environmental conditions gradually and homogeneously change such that all sites remain within the tolerance limit of all species. Two subcases are distinguished: in 2a all species scores increase (positive unit correlations, only the first three rows considered), whereas in 2b there is a contrast between the trends for two sets of species (negative unit correlations also appear, all the six rows considered). A non-ecological alternative to this perfectly linear data structure comes from morphometry. Table A1 in Appendix A summarizes five measurements for 20 individuals of *Leptograpsus* crabs (Reyment 1991), providing slightly noisy linear data implying a very short gradient, used in our present paper to call attention to published results misinterpreted due to programming errors.

3) The richness gradient is a different kind of ''linear'' arrangement (Fig. 3). Here, linearity refers to a gradual change from a species-rich site on the left end, which loses one species at a time along the gradient so that at the endpoint only one species remains. Presence/absence data (Appendix B) for 80 quadrats (each measuring 4 m $\times$ 4 m) located in a dolomite grassland community in the Sashegy Nature Reserve (Budapest, Hungary) provide a basis for comparison with this artificial data set. For more on this community and results of its numerical analyses, see Podani (1985) and Podani et al. (2000).

Resemblance coefficients

Let $\mathbf{X} = \{x_{ij}\}$ denote the raw data matrix with n species (rows) and m sites or objects (columns). Let further x_{i+}, x_{+j}, and x_{++} represent row, column, and grand totals of **X,** respectively. To express distance and dissimilarity (collectively termed here as ''resemblance'') between any pair of objects j and k, the following coefficients will be used to calculate input **D** matrices for PCoA:

Euclidean distance:

$$d_{jk} = \sqrt{\sum_i (x_{ij} - x_{ik})^2} \tag{1}$$

Manhattan distance (or city-block metric):

$$d_{jk} = \sum_i |x_{ij} - x_{ik}| \tag{2}$$

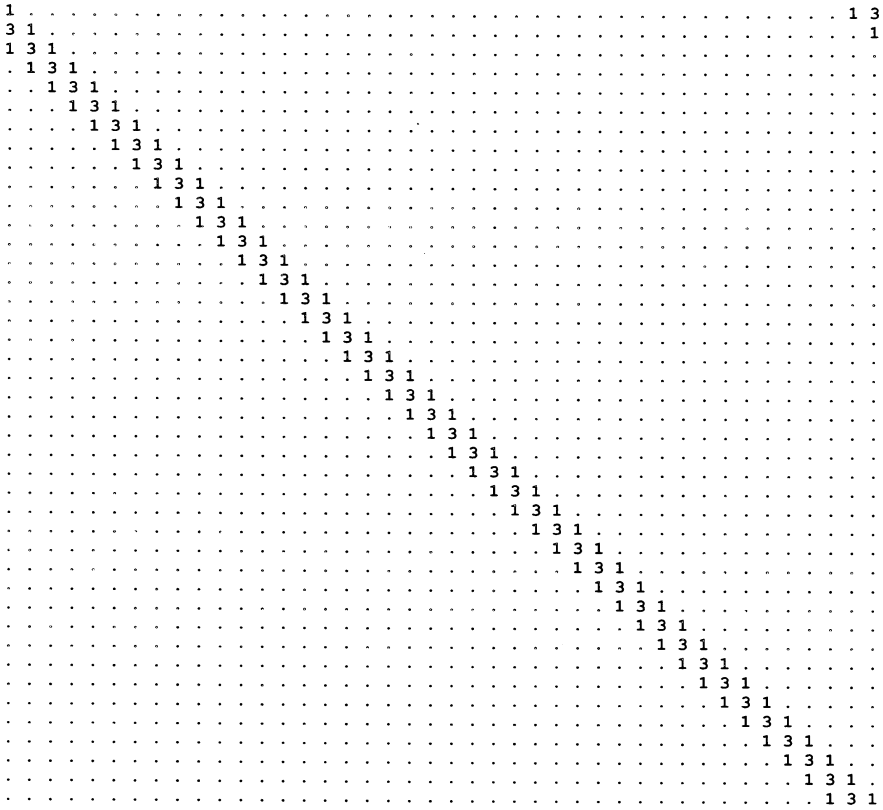

FIG. 1. Artificial data matrix representing a long one-dimensional background gradient with simplified unimodal species response. ''Species'' are in rows (42), units (objects) are the columns (40 + 2). Note that zeros are replaced by dots to emphasize the banded appearance of the main diagonal.

Chord distance:

$$d_{jk} = \sqrt{2\left[1 - \frac{\sum_i x_{ij}x_{ik}}{\sqrt{\sum_i x_{ij}^2 \sum_i x_{ik}^2}}\right]} \tag{3}$$

Complement of similarity ratio (Wishart 1969, van der Maarel 1979):

$$d_{jk} = 1 - \frac{\sum_i x_{ij}x_{ik}}{\sum_i x_{ij}^2 + \sum_i x_{ik}^2 - \sum_i x_{ij}x_{ik}} \tag{4}$$

Marczewski-Steinhaus coefficient (= complement of Ruzicka index; = percentage remoteness if multiplied by 100) (Pielou 1984):

$$d_{jk} = \frac{\sum_i |x_{ij} - x_{ik}|}{\sum_i \max\{x_{ij}, x_{ik}\}} \tag{5}$$

Bray-Curtis formula (= complement of Steinhaus similarity; = percentage difference if multiplied by 100) (Pielou 1984):

$$d_{jk} = \frac{\sum_i |x_{ij} - x_{ik}|}{\sum_i (x_{ij} + x_{ik})} \tag{6}$$

Canberra metric:

$$d_{jk} = \sum_i \frac{|x_{ij} - x_{ik}|}{(x_{ij} + x_{ik})} \tag{7}$$

Balakrishnan-Shangvi measure:

$$d_{jk} = \sum_i \frac{(x_{ij} - x_{ik})^2}{(x_{ij} + x_{ik})} \tag{8}$$

Complement of Horn index: (see Eq. 9 at bottom of page)

$$d_{jk} = 1 - \frac{\sum_i (x_{ij} + x_{ik})\log(x_{ij} + x_{ik}) - \sum_i x_{ij}\log x_{ij} - \sum_i x_{ik}\log x_{ik}}{(x_{\cdot j} + x_{\cdot k})\log(x_{\cdot j} + x_{\cdot k}) - x_{\cdot j}\log x_{\cdot j} - x_{\cdot k}\log x_{\cdot k}} \tag{9}$$

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

FIG. 2. Artificial data matrix representing perfectly linear response to a one-dimensional underlying gradient. The number of species (rows) is 6, the number of objects (columns) is 40.

Hellinger distance (Rao 1995, Legendre and Gallagher 2001):

$$d_{jk} = \sqrt{\sum_i \left[\sqrt{\frac{x_{ij}}{x_{+j}}} - \sqrt{\frac{x_{ik}}{x_{+k}}}\right]^2} \quad (10)$$

Chi-square distance:

$$d_{jk} = \sqrt{x_{++}} \sqrt{\sum_i \frac{1}{x_{i+}}\left[\frac{x_{ij}}{x_{+j}} - \frac{x_{ik}}{x_{+k}}\right]^2} \quad (11)$$

(see Podani [2000], for more on coefficients mentioned above without reference). We suggest a *new distance* formula, conceptually ''orthogonal'' to the Hellinger distance; it implies standardization by species totals and square-root transformation. The ecological meaning of these operations is underweighting of common species and overemphasizing the less abundant ones. Its formula is given by

$$d_{jk} = \sqrt{\sum_i \left[\sqrt{\frac{x_{ij}}{x_{i+}}} - \sqrt{\frac{x_{ik}}{x_{i+}}}\right]^2} \quad (12)$$

These indices incorporate various forms of data transformation, standardization, and normalization and therefore provide a reasonably wide range of possibilities to examine gradient recovery of PCoA.

Principal coordinates analysis

Perhaps PCoA (alias metric multidimensional scaling, Gower 1966, Torgerson 1958) is the most general metric ordination procedure. The advantage of PCoA is its ability to start from any symmetric resemblance matrix, which is not so with other metric methods. PCoA attempts to embed the resemblance structure into a Euclidean space—its success depending on the type of coefficient. If the input is Euclidean distances (Eq. 1), as originally suggested, then the first step comprises the transformation of d values according to

$$a_{jk} = -0.5 d_{jk}^c \quad (13)$$

with the exponent $c = 2$. Matrix **A** thus obtained is centered and then subjected to eigenanalysis to derive eigenvalues and eigenvectors that in turn provide the ordination scores. A major issue that concerns us in this paper is the appropriate choice of c for a wide range of coefficients. As we shall demonstrate below, this question is not as trivial as earlier thought.

Computations were performed using the SYN-TAX 2000 package (Podani 2001) and its PRINCOOR module modified to allow run-time definition of exponent c.

RESULTS

When horseshoes are typical: unimodal response to a long gradient

Principal coordinates analysis (PCoA) of the data in Fig. 1 resulted in the same general picture for distance coefficients 1–10. This is perhaps unexpected, if we consider how different manipulations with the data are involved in these coefficients. On the plane of axes 1 and 2, the familiar horseshoe is observed with involutions directed towards the origin (Fig. 4a). If higher axes are portrayed together with the first axis, then more complex configurations emerge (Appendix A: Figs. A1 and A2), all representing polynomial relationship to axis 1, complicated by involutions. Raising d to powers higher than 2 led to a very slight increase in the first few eigenvalues, with the ordination con-

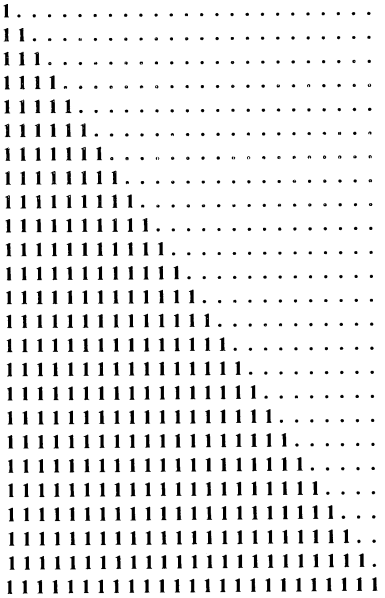

FIG. 3. Artificial data representing a species richness gradient. ''Species'' are in rows; sites (objects) are the columns. Note that zeros are replaced by dots to emphasize the presence scores.

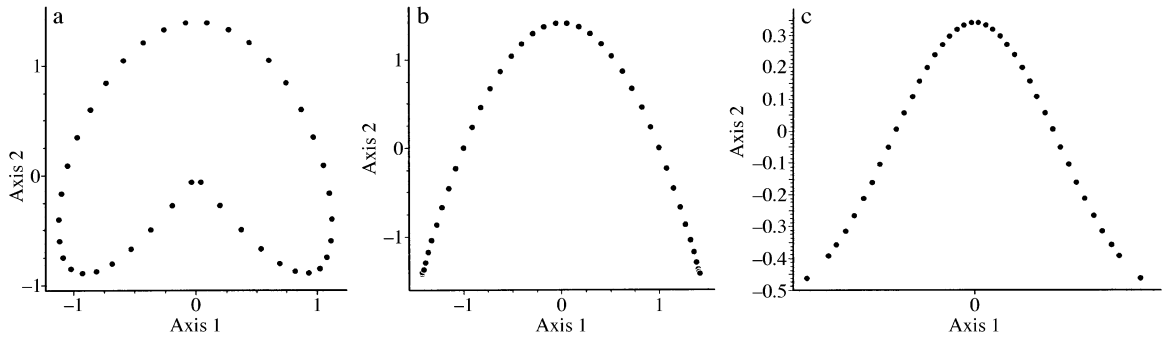

FIG. 4. Horseshoe effect in PCoA (principal coordinates analysis) ordinations of nonlinear data from long background gradient with unimodal species response (Fig. 1) for axes 1–2 and for coefficients 1–10 (a), 11 (b), and 12 (c). Each ordination comprises 40 objects, numbered 1 to 40 from left end to right end in each diagram.

figuration remaining largely intact. However, negative eigenvalues appeared, and their magnitude increased along with increases of *c,* a phenomenon to be examined and discussed later (see *The horseshoe paradox*, below). The chi-square distance (Eq. 11) yielded a parabola without involutions (Fig. 4b). This is an expected result if we recall the relationship between this measure and correspondence analysis (CoA), the latter being known to be free from involutions in most cases (Legendre and Legendre 1998). The new formula (Eq. 12) produced a unique configuration—a ''bell-shaped'' arrangement of points in the first two PCoA dimensions (Fig. 4c), with extremities folded mildly outwards.

It is suggestive to examine what happens if the two endpoints of the gradient are ''connected,'' i.e., the last two columns of Fig. 1 are also included in analysis. This change may be interpreted ecologically that some cyclic (or periodical) processes operate in the background and sampling units are collected evenly from all stages of that process. The result is an entire regular circle of points on axes 1 and 2, a completely adequate representation of the cyclic structure, with malformations added only in higher dimensions. Thus, a relatively small change in the data had enormous consequence as to the shape of the point scatter. Iwatsubo (1984) has analyzed this problem for a special kind of matrix; a generalization is given in Appendix C.

A further problem that deserves our attention here is how the nature of the end points influences the resulting ordinations. To examine this, the first two and the last two rows of Fig. 1 were deleted. As a result, all species have their optima within the range of study and therefore extreme sites are poorer in species than the middle sites. The PCoA ordination was not or little affected for coefficients 1, 2, 4–9, and 11. Chord distance and Hellinger distance, however, turned out to be very sensitive to this change: a ''bell-shaped'' arch was produced like the one in Fig. 4c, complicated with abrupt involutions of the extreme sites at both ends. The new distance formula (Eq. 12) was also strongly affected by the poorer end points, in this case producing an ordination almost indistinguishable from the one in Fig. 4a.

The horseshoe paradox

Surprisingly, whereas most coefficients produced similar results in the analysis of strongly nonlinear data, this was not so with the perfectly linear data at all. Dependence on the coefficient is further complicated by the exponent used in Eq. 13 and by the relationships among variables as well. To demonstrate these issues, we first used variables 1–3 from Fig. 2, which have positive unit correlations with one another. Linearity was perfectly recovered for the Euclidean, Manhattan, Hellinger, chi-square, and the new distance measure whereas chord distance and the Horn index could not be used because all distances were 0. What is of more importance here is that the other coefficients (4–8) gave rise to a paradoxical horseshoe-like configuration with interpoint distances gradually shortened

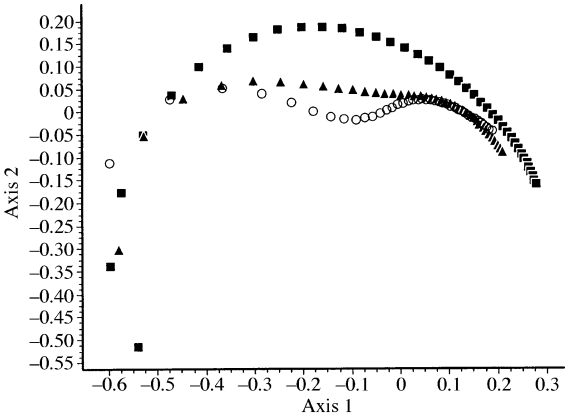

FIG. 5. The horseshoe paradox and the flattening effect of power transformation of dissimilarities. PCoA (principal coordinates analysis) ordination of linear data (Fig. 2, first 3 variables) using the Bray-Curtis dissimilarity raised to the power of $c = 2$ (solid squares), $c = 4$ (solid triangles), and $c = 6$ (open circles) before eigenanalysis. Objects are numbered 1 to 40 from left end to the right end of the arch.

TABLE 1. The efficiency of principal coordinates analysis in recovering linear data structures (Fig. 2) for various resemblance coefficients and different values of the exponent *c*.

Resemblance coefficient	*c*	λ_1	λ_2	λ_3	No. negative eigenvalues	Deviation (%)
a) Three-species case						
Eq. 4., Similarity ratio	2	67.8	25.0	5.0	6	7
	4	82.0	17.0	0.5	4	18
	6	88.6	9.3	1.0	3	24
Eq. 5., Marczewski-Steinhaus	2	57.8	24.1	8.7	0	0
	4	77.8	20.8	1.4	6	7
	6	86.4	13.5	0.1	4	15
Eq. 6., Bray-Curtis and Eq. 7, Canberra	2	72.9	21.9	4.1	0	0
	4	90.9	9.1	0	4	16
	6	96.3	2.4	0.7	2	28
Eq. 8., Balakrishnan-Shangvi	2	92.5	6.6	0.8	0	0
	4	100	0	0	7	31
	6	98.4	1.2	0.4	7	48
b) Six-species case						
Eq. 3., Chord	2	96.3	3.7	0	0	0
	4	100	0	0	2	28
	6	98.7	1.0	0.2	2	42
Eq. 4., Similarity ratio	2	99.2	0.8	0	4	12
	4	97.7	1.5	0.8	2	36
	6	96.7	2.5	0.7	2	48
Eq. 5., Marczewski-Steinhaus	2	78.5	14.9	3.8	0	0
	4	100	0	0	3	15
	6	99.6	0.2	0.1	2	33
Eq. 7., Canberra	2	77.0	14.5	5.2	0	0
	4	100	0	0	6	24
	6	98.9	0.6	0.3	5	45
Eq. 8., Balakrishnan-Shangvi	2	90.7	7.3	1.4	0	0
	4	100	0	0	12	30
	6	98.4	1.2	0.4	15	49
Eq. 9., Horn	2	100	0	0	2	36
	4	97.5	2.1	0.4	2	65
	6	97.6	2.2	0.2	2	77
Eq. 10., Hellinger	2	97.1	2.9	0	0	0
	4	100	0	0	2	40
	6	97.9	1.7	0.4	2	58
Eq. 12., New distance	2	97.1	2.9	0	0	0
	4	100	0	0	2	50
	6	97.9	1.7	0.4	1	90

Notes: Eigenvalues (λ) are given as percentages of total of positive eigenvalues. Coefficients that could not be used or provided perfect recovery for $c = 2$ without negative eigenvalues are not listed (see columns 4–5 in Table 2).

from object 1 to 40 (squares in Fig. 5 represent this ordination based on the Bray-Curtis coefficient). By ''paradoxical'' we mean that a linear ordination method from linear data produced a curvilinear arrangement and that the coordinates on axis 1 do not reproduce the correct sequence of objects. This arch was more and more flattened and thus linearity approximated for all coefficients when *c* was raised to 4 and 6 (triangles and circles, respectively, in Fig. 5). Therefore, using increased values of *c* may be considered as a detrending operation; this process is associated with an increased concentration of total variance into the first eigenvalue (Table 1a), a phenomenon most conspicuous for the Balakrishnan-Shangvi measure—the curvature obtained using $c = 2$ is completely unfolded with $c = 4$. Increases of variance on the first axis are compensated for by negative eigenvalues indicating that the input dissimilarities cannot be embedded perfectly into a Euclidean space (Gower and Legendre 1986, Legendre and Legendre 1998). As a measure of deviation from the Euclidean condition, we suggest the ratio of the sum of negative eigenvalues on the sum of positive eigenvalues, multiplied by 100 (Table 1, last column). Low percentage indicates that the two-dimensional configuration is little affected by negative eigenvalues and, as suggested by Digby and Kempton (1987), these can be ignored as long as they remain as small as the minor positive values. Great deviation implies a strongly non-Euclidean resemblance structure, however, although detrending by power transformation in such cases can still provide meaningful results. We have tried the method suggested by Gower and Legendre (1986, see also Legendre and Anderson 1999) to eliminate the negative eigenvalues (by adding a constant *b*

to all non-diagonal elements of the dissimilarity matrix) to see if this operation improves eigenvalue extraction (we used $c = 6$). The inflation of the positive eigenvalues achieved this way did eliminate the negative ones, but increasing values of b caused the ordination to approximate the configuration obtained for $c = 2$. At the end, with high enough value of b, the analysis completely reproduced that starting ordination! Therefore, no correction for negative eigenvalues is suggested if the objective is gradient recovery.

The data matrix of Fig. 2 with all the six species (rows) included implies a correlation matrix with a mixture of positive and negative unit correlations. Losses for the first three variables are compensated for by the gains of the second three variables along the gradient. This structure leads to a point scatter completely different from the three-variable case for many of the coefficients, while Euclidean, Manhattan, and chi-square distances remain unaffected as expected. The Bray-Curtis formula and the Horn index now resulted in a 100% gradient recovery, the latter with negative eigenvalues. Hellinger distance, the similarity ratio, chord distance, and the new distance measure (Eq. 12) yielded a very high first eigenvalue (Table 1b), with approximately linear arrangement of points on the first axis, and a weak arch on the second. The Canberra metric, the Marczewski-Steinhaus coefficient, and the Balakrishnan-Shangvi measure still produce an arch, although more flattened than for the three-variable case discussed above. It must be pointed out that these small curvatures were entirely removed from the analysis when d was raised to the power of 4. Results for the six-species case suggest that once 100% gradient recovery is reached, further increases of c can only worsen the situation.

True artifacts obtained from linear data

The morphometric data for *Leptograpsus* crabs (Appendix A: Table A1) lead to pairwise correlations ranging from 0.92 to 0.99 for the five variables. This is a strongly linear data structure with a single overwhelming background ''gradient,'' i.e., overall animal size. PCA (principal components analysis) confirms this expectation: component 1 explains >97% of the total variance. As an intriguing contrast, Reyment (1991) reported that a PCoA from a matrix computed by Gower's formula (Gower 1971) from the raw data yielded a ''horseshoe'' plot of the individuals for axes 1 and 2 (Reyment 1991: Fig. 3.2). This is due to a serious programming error, namely the transformation given by Eq. 13 is missing from Reyment's (1991: S8) computer program. The error is ''inherited'' probably from the program by Blackith and Reyment (1971), from which the transformation is also missing (P. Legendre, *personal communication*).

The use of Gower's formula in PCoA needs a deeper inspection, however. An additional source of artifacts is the use of $c = 1$ in the transformation (Eq. 13) on the assumption that d_{jk} is already squared (i.e., it is a *dissimilarity*). The PcoA of Manhattan distances among the crabs introduces an artificial arch into the result with $c = 1$ (Appendix A: Fig. A3). The analysis with $c = 2$ removes the polynomial shape and the ordination will be dominated by a linear axis, leaving stochastic variation to the second one (eigenvalues as percentages are $\lambda_1 = 98.9\%$, $\lambda_2 = 0.5\%$). Gower's index, if applied to the quantitative case, implies the same absolute difference operation as the Manhattan metric, albeit standardized by range, which is immaterial here. Consequently, this index requires the same treatment in PCoA as the Manhattan metric. Proper analyses (i.e., $c = 2$) of this data set by other coefficients (except Horn and chord distance) completely recover the one-dimensional size component with very high first eigenvalues (>87%). It should be pointed out, however, that notwithstanding the high first eigenvalue the Bray-Curtis measure, the Marczewski-Steinhaus coefficient, and the Canberra metric still produce a weak, hardly noticable arch along axes 1–2.

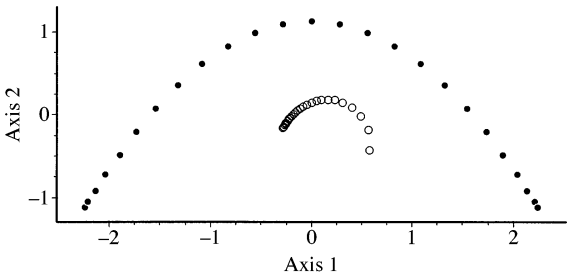

FIG. 6. Standard ($c = 2$) PCoA (principal coordinates analysis) of data with a species-richness gradient (Fig. 3) using Euclidean distance (solid circles) and the Bray-Curtis formula (open circles). Numbering of objects is 1 to 25 from left to right.

The richness arch

The PCoA of the data of Fig. 3 (with $c = 2$) demonstrates that Euclidean distance and Manhattan metric, in agreement until now, can produce different results under certain circumstances. Euclidean distance yields an arch ($\lambda_1 = 62\%$; $\lambda_2 = 16\%$, and $\lambda_3 = 7\%$, Fig. 6). Manhattan metric yields what a general surveyor would expect for this kind of data: the gradual change is manifested as a straight line in the ordination diagram ($\lambda_1 = 100\%$). Of the remaining coefficients, only the Canberra metric produced this result, due to its identity with Manhattan metric in the presence/absence case. All other coefficients (including the chi-square metric and therefore CoA) resulted in more or less pronounced horseshoe shapes, with λ_1 ranging from 43 to 76%, and leaving substantial percentages for uninterpretable subsequent dimensions. As an example, the PCoA ordination based on the Bray-Curtis index (= 1 − Sørensen index in this case) is depicted in Fig. 6 ($\lambda_1 = 76\%$, $\lambda_2 = 21\%$, and $\lambda_3 = 3\%$).

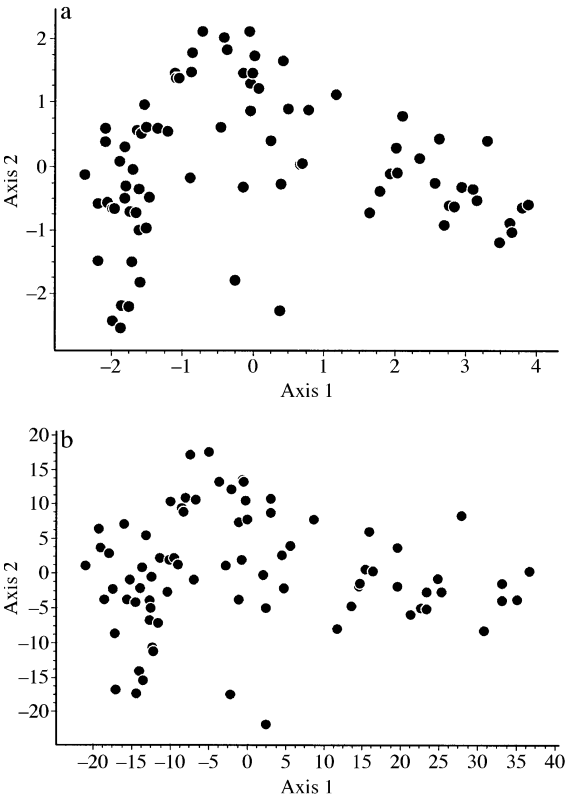

FIG. 7. PCoA (principal coordinates analysis) recovery of a species-richness gradient for the Sashegy grassland data using Euclidean distance for presence/absence, with two different exponents: (a) $c = 2$, $\lambda_1 = 22\%$, $\lambda_2 = 8\%$. (b) $c = 4$, $\lambda_1 = 35\%$, $\lambda_2 = 10\%$. Note that ordination (b) is equivalent to the PCoA of the same data with the Manhattan metric at $c = 2$.

The influence of modifying c is apparent for this gradient type. (Euclidean distance with $c = 4$ need not be mentioned, because in the presence/absence case it is identical to Manhattan metric with $c = 2$, discussed already.) The ordinations improved greatly for the other coefficients; the curves became more flattened and the first eigenvalues were higher when c was increased to 4, but negative eigenvalues appeared.

Standard PCoA of the Sashegy Nature preserve presence/absence data (Appendix B) using Euclidean distance ($c = 2$) provides a noticable arch of the points, with two outliers only (Fig. 7a). Remarkably, PCoA from Manhattan measures with $c = 2$ removes much of the arch from the configuration: the right arm associated with axis 2 becomes shorter while the right arm is relatively longer (Fig. 7b). That the arrangement of points on the first axis follows a species-richness gradient is confirmed by the very high product moment correlation ($r = 0.83$) between species richness and axis 1. Actually, the quadrats with low coordinate originate from open areas with southern exposure and steep slope. On the other end, we find quadrats of a closed community from northeastern exposition with gentle slope.

Matrix shape vs. gradient recovery

The above analyses allow a generalization to be made between the input matrix and the resulting ordination. This is possible because the artificial data are prepared on purpose in such a way that the **D** matrix is optimally arranged. It is then most suggestive to examine the change of d^c in the input matrices in a direction perpendicular to the main diagonal, from the lower left to the upper right corner. If d^c first increases and then levels off (Fig. 8a), as in case of the long gradient with unimodal species response, then the re-

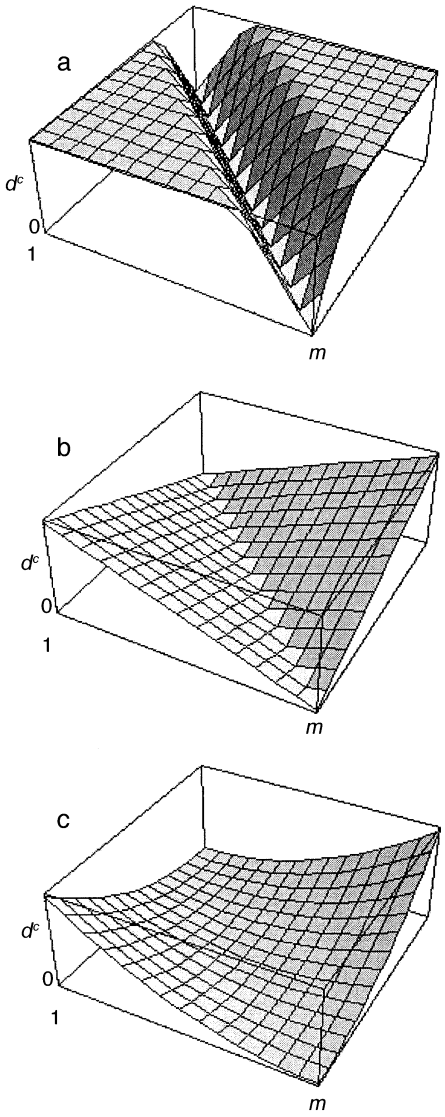

FIG. 8. Graphical illustration of the relationship between d^c and its position in dissimilarity matrix $\mathbf{D}_{m,m}$ with reference to gradient recovery by PCoA (see *Results: Matrix shape* vs. *gradient recovery* for explanation).

TABLE 2. Gradient recovery of various resemblance coefficients in principal coordinates analysis (PCoA) based on artificial data used in this paper. For linear data, r is the correlation coefficient between species.

Coefficient	Unimodal response		Linear data		Richness gradient
	Rich ends	Poor ends	$r = 1.0$	$r = 1.0$ or -1.0	
1. Euclidean	ooo	ooo	†††	†††	o
2. Manhattan	ooo	ooo	†††	†††	†††
3. Chord	ooo	ooo	x	††	o
4. Similarity ratio	ooo	ooo	o	††	o
5. Marczewski-Steinhaus	ooo	ooo	o	o	o
6. Bray-Curtis	ooo	ooo	o	†††	o
7. Canberra	ooo	ooo	o	o	†††
8. Balakrishnan-Shangvi	ooo	ooo	†	†	o
9. Horn	ooo	ooo	x	†††	o
10. Hellinger	ooo	ooo	†††	††	o
11. Chi-square	oo	oo	†††	†††	o
12. New distance	oo	ooo	†††	††	o

Note: Gradient-recovery key symbols: ooo = extremely poor recovery (very low first eigenvalue, strong involutions), oo = moderately poor recovery (very low first eigenvalue, strong arch without involutions), o = poor recovery (first eigenvalue between 40–80%, arch sometimes with slight involutions), † = flattened arch (first eigenvalue 80–95%), †† = almost straight line (95% < first eigenvalue < 100%), ††† = perfect recovery (first eigenvalue = 100%), x = could not be tested.

sulting ordination will be an involuted horseshoe. This is in agreement with Kendall's (1971) finding: he attributed the horseshoes to the use of bounded metrics where beyond some threshold all distances are equal. If d^c increases linearly towards the corners (Fig. 8b), then the configuration will be a parabola. This is the case when Manhattan distances are applied to perfectly linear data and the exponent is only 1 (a proof is given in Appendix D). However, when the relationship is a parabola (Fig. 8c), then the resulting ordination will reproduce an entirely linear arrangement (Manhattan distance with $c = 2$).

DISCUSSION

The main message in the present paper is that the horseshoe (or arch) effect often encountered in principal coordinates analysis (PCoA) ordinations can be a manifestation of several factors. They appear as true artifacts whenever transformation (Eq. 13) of the input resemblance matrix is omitted or inappropriate—a problem easily avoided by careful programming. Contrarywise, appearance of arches or horseshoes, even though often paradoxical, is a *mathematical* necessity in the other situations examined in this study. There is no single resemblance measure that always performs well; efficiency of gradient recovery depends primarily on gradient type (Table 2). This observation has far-reaching consequences in numerical ecology, suggesting that researchers should always be careful when selecting resemblance coefficients for PCoA ordinations.

For unimodal data structures, most resemblance coefficients (collectively expressing both distance and dissimilarity) led to the familiar horseshoe with strong involutions, except for chi-square distance and some other coefficients that proved to be sensitive to the nature of gradient ends. For high turnover, we demonstrated that the horseshoe is in fact a transition towards a circle: a mathematically slight modification of the input data, interpreted ecologically as a manifestation of periodical changes, reveals the underlying circularity of such data. For perfectly linear data structures, arches and expected linear arrangements may both appear, depending on the choice of the resemblance coefficient. For example, the Bray-Curtis index produces an arch if all variables are positively correlated. This is because for a given pair of objects the same *absolute* difference is more influential when the totals are also small. This asymmetry causes the dissimilarities between neighboring object pairs to gradually decrease from the low-abundance (or species-poor) sites towards the high-abudance (or species-rich) sites, as do dissimilarities between the second, third, etc., neighbors. As a result, the distances can be best reproduced by an arch in the ordination plane defined by axes 1–2. This arch, however, is much less conspicuous and the ordination provides an acceptable gradient recovery if proportional differences among objects are much smaller—as exemplified by analyses of the *Leptograpsus* data (see *Results: True artifacts obtained from linear data*, above; Appendix A). Furthermore, if losses are compensated for by gains in other species along the gradient (balanced mixture of positive and negative unit correlations), the Bray-Curtis coefficient-based PCoA will be entirely free of the arch.

In the case of a hypothetical richness gradient, the Manhattan and Canberra metrics correctly indentified the background change, whereas the other measures, including the chi-square distance, responded by more or less pronounced arches in the ordination. We ob-

served that increasing the value of the c exponent always has some ''detrending'' effect. Curved arrangement and eigenvalue extraction were the least influenced by increasing exponents for the unimodal response on long gradient, whereas detrending was more significant for the other two gradient types examined here. However, raising input dissimilarities to the power >2 introduces negative eigenvalues that become relatively large when c is increasing. The reason is that for any triplet i, j, k of objects, provided that they can be labeled such that $d_{jk} > d_{ik}$, d_{ij}, the triangle inequality is always violated if $c \gg 2$. The problem of negative eigenvalues resulting from the use of nonmetric coefficients in PCoA has been discussed in detail in the literature (e.g., Gower and Legendre 1986). Our studies suggest that negative eigenvalues arising from power transformation are unavoidable concomitants of the procedure; this is the price we have to pay in order to get a detrended point configuration. Therefore, the effect of power transformation must always be evaluated through comparisons with the original PCoA results.

In analyses of actual community data, the relative importance of factors that can lead to arched arrangements (e.g., gradient type, interplay of variables) is unknown to the user a priori. In fact, unimodal-response and species-richness changes may be present simultaneously, and therefore attributing the same ecological explanation to an observed horseshoe may be wrong. The current literature of numerical ecology appears to overemphasize the ''unimodal response type'' of species as the primary reason behind ordination horseshoes, and does not observe other reasons (to mention a few: Gauch 1982, Digby and Kempton 1987, ter Braak 1987, Ludwig and Reynolds 1988, ter Braak and Prentice 1988). It is therefore not surprising that all works evaluating the performance of ordination methods use coenoclines and coenoplanes simulated on this logical foundation (from Swan [1970] to Økland [1999]). We demonstrated that, depending on the resemblance coefficient used, linear response within a shorter ecological gradient or gradual richness change may just as well be behind the ''failure'' of PCoA to recognize unidirectional changes in a ''comfortable'' way. Also, due to the close relationship between Euclidean distance-based PCoA and species-centered PCA (principal components analysis), our comments regarding Euclidean distance in PCoA directly apply to the evaluation of PCA results.

The complexity of this issue was illustrated by the PCoA of actual grassland data, confirming that the richness example is not speculative at all. The high correlation between axis 1 and species richness and the fact that for Manhattan distance the horseshoe effect diminishes reveal that it is an actual analogue to the hypothetical richness gradient of Fig. 3. Of course, this is not to say that there is no ecological change behind species richness changes (in the present case: microclimate as influenced by exposition and slope), but the nature of collective species behavior is different from what most ecologists assume when a horseshoe-shaped configuration arises. Species richness changes can appear in many kinds of ecological data, for example, in a sample representing post-fire regeneration processes in grassland vegetation or in a transect running into an old-growth forest. In a review, Pausas and Austin (2001) present many examples for high correlations between species richness and different kinds of environmental gradients, with no reference to the actual data pattern associated with the change. The data in Fig. 3 represent one possibility, while—as already mentioned—species richness changes may appear simultaneously with other types of the gradient, such as the one in Fig. 1. The problem whether an ordination horseshoe is explained by unimodal response of species or by mere species-richness changes can be resolved by the comparison of Euclidean and Manhattan distance-based ordinations and by calculating correlation between the first axis and species richness. Such an analysis should always complement plain correspondence analysis (CoA) as well, because CoA (or the use of chi-square distance in PCoA) also confounds unimodal response of species with species-richness changes. Of course, if quantitative data rather than presence/absence scores are used, the situation becomes more complex so that comparisons between analyses based on these two data types are also recommended.

The present study by no means exhausted all the possibilities for defining background gradient types, so that shapes different from those illustrated and evaluated here can also appear. For example, we used a very simple unimodal response model with regular species displacement, which can be altered by changing gradient length (species turnover), niche breadth of individual species, and the relative position of their optima, to mention just a few possibilities. We may refer now to Legendre and Legendre (1998) who used only three species with a wider niche and large spacing on a short gradient to demonstrate the performance of PCoA with the Bray-Curtis coefficient (Legendre and Legendre 1998: Table 9.9 and Fig. 9.17e–h, boldface values in our Fig. 9a). In the resulting ordination, reproduced here as Fig. 9b, there is a small twist outwards, most striking in dimension 3, before the ends are involuted on both arms of the horseshoe. This twist, uninterpreted by the authors, is missing from the Euclidean-distance-based analysis and may be attributed to the superposition of two ''effects'': the main involuted arch resulting from the unimodal response type (Fig. 4a) and the fact that distances do not increase monotonically before reaching the plateau (Fig. 9d). This data set, if more species are used to fill the ''gaps'' between the three original species (all values in Fig. 9a), may be used to demonstrate one more interesting phenomenon: PCoA with Canberra metric produces an almost regular circle, albeit with unequal spacing be-

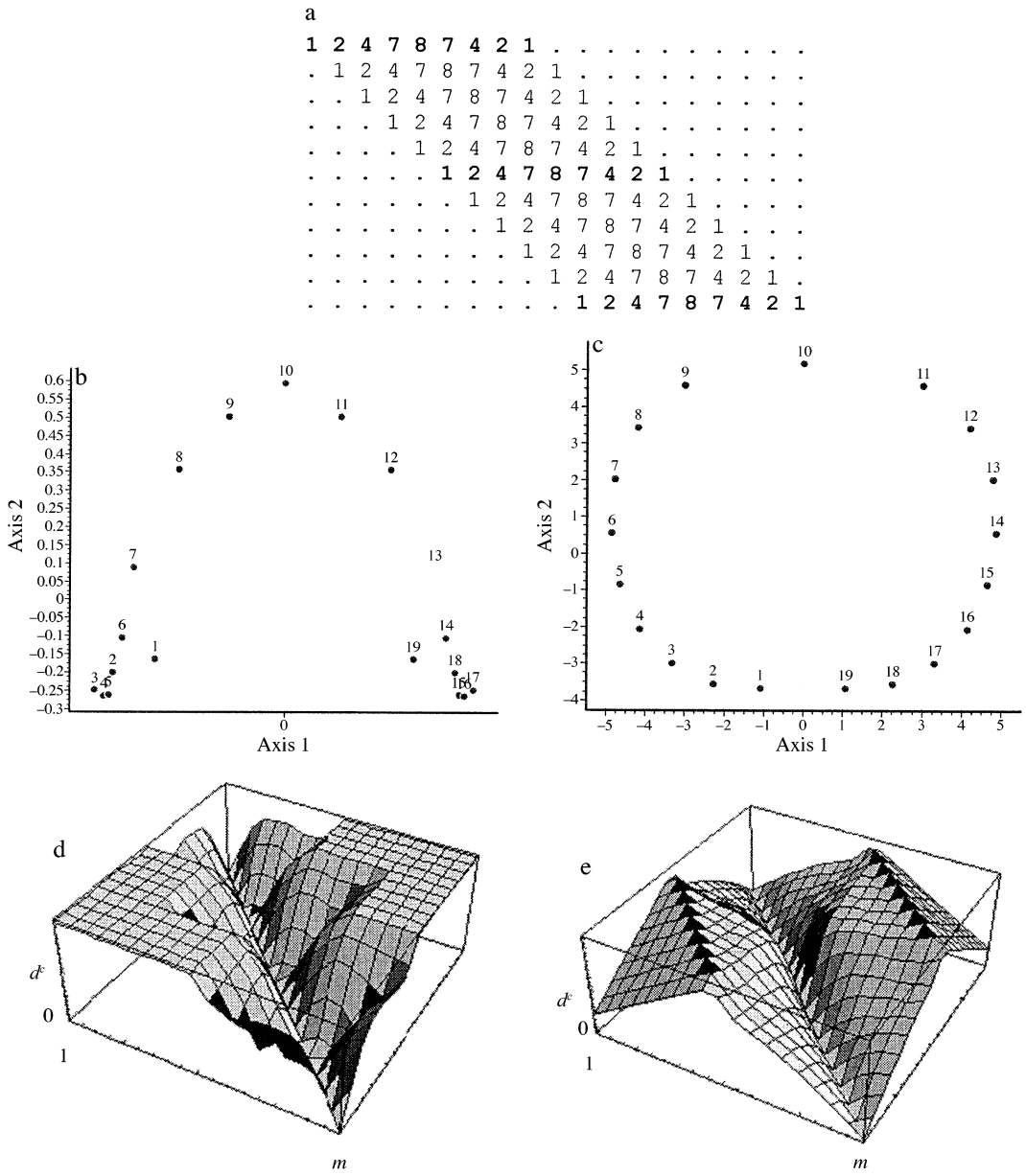

FIG. 9. PCoA (principal coordinates analysis) performance for data not treated in this paper in detail. (a) Raw data for three species (rows) on a short gradient (from Legendre and Legendre [1998: Table 9.9], here in boldface), and the data matrix expanded to ensure a more gradual species turnover (all scores). Zeros are replaced by dots. (b) PCoA based on Bray-Curtis coefficient for Legendre and Legendre's (1998) original three-species data, exhibiting a twisted horseshoe; (c) PCoA based on Canberra metric for the expanded data matrix, giving a paradoxical circle. (d, e) Perspective views of the two respective $m \times m$ dissimilarity matrices (with the number of objects running from 1 to m [x-axis and z-axis] and d^c, the input dissimilarity raised to the power of c [and running from 0 to other values not shown]).

tween the neighboring points (Fig. 9c, while Fig. 9e provides a perspective view on the dissimilarity matrix). That is, a circle is obtained from data implying unimodal response for all species with low turnover rate, a potential source of misinterpretation if the researcher is not careful enough! Note that a quite similar circle, with different interpoint spacing is obtained by Euclidean distance and the Bray-Curtis coefficient as well, whereas the chi-square distance produces an arch as expected, further illustrating the confounding effect of gradient type and coefficient behavior. We have a good reason to suppose that other complex shapes may also appear for different combinations of gradient type and coefficient (the reader may verify, for example,

that for the three-species case Canberra metric produces a square-like arrangement of points).

Although the current literature appears to distinguish between arches (without involutions) and horseshoes (with involutions) only, there are more types of curves that may appear in ecological ordinations if the method is PCoA. The configurations in Fig. 5, for example, are neither arches nor horseshoes in the usual sense: there is an apparent asymmetrically curved series of points with interpoint distances gradually diminishing towards the end of the gradient. Also, depending on the properties of gradient ends, the newly proposed coefficient may generate a bell-shaped configuration from data representing the unimodal response type of species. The results suggest that this coefficient improves gradient recovery on axis 1 (i.e., a bell-shaped point scatter arises) if some species have their optima outside each end of the gradient. For the other gradient types examined in the present paper, the new formula (Eq. 12) does not perform better than the other distance functions. Further studies are required to evaluate this coefficient under circumstances different from those evaluated here.

Ecologists seem to be divided into two opposing parties based on their attitude towards the treatment of ordination curvatures. Many of them believe that this phenomenon is harmful in indirect gradient analysis and therefore all effort must be made to eliminate the arch a priori from the ordination. The alternative view considers the ''horseshoe'' as an unavoidable by-product of ordinations of data which violate the assumption of linearity, and insists that its recognition in the final results is easy and does not hinder their interpretability (cf. Dale 1994). Again, what is usually missing from pro and contra arguments is a clear explanation of the reasons and a detailed typification of this phenomenon. The findings of the present paper complement significantly the debate, even though it is mostly around CoA (and occasionally PCA). Our examples demonstrated fairly convincingly that automatic detrending a PCoA horseshoe would be unwise without knowing the results and without identifying the reasons behind curved arrangements.

Acknowledgments

Many thanks are due to P. Legendre and two anonymous referees for their thorough and constructive criticism. We are grateful to M. B. Dale for his extremely useful comments on an earlier version of the manuscript. Our research was supported by the Hungarian Scientific Research Fund (OTKA), grant number 29784 given to the first author.

Literature Cited

Austin, M. P., and I. Noy-Meir. 1971. The problem of nonlinearity in ordination: experiments with two-gradient models. Journal of Ecology **59**:763–773.

Blackith, R. E., and R. A. Reyment. 1971. Multivariate morphometrics. Academic Press, London, UK.

Bradfield, G. E., and N. C. Kenkel. 1987. Nonlinear ordination using flexible shortest path adjustment of ecological distances. Ecology **68**:750–753.

Dale, M. B. 1994. Straightening the horseshoe: a Riemannian resolution? Coenoses **9**:43–53.

Dale, M. B. 2000. On plexus representation of dissimilarities. Community Ecology **1**:43–56.

De'ath, G. 1999*a*. Principal curves: a new technique for indirect and direct gradient analysis. Ecology **80**:2237–2253.

De'ath, G. 1999*b*. Extended dissimilarity: a method of robust estimation of ecological distances from high beta diversity data. Plant Ecology **144**:191–199.

Digby, P. G. N., and R. A. Kempton. 1987. Multivariate analysis of ecological communities. Chapman and Hall, London, UK.

Gauch, H. G. 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge, UK.

Goodall, D. W. 1954. Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. Australian Journal of Botany **2**:302–324.

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

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

Gower, J. C., and P. Legendre. 1986. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification **3**:5–48.

Greenacre, M. J. 1984. Theory and applications of correspondence analysis. Academic Press, London, UK.

Heiser, W. J. 1987. Joint ordination of species and sites: the unfolding technique. Pages 189–121 *in* P. Legendre and L. Legendre, editors. Developments in numerical ecology. Springer-Verlag, Berlin, Germany.

Iwatsubo, S. 1984. The analytical solutions of an eigenvalue problem in the case of applying optimal scoring method to some types of data. Pages 31–40 *in* E. Diday et al. editors. Data analysis and informatics. North Holland, Amsterdam, The Netherlands.

Kendall, D. G. 1971. Seriation from abundance matrices. Pages 215–252 *in* F. R. Hodson, D. G. Kendall, and P. Tautu, editors. Mathematics in the archaeological and historical sciences. Edinburgh University Press, Edinburgh, UK.

Kenkel, N. C., and L. Orlóci. 1986. Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology **67**:919–928.

Legendre, P., and M. J. Anderson. 1999. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological Monographs **69**:1–24.

Legendre, P., and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia **129**:271–280.

Legendre, P., and L. Legendre. 1998. Numerical ecology. Second edition. Elsevier, Amsterdam, The Netherlands.

Ludwig, J. A., and J. F. Reynolds. 1988. Statistical ecology. Wiley Interscience, New York, New York, USA.

Økland, R. H. 1999. On the variation explained by ordination and constrained ordination axes. Journal of Vegetation Science **10**:131–136.

Pausas, J. G., and M. P. Austin. 2001. Patterns of plant species richness in relation to different environments: an appraisal. Journal of Vegetation Science **12**:153–166.

Pielou, E. C. 1984. The interpretation of ecological data. Wiley Interscience, New York, New York, USA.

Podani, J. 1985. Syntaxonomic congruence in a small-scale vegetation survey. Abstracta Botanica **9**:99–128.

Podani, J. 2000. Introduction to the exploration of multivariate biological data. Backhuys, Leiden, The Netherlands.

Podani, J. 2001. SYN-TAX 2000. User's manual. Scientia, Budapest, Hungary.

Podani, J., P. Csontos, and J. Tamás. 2000. Additive trees in the analysis of community data. Community Ecology **1**: 33–41.

Rao, C. R. 1995. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Quaderns d'Estadística i Investigacaó Operatíva **19**: 23–63.
Reyment, R. A. 1991. Multidimensional palaeobiology. Pergamon, Oxford, UK.
Swan, J. M. A. 1970. An examination of some ordination problems by use of simulated vegetational data. Ecology **51**:89–102.
ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology **67**:1167–1179.
ter Braak, C. J. F. 1987. Ordination. Pages 81–173 *in* R. H. G. Jongman, C. J. F. ter Braak, and O. F. R. van Tongeren, editors. Data analysis in community and landscape ecology. Center for Agricultural Publishing and Documentation, Wageningen, The Netherlands.
ter Braak, C. J. F., and I. C. Prenctice. 1988. A theory of gradient analysis. Advances in Ecological Research **18**: 271–317.
Torgerson, W. S. 1958. Theory and methods of scaling. John Wiley and Sons, New York, New York, USA.
van der Maarel, E. 1979. Multivariate methods in phytosociology, with reference to the Netherlands. Pages 163–225 *in* M. J. A. Werger, editor. The study of vegetation. Junk, The Hague, The Netherlands.
Wartenberg, D., S. Ferson, and F. J. Rohlf. 1987. Putting things in order: a critique of detrended correspondence analysis. American Naturalist **129**:434–448.
Williamson, M. H. 1978. The ordination of incidence data. Journal of Ecology **66**:911–920.
Wishart, D. 1969. An algorithm for hierarchical classifications. Biometrics **25**:165–170.

APPENDIX A

The *Leptograpsus* data plus their ordination and the principal coordinates analysis results for the long gradient data on axes 1–3 and 1–4 are available in ESA's Electronic Data Archive: *Ecological Archives* E083-062-A1.

APPENDIX B

The presence/absence data matrix from the Sashegy Nature Reserve (Budapest, Hungary) may be downloaded as an ASCII text file from ESA's Electronic Data Archive: *Ecological Archives* E083-062-A2.

APPENDIX C

A proof showing how and why Lissajous curves are obtained from a cyclic gradient is available in ESA's Electronic Data Archive: *Ecological Archives* E083-062-A3.

APPENDIX D

A proof showing why polynomial-shaped curves may arise from a linear gradient is available in ESA's Electronic Data Archive: *Ecological Archives* E083-062-A4.