Oecologia (2001) 129:271–280
DOI 10.1007/s004420100716

Pierre Legendre · Eugene D. Gallagher

Ecologically meaningful transformations for ordination of species data

Received: 25 September 2000 / Accepted: 17 March 2001 / Published online: 11 July 2001

Abstract This paper examines how to obtain species biplots in unconstrained or constrained ordination without resorting to the Euclidean distance [used in principal-component analysis (PCA) and redundancy analysis (RDA)] or the chi-square distance [preserved in correspondence analysis (CA) and canonical correspondence analysis (CCA)] which are not always appropriate for the analysis of community composition data. To achieve this goal, transformations are proposed for species data tables. They allow ecologists to use ordination methods such as PCA and RDA, which are Euclidean-based, for the analysis of community data, while circumventing the problems associated with the Euclidean distance, and avoiding CA and CCA which present problems of their own in some cases. This allows the use of the original (transformed) species data in RDA carried out to test for relationships with explanatory variables (i.e. environmental variables, or factors of a multifactorial analysis-of-variance model); ecologists can then draw biplots displaying the relationships of the species to the explanatory variables. Another application allows the use of species data in other methods of multivariate data analysis which optimize a least-squares loss function; an example is *K*-means partitioning.

Keywords Biplot diagram · Canonical correspondence analysis · Correspondence analysis · Principal-component analysis · Redundancy analysis

Introduction

Correspondence analysis (CA) and canonical correspondence analysis (CCA) are widely used to obtain unconstrained or constrained ordinations of species abundance data tables and the corresponding biplots or triplots which are extremely useful for ecological interpretation (Fig. 1a, c). Empirical work during the 1970s established that CA was appropriate for such data, while ter Braak (1985) showed that the chi-square distance preserved in CA provided a good approximation for species with unimodal distributions along a single environmental gradient. There is a problem with this metric, however: a difference between abundance values for a common species contributes less to the distance than the same difference for a rare species, so that rare species may have an unduly large influence on the analysis (Greig-Smith 1983; ter Braak and Smilauer 1998; Legendre and Legendre 1998). To avoid this, users of CA and CCA may remove the rarest species from the analysis, or resort to empirical methods giving small weights to rare species, as found for instance in the ordination program Canoco (ter Braak and Smilauer 1998). The chi-square distance is not unanimously accepted among ecologists: using simulated data, Faith et al. (1987) concluded that it was one of the worst distances for community composition data.

Alternatives to CA and CCA are principal-component analysis (PCA, for unconstrained ordination) and redundancy analysis (RDA, for constrained ordination) (Fig. 1a, c). In the full-dimensional space, these methods preserve the Euclidean distances among sites. For the analysis of sites representing short gradients, PCA and RDA may be suitable. For longer gradients, many species are replaced by others along the gradient and this generates many zeros in the species data table. Community ecologists have repeatedly argued that the Euclidean distance (and thus PCA and RDA) is inappropriate for raw species abundance data involving null abundances (e.g. Orlóci 1978; Wolda 1981; Legendre and Legendre 1998; Table 1). For that reason, CCA is often the method favoured by researchers who are analysing compositional data, despite the problem posed by rare species.

Other alternatives are available for unconstrained ordination analysis. One may compute a resemblance matrix (similarity or distance) among sites using any of a

P. Legendre (✉)
Département de sciences biologiques,
Université de Montréal, C.P. 6128, succursale Centre-ville,
Montréal, Québec, H3C 3J7, Canada
e-mail: pierre.legendre@umontreal.ca
Fax: +1-514-3432293

E.D. Gallagher
Department of Environmental, Coastal & Ocean Sciences,
University of Massachusetts at Boston, Boston, MA 02125, USA

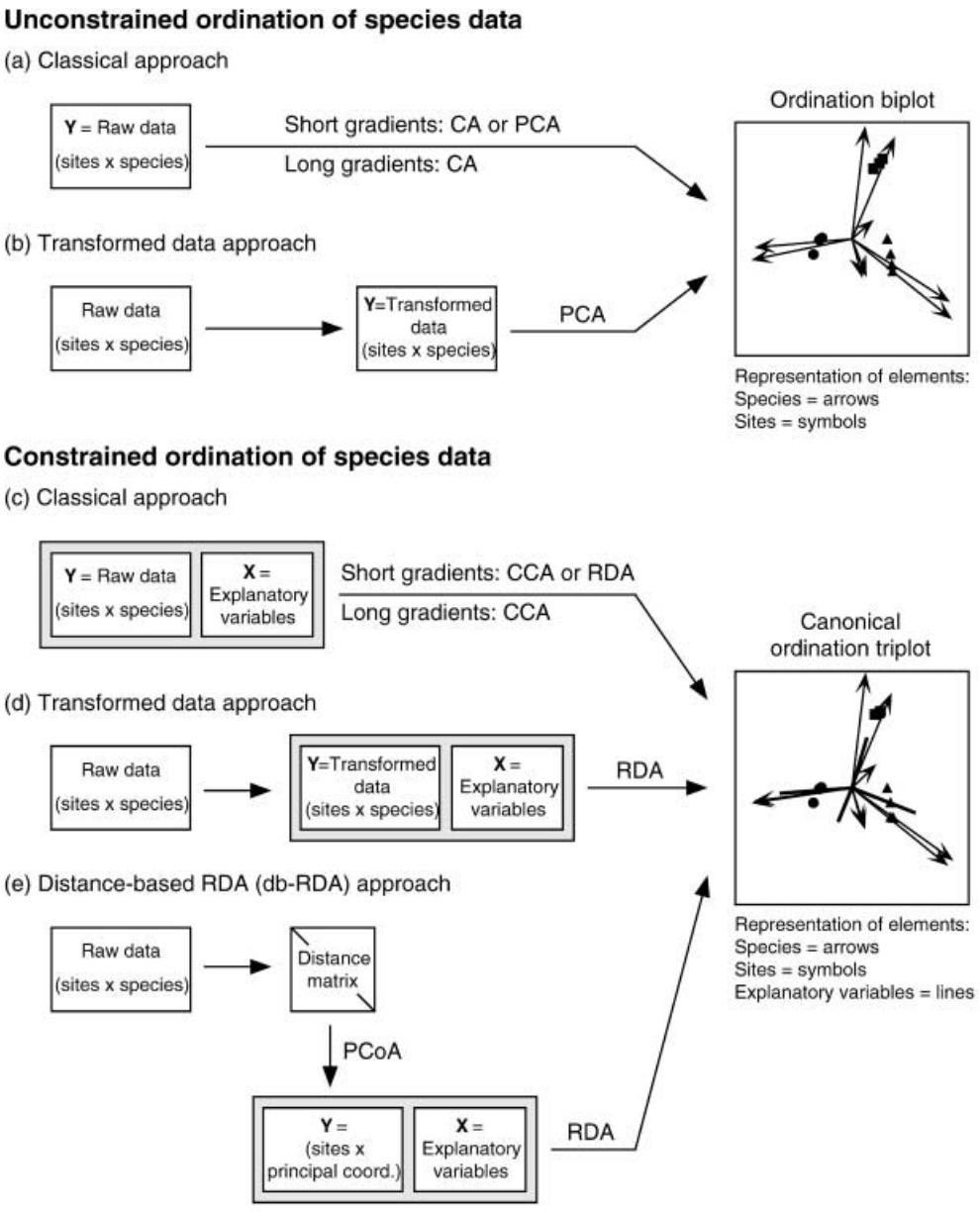

Fig. 1 Schematic comparison of techniques that can be used to obtain unconstrained (**a,b**) or constrained (**c–e**) ordination biplots or triplots of species data tables

Table 1 Species abundance paradox, modified from Orlóci (1978). The paradox is that the Euclidean distance between sites 1 and 2, which have no species in common, is smaller than that between sites 1 and 3 which share species 2 and 3. This example shows that the Euclidean distance is not appropriate for species community composition data containing zeros. With the other coefficients used here, the distance between sites 1 and 2 is larger than between sites 1 and 3, and the distance between sites 1 and 2 is the same as between sites 2 and 3, or very nearly so

		Species 1	Species 2	Species 3
Species abundance paradox data (three sites, three species)	Site 1	0	1	1
	Site 2	1	0	0
	Site 3	0	4	8
Distance function		*D*(site 1, site 2)	*D*(site 1, site 3)	*D*(site 2, site 3)
$D_{\text{Euclidean}}$		1.7321	7.6158	9.0000
D_{chord}		1.4142	0.3204	1.4142
$D\chi^2_{\text{metric}}$		1.0382	0.0930	1.0352
$D\chi^2_{\text{distance}}$		4.0208	0.3600	4.0092
$D_{\text{species profiles}}$		1.2247	0.2357	1.2472
$D_{\text{Hellinger}}$		1.4142	0.1697	1.4142

number of resemblance coefficients that are appropriate for species presence-absence or abundance data (see Legendre and Legendre 1998 for a review). Following this, principal-coordinate analysis (PCoA) or non-metric multidimensional scaling (NMDS) can be used to obtain an ordination in a small number of dimensions, usually two or three. To obtain biplots of species and sites from PCoA or NMDS, one can (1) compute correlations between the original species vectors (i.e. the vectors whose *i*th components are the counts of a species at site *i*) and the site scores along the PCoA or NMDS ordination axes and scale these correlations as described in Eq. 14 (below), or (2) use the site scores along the two or three PCoA or NMDS ordination axes retained for ordination, together with the original species data, and carry out a PCA of this larger data table in which the original species will be treated as supplementary variables having weights of zero in the analysis; the use of supplementary variables in PCA is described, for example, in ter Braak and Smilauer (1998) and Legendre and Legendre (1998).

Resemblance matrices cannot be used directly in canonical ordination, however. Legendre and Anderson (1999) have proposed a solution to this problem, called distance-based redundancy analysis (db-RDA; Fig. 1e): (1) compute a matrix of distances D_{ij} among sites using a measure appropriate to species data, e.g. the Steinhaus/Odum/Bray-Curtis measure (called Bray-Curtis for simplicity) which is often the preferred choice of ecologists; (2) compute all the principal coordinates, using PCoA; they preserve the original distances D_{ij} in full ordination space; if negative eigenvalues and corresponding complex-number axes are produced during eigenvalue decomposition, a correction can be applied to the distance matrix to eliminate them; (3) use RDA to analyse the relationship between the principal coordinates, representing the species data, and the explanatory variables. db-RDA is well-suited to test the significance of relationships between the explanatory and response data tables, but not to produce biplots or triplots of species, sites and environmental variables, which may be needed for interpretation (Gabriel 1982; ter Braak 1994), because the species matrix is replaced in step 3 by another matrix whose columns are principal coordinates. Each column now represents a non-linear combination of the original species, so that their roles cannot easily be untangled. The db-RDA approach can be used either for regular redundancy analysis of community composition against environmental variables, or to obtain a Manova-like analysis in which the factors of the Manova are coded in the matrices of environmental variables and covariables of the canonical analysis. See Legendre and Anderson (1999) for details.

The present paper describes transformations of the species data that allow ecologists to use ordination methods such as PCA and RDA, which are Euclidean-based, with community composition data containing many zeros (long gradients). These transformations offer alternatives, for ordination analysis of community data, to CA and CCA, which are based upon the chi-square metric (Fig. 1b). They allow the use of the original (transformed) species data in RDA to test the relationships with explanatory variables (i.e. environmental variables or factors of a multifactorial analysis-of-variance model; Fig. 1d), as an alternative to db-RDA (Fig. 1e), thus allowing one to draw biplots displaying the relationships of the species to the explanatory variables. An additional application allows the use of community composition data in other methods of multivariate analysis which optimize a least-squares loss function. An example is *K*-means partitioning which separates the objects (e.g. sampling sites) into groups obtained by minimizing the sum of the squared Euclidean distances of the objects to the group centroids.

Transform species composition data to obtain targeted (dis)similarity coefficients

Some (dis)similarity measures commonly used by community ecologists can be obtained by first modifying the species data, then computing the Euclidean distance among sites on the modified data set. We are proposing here to transform the species presence-absence or abundance data and use the transformed data in PCA, RDA or *K*-means partitioning. The net result is an analysis that will preserve the chosen distances among objects. Not all similarity coefficients that have been proposed to analyse community structure data can be obtained through such a transformation, however (see Discussion).

Consider a species abundance data table $Y=[y_{ij}]$ of size ($n \times p$) with sites (rows) $i=\{1...n\}$ and species (columns) $j=\{1...p\}$; the row sums are noted y_{i+} and the column sums y_{+j}; the overall sum is y_{++}. We will define transformations of the data $Y \rightarrow Y'$| such that the Euclidean distance (Eq. 4) among the rows of the transformed data table is equal to some other distance computed among the rows of the original data table Y' (Fig. 2). In the remainder of this paper, we will consider only distance coefficients; if required by the software, the corresponding similarities can be obtained by $S(i, j)=1-D(i, j)$ or $S(i, j)=1-D^2(i, j)$, after ranging the distances to the interval [0, 1] if necessary.

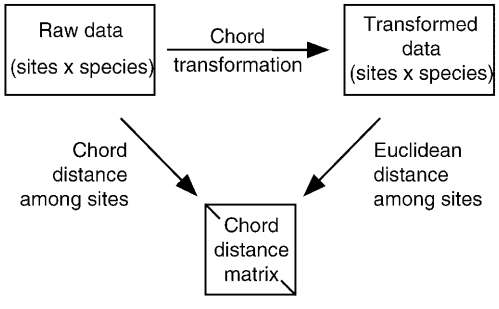

Fig. 2 Illustration of the role of the data transformations as a way of obtaining a given distance function. The example uses the chord distance

The transformations described below are precursors of distances that have all been described as appropriate for community composition data. They allow users to retain the identity of the individual species in PCA or RDA biplots.

(1) Chord distance

The chord distance, proposed by Orlóci (1967) and Cavalli-Sforza and Edwards (1967), is the Euclidean distance computed after scaling the site vectors to length 1, i.e. dividing each value by the norm, or length, of the vector. After normalization, the Euclidean distance between two objects (sites) is equivalent to the length of a chord joining two points within a segment of a hypersphere of radius 1. The formula for the chord distance between sites x_1 and x_2 across the p species is thus:

$$D_{chord}(x_1,x_2) = \sqrt{\sum_{j=1}^{p}\left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{2j}^2}}\right)^2} \quad (1)$$

The chord distance may also be computed using the following formula found in several textbooks and papers (e.g. Orlóci 1967):

$$D_{chord}(x_1,x_2) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} y_{1j}y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2 \sum_{j=1}^{p} y_{2j}^2}}\right)} \quad (2)$$

It is clear from Eq. 1 that if the data $[y_{ij}]$ are first transformed into $[y'_{ij}]$ as follows:

$$y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}} \quad (3)$$

then the Euclidean distance

$$D_{Euclidean}(x'_1,x'_2) = \sqrt{\sum_{j=1}^{p}\left(y'_{1j} - y'_{2j}\right)^2} \quad (4)$$

between row vectors of transformed data is identical to the chord distance between the original row vectors of species abundances. The inner part of Eq. 2 is actually the cosine of the angle (θ) between the two site vectors, normalized or not; this is easily derived from the scalar product of two vectors: $b \cdot c$=(length of b)×(length of c)×cos θ. So the chord distance may be written as:

$$D_{chord} = \sqrt{2(1-\cos\theta)} \quad (5)$$

This distance is maximum when the two sites have no species in common; the normalized site vectors may then be represented by points at 90° from each other on the circumference of a sector of a circle (for two species) or the surface of a segment of a hypersphere (for p species) and the distance between the two sites is $\sqrt{2}$. Trueblood et al. (1994) used a form of Eq. 3 in their PCA-H method, with y_{ij} being the probability of sampling species j in sample i with a random draw of m individuals from the sample. They called the Euclidean distances between these transformed row vectors CNESS, the chord normalized expected species shared. CNESS is a metric version of Grassle and Smith's (1976) NESS similarity index. Orlóci's chord distance equals CNESS when m=1.

(2) Chi-square metric and chi-square distance

The chi-square metric is often used for clustering or ordination of species abundance data. Although this measure has no upper limit, it produces distances smaller than 1 in most cases. The formula is:

$$D_{\chi^2 metric}(x_1,x_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}}\left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}}\right)^2} \quad (6)$$

The inner part is the Euclidean distance computed on relative abundances, weighted by the inverse of the column (species) sums y_{+j}. If a species j is rare, its column sum y_{+j} is small and this species contributes a great deal to the sum of squares. If the data $[y_{ij}]$ are transformed into $[y'_{ij}]$ as follows:

$$y'_{ij} = \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \quad (7)$$

then the Euclidean distance (Eq. 4) between row vectors of transformed data is identical to the chi-square metric (Eq. 6) between the original row vectors of species abundances.

The chi-square distance is the chi-square metric multiplied by the square root of the sum of abundances in the data table, $\sqrt{y_{++}}$. This distance is particularly important in numerical ecology because it is the distance preserved in CA and CCA. These two distances are treated together because they only differ by a multiplicative constant. The formula for the chi-square distance is:

$$\begin{aligned} D_{\chi^2 distance}(x_1,x_2) &= \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}/y_{++}}\left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}}\right)^2} \\ &= \sqrt{y_{++}}\sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}}\left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}}\right)^2} \end{aligned} \quad (8)$$

If the data $[y_{ij}]$ are transformed into $[y'_{ij}]$ as follows:

$$y'_{ij} = \sqrt{y_{++}}\frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \quad (9)$$

a little algebra shows that the Euclidean distance (Eq. 4) between row vectors of transformed data is identical to the chi-square distance (Eq. 8) between the original row vectors of species abundances. This is not to say that a

PCA of data transformed as in Eq. 9, or a PCoA of a matrix of chi-square distances, will exactly reproduce a CA ordination; the rotation of the data point distribution in CA does not maximize inertia in the same way as in PCA or PCoA.

The transformation proposed here (Eq. 9) had been described by Chardy et al. (1976) and used by Pinel-Alloul et al. (1995) to prepare matrices of phytoplankton (87 taxa) and fish (18 taxa) data, prior to using them as matrices of explanatory variables in a CCA where the dependent matrix was a table of zooplankton abundances (54 taxa).

(3) Distance between species profiles

A variant of the chi-square metric can be obtained by removing the standardization by the inverse of y_{+j}. The species data are simply transformed into profiles of relative frequencies before computing Euclidean distances. This equation does not give extra weight to the rare species; the most abundant species contribute predominantly to the sum of squares. The formula is:

$$D_{species\ profiles}(x_1, x_2) = \sqrt{\sum_{j=i}^{p} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} \qquad (10)$$

This formula is constructed in the same way as Eq. 1; it is the Euclidean distance (Eq. 4) between species profiles. If the data $[y_{ij}]$ are first transformed into $[y'_{ij}]$as follows:

$$y'_{ij} = \frac{y_{ij}}{y_{i+}} \qquad (11)$$

then the Euclidean distance (Eq. 4) between rows of transformed data (which are also called "compositional data") is identical to the distance between species profiles computed on the original species abundance data (Eq. 10).

(4) Hellinger distance

The Hellinger distance is also a measure recommended for clustering or ordination of species abundance data (Rao 1995). The formula is:

$$D_{Hellinger}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2} \qquad (12)$$

Rao (1995) recommends it as a basis for a new ordination method. Using simulations, Legendre and Legendre (1998) concluded that for linear ordination, the Hellinger distance offers a better compromise between linearity and resolution than the chi-square metric and the chi-square distance. If the data $[y_{ij}]$ are first transformed into $[y'_{ij}]$ as follows:

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}} \qquad (13)$$

then the Euclidean distance (Eq. 4) between row vectors of transformed data is identical to the Hellinger distance between the original row vectors of species abundances.

The transformations described in this section form a set in the sense that the corresponding distances have all been recommended for analysis of community composition data and can all be obtained by transforming the species abundance data followed by computation of Euclidean distances between the rows of transformed data. Other coefficients, described for instance in Legendre and Legendre (1998), that are appropriate for community composition analysis cannot be obtained using simple transformations of the species abundance data.

Example

To illustrate the differences among ordination methods, an artificial ecological gradient was created by generating abundances for nine species at 19 sites along a transect (Fig. 3a). Species 2 to 4, represented by 36 individuals each, replace each other along the gradient. Species 1 and 5 have the same kind of distribution and appear at the ends of the transect. Rare species 6–9 occur in narrow ranges of conditions along the gradient; they are represented by 2–5 individuals.

Distance functions that are suitable for the analysis of community composition data should, minimally, be able to produce reasonable reconstructions of such a simple gradient. This can be assessed by looking either at the distance matrices themselves (Fig. 3) or at the biplots (Fig. 4).

When examining distance matrices, one expects distances to increase monotonically as sites get further apart, until a maximum is reached for sites that have no species in common. This can be displayed using graphs of the computed ecological distances (ordinate) against the true geographic distances along the transect (abscissa). A model of the relationship can be drawn by joining the mean ecological distances computed for each geographic distance. If sampling has taken place on a geographic surface instead of a transect, the geographic distances will not fall into a small number of discrete distances; a smooth function can be drawn using moving averages, splines, or LOWESS smoothing. We call this type of graph a *diastemogram*, from the Greek διαστεμα (diastema) distance, and γραμμα (gramma) drawing.

In Fig. 3, the diastemogram function is monotonically increasing for the chord, chi-square, Hellinger and Bray-Curtis distances, as sites get farther away along the gradient. Even though the Bray-Curtis distance cannot be obtained using one of the transformations of the previous section, it has been included in Fig. 3 for comparison and reference because of its wide use in community ecology. The diastemogram function is not monotonic for the Euclidean distance and the distance between species profiles. On the other hand, the coefficient of determination (R^2) measures how much of the variance of the

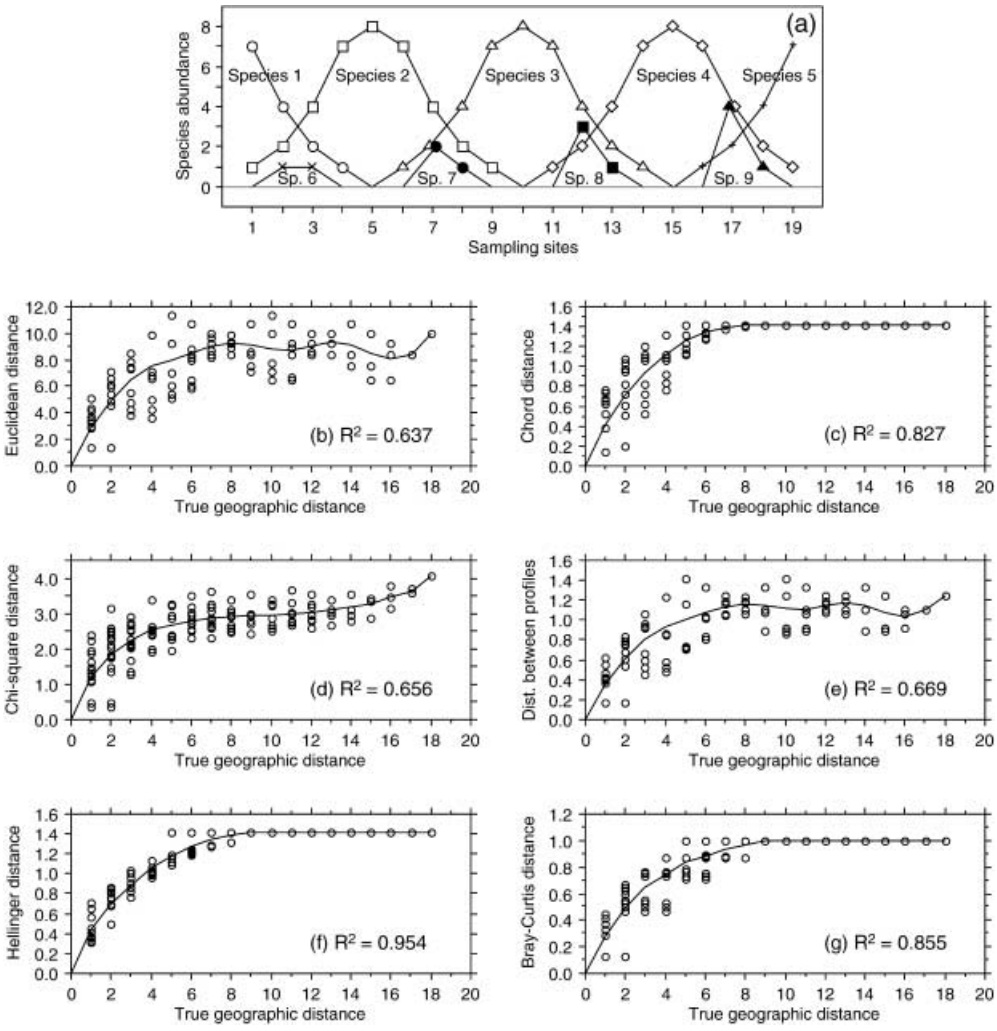

Fig. 3a–g Analysis of artificial gradient data. **a** The gradient comprises 19 sites (numbers along abscissa) and nine species (different symbols). **b–g** Diastemograms comparing true geographic distances (abscissa) to the computed ecological distances among sites (ordinate). The construction and interpretation of these graphs is described in the text

ecological distance matrix is explained by the diastemogram function; the value of R^2 is low for the Euclidean distance and the distance between species profiles. These are indications that these two distances should not be used to represent at least this particular ecological gradient. Note, however, that PCA based on species profiles can be interpreted in terms of alpha and beta diversity (ter Braak 1983).

In our example, R^2 is highest for the Hellinger distance, followed by the Bray-Curtis and chord distances; the chi-square distance, which does not reach an asymptote, comes last by that criterion. The diastemogram function is monotonic for these three distances. So, the best choices for this example seem to be the Hellinger and chord distances, for which ordinations can be obtained through the simple transformations described in the previous section followed by PCA or RDA (Fig. 1b, d), or the Bray-Curtis distance for which ordination diagrams can be obtained by PCoA; biplots of species and sites are more difficult to obtain, however, in that case.

Another criterion for the comparison of distance measures is the importance given to rare species in the analysis. If rare species are well sampled and truly rare, they may be used as indicators of the conditions that may exist at some sites only. In that case, they should receive high weight, as they do in CA and, to a lesser extent, in PCA after the chi-square transformation (Table 2). In contrast, when rare species are observed sporadically at some sites but, as the result of "sampling error", not at others where they are also present, giving them high weight in the analysis is unwise. This phenomenon is exacerbated in environments where sampling is conducted blindly – for example, in aquatic and soil ecology. Coefficients such as the Euclidean, chord, species profile or Hellinger distances do not give high weights to the rare species (Table 2).

The combined effect of the various types of transformation and the most commonly used ordination methods was assessed by carrying out ordination of the data, first by PCA and CA, then by PCA after transforming the data as described above. We observe the following:

- After PCA of the raw data (Fig. 4a), axis 1 displays the gradient with strong inward folding of the sites at the ends of the gradient (horseshoe effect). Because the Euclidean distance function considers double ze-

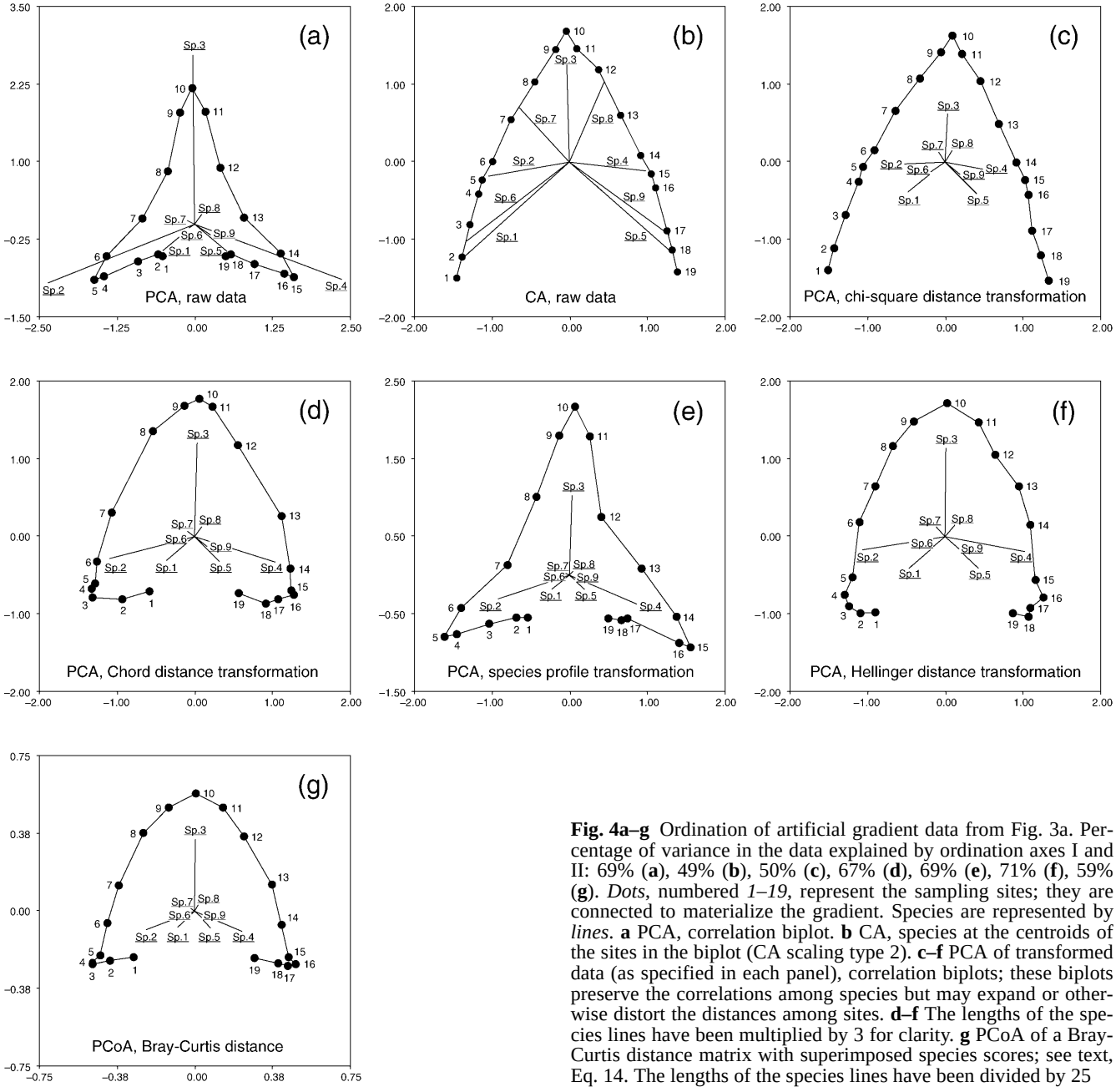

Fig. 4a–g Ordination of artificial gradient data from Fig. 3a. Percentage of variance in the data explained by ordination axes I and II: 69% (**a**), 49% (**b**), 50% (**c**), 67% (**d**), 69% (**e**), 71% (**f**), 59% (**g**). *Dots*, numbered *1–19*, represent the sampling sites; they are connected to materialize the gradient. Species are represented by *lines*. **a** PCA, correlation biplot. **b** CA, species at the centroids of the sites in the biplot (CA scaling type 2). **c–f** PCA of transformed data (as specified in each panel), correlation biplots; these biplots preserve the correlations among species but may expand or otherwise distort the distances among sites. **d–f** The lengths of the species lines have been multiplied by 3 for clarity. **g** PCoA of a Bray-Curtis distance matrix with superimposed species scores; see text, Eq. 14. The lengths of the species lines have been divided by 25

Table 2 Fraction of the total variance occupied by each species, either for the raw species data from Fig. 3a (in the case of PCA and CA), or after the stated transformation (tr.) of the data. Each row sums to 1. The variance of a species vector measures its relative importance in the analysis. Species (Sp.) 6–9 are the rare species

	Sp. 1	Sp. 2	Sp. 3	Sp. 4	Sp. 5	Sp. 6	Sp. 7	Sp. 8	Sp. 9
PCA (original data)	0.1070	0.2434	0.2434	0.2434	0.1070	0.0032	0.0081	0.0164	0.0281
CA (original data)[a]	0.1060	0.0355	0.0349	0.0344	0.1037	0.1725	0.1725	0.1725	0.1679
Chord tr. (Eq. 3)	0.1248	0.2303	0.2245	0.2192	0.1208	0.0065	0.0148	0.0249	0.0343
Chi-square tr. (Eq. 9)	0.1686	0.1410	0.1393	0.1382	0.1644	0.0428	0.0584	0.0700	0.0773
Profile tr. (Eq. 11)	0.1143	0.2458	0.2427	0.2409	0.1114	0.0041	0.0085	0.0136	0.0187
Hellinger tr. (Eq. 13)	0.1216	0.2161	0.2126	0.2100	0.1166	0.0206	0.0284	0.0346	0.0395

[a] ter Braak and Smilauer (1998, Eq. 6.36)

ros as an indication of similarity, PCA brings together sites from the two ends of the gradient that have no species in common. PCA after the transformation into species profiles (Eq. 11) produces similar results (Fig. 4e). The inadequacy of ordinations displaying strong horseshoes has been discussed in the ecological literature since Goodall (1954).

- CA displays the gradient correctly along axis 1 (Fig. 4b); so does PCA on data transformed by Eq. 9, where the ordination preserves the chi-square distance among sites (Fig. 4c); Eq. 7 leads to similar results (not shown). The small differences in the ordination of sites between Figs. 4b and c is due to the fact that CA does not define inertia in the same way and, so, does not apportion it among axes in the same way as PCA.
- PCAs on data transformed using Eqs. 3 (chord transformation; Fig. 4d) and 13 (Hellinger transformation; Fig. 4f) produce good representations of the gradient along axis 1, with some inward folding of three sites at both ends of the transect.
- PCoA of a Bray-Curtis distance matrix is shown in Fig. 4g for comparison. Linear correlations were computed between the original species vectors and the first two principal coordinates. The correlations were weighted as follows to obtain the species scores for the biplot:

$$SpeciesScore_{jk} = r_{jk} s_j / s_k \qquad (14)$$

where r_{jk} is the correlation between species j and site score vector k, s_j is the standard deviation of species j, and s_k is the standard deviation of site score vector k. The term s_k adjusts the species scores to the scaling used in any particular analysis: the variance of site score vector k is λ_k in PCoA and in PCA distance biplots (λ_k=eigenvalue), whereas it is 1 in PCA correlation biplots.

For the present example, the ordinations of sites obtained for the chord and Hellinger transformations have inward folding of three sites at both ends of the transect (horseshoe effect); the horseshoe is not stronger than in the ordination obtained using the Bray-Curtis distance (Fig. 4). The horseshoe effect is much stronger in the case of the Euclidean distance and the species profile transformation. CA as well as PCA after chi-square distance transformation produced no horseshoe effect in this example. Noticeably, the fraction of variance of the species data accounted for by the first two ordination axes is much lower in CA and in the PCA following chi-square distance transformation (49–50%) than in the other PCAs (67–71%). The Bray-Curtis PCoA ordination is intermediate in that respect (57%).

In the biplots, where only the first two axes were used, all methods based upon PCA gave a fair representation of the relative numerical importance of the rare species. This included the PCA after chi-square distance transformation of the data: even though the rare species have high weights in the analysis (higher variance in Table 2 than in the other PCA results), they only load heavily on the lesser ordination axes. In CA, where the rare species have high weights (Table 2), the importance of a species for an ordination subspace is given, e.g. in the program Canoco, by the "cumulative fit per species as fraction of variance of species". Here again, the rare species only load heavily on the last ordination axes. Fig. 4b (CA) shows the species at the centroids of the sites where they are present; compare Figs. 3a and 4b. The lengths of the lines joining the species to the origin are not a measure of their importance in the analysis; in CA biplots, species scores are slopes with respect to the axes (ter Braak and Smilauer 1998, Eq. 6.17).

Discussion

The transformations described above are precursors of distances that have all been described as appropriate for community composition data. This is not to say that these five are the only appropriate distances for species abundance data; see for example Faith et al. (1987) or Legendre and Legendre (1998) for reviews of appropriate coefficients. These distances, however, can be obtained through transformations that allow users to retain the identity of the individual species in biplots. Prior to computing these transformations, any of the standardizations investigated by Faith et al. (1987) may also be used.

Theoretical considerations

The distance functions that can be obtained by transformation of the species data followed by calculation of Euclidean distances have some interesting mathematical properties:

1. All distance functions pertaining to this group are Euclidean, meaning that the distances among objects that they produce can be entirely represented in Euclidean space. A representation is Euclidean when PCoA of the distance matrix does not produce negative eigenvalues. These concepts are explained, for example, by Legendre and Legendre (1998).
2. The distances corresponding to the transformations described above can be computed on presence-absence species data. Hence the transformations can also be applied to this type of data.
3. Coefficients that are non-Euclidean cannot be obtained through transformation of the data followed by calculation of Euclidean distances. If it were possible to do so, the resulting distances would be Euclidean, while they are not. In particular, distances which are one-complements of similarity coefficients for binary (0–1) data, such as the simple matching, Jaccard or Sørensen coefficients, cannot be obtained through such transformations since none of them are Euclidean (Gower and Legendre 1986; Legendre and Legendre 1998, Table 7.2); Sørensen's coefficient is not even a metric. The widely used Bray-Curtis coefficient for community data, which ranked among the best of the coefficients studied by Faith et al. (1987), cannot be obtained through such a transformation because it is

neither Euclidean nor metric; likewise for the NESS similarity coefficient of Grassle and Smith (1976). To use these otherwise excellent coefficients in constrained ordination, one should use either the db-RDA approach of Legendre and Anderson (1999) or the alternative computation procedure proposed by McArdle and Anderson (2001).

4. The coefficients that can be obtained by data transformation are those that can be expressed as an equation where each value y_{ij} is weighted by some function of the values in the same row and/or column (e.g. the sum, or the sum of squares, of the values). The sum of all values y_{++} may also be included in the transformation. As a result, these coefficients only compare species profiles, normalized in various ways. To obtain distances that preserve total abundances at the sites, instead of normalized profiles, one has to use the coefficient of Bray-Curtis or that of Kulczynski; see Faith et al. (1987) or Legendre and Legendre (1998) for details.

Practical considerations

For the analysis of community gradients, it does not matter that an analysis gives high weights to the rare species when the end-product is simply a reduced-space ordination diagram. In CA or CCA, and in PCA or RDA based upon the chi-square metric (Eq. 7) or the chi-square distance transformation (Eq. 9), the rarest species are well fitted by the axes with the smallest eigenvalues. The contributions of these species to the first few axes, used for reduced-space ordination, are small. The weights given to rare species do matter, however, when the end-product is a test of significance of the relationship between species composition and a set of explanatory variables, or a test of factors in a multiple analysis-of-variance model following the db-RDA approach of Legendre and Anderson (1999). CCA, or the chi-square distance transformation followed by RDA, should not be used unless one specifically wants to give high weight to the rare species that may indicate the presence of particular environmental conditions.

The chord and Hellinger transformations are appropriate alternatives giving low weights to rare species. In our example, the diastemogram functions for these two transformations showed that the resulting distances were monotonically related to the geographic distances along the gradient, as in the case of the Bray-Curtis coefficient; they reached an asymptote for sites that had no species in common; they produced little horseshoe effect in ordinations; and they allowed the representation of species and sites in biplots.

For simple ordination analysis, the difference between CA and PCA after the chi-square distance transformation is small. With appropriate scalings of the species and site scores, the same distance is preserved in the two forms of analysis, but the eigenvalues may differ slightly. There is a more important difference in canonical ordination, however: the multiple regressions of the species on the set of explanatory variables are done with weights in CCA; this is not the case in RDA. The weights in CCA are given by a diagonal matrix containing the square roots of the row sums of the species data table. This means that a site where many individuals have been observed contributes more to the regression than a site with few individuals. CCA should only be used when the sites have approximately the same number of individuals, or when one explicitly wants to give high weight to the richest sites. This problem of CCA was one of our incentives for looking for alternative methods for canonical ordination of community composition data. RDA based upon the transformations proposed in this paper (including the chi-square metric or distance transformations) offers an alternative solution.

PCA of raw data, which preserves the Euclidean distance among sites in full-dimensional space, is inappropriate for species composition data. PCA of the species profiles allows a quick view of both alpha diversity (as Simpson's diversity) and beta diversity (at the cost of severe distortion at the ends of long gradients).

Applications

Here are some cases where the proposed data transformations may be useful for the analysis of species abundance tables, or other types of frequency data:

- In unconstrained or canonical ordination, when one does not wish to use the chi-square distance preserved by CA and CCA because of the differential weighting of rare species.
- In canonical ordination, when one does not wish to use the CCA weighting system for sites.
- When RDA or CCA is used to depict biotic control through top-down or bottom-up interactions (Lindeman 1942; Southwood 1987). In such studies, the analysis of a species matrix Y is constrained using another species matrix X representing predators or prey. The species data in X need to be transformed in such a way that the Euclidean distances among rows of X correspond to meaningful distances in species space, before X is used in a linear model; the transformations proposed in this paper can be used to do this. See Pinel-Alloul et al. (1995) for an example.
- The transformations described in this paper offer new ways of partitioning sites described by species abundance data, i.e. dividing them into groups. One commonly used partitioning method is *K*-means, which is a Euclidean method minimizing a least-squares loss function. To preserve a distance function which is appropriate for community composition data, instead of the Euclidean distance which is inappropriate (Table 1), one can, prior to *K*-means, transform the species abundances using Eqs. 3, 7, 9, 11 or 13.

Theoretical criteria are not known at the moment that would allow one to select the best distance function (or

data transformation) for any specific situation. In canonical analysis, one may empirically select the transformation which leads to the highest fraction of explained variation. Computer programs to carry out these transformations are available on the WWW sites http://www.fas.umontreal.ca/biol/legendre/ (Fortran source code and compiled versions) and http://www.es.umb.edu/edgwebp.htm (Matlab code).

Acknowledgements Thanks to Elaine Hooper who motivated this investigation by asking how to represent the original species in a biplot after db-RDA; to David W. Roberts who proposed improvements to the example data set used in this paper and to the ways of showing differences among distance functions; to Philippe Casgrain who incorporated diastemograms into The R Package; and to Daniel Borcard and Mark Burgman for comments on the manuscript. This research was supported by NSERC grant number OGP7738 to P. Legendre.

References

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis: models and estimation procedures. Evolution 21:550–570

Chardy P, Glemarec M, Laurec A (1976) Application of inertia methods to benthic marine ecology: practical implications of the basic options. Estuarine Coastal Shelf Sci 4:179–205

Faith DP, Minchin PR, Belbin L (1987) Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69: 57–68

Gabriel KR (1982) Biplot. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, vol 1. Wiley, New York, pp 263–271

Goodall DW (1954) Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. Aust J Bot 2:304–324

Gower JC, Legendre P (1986) Metric and Euclidean properties of dissimilarity coefficients. J Classif 3:5–48

Grassle JF, Smith W (1976) A similarity measure sensitive to the contribution of rare species and its use in investigation of variation in marine benthic communities. Oecologia 25:13–25

Greig-Smith P (1983) Quantitative plant ecology, 3rd edn. Blackwell, London

Legendre P, Anderson MJ (1999) Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecol Monogr 69:1–24

Legendre P, Legendre L (1998) Numerical ecology, 2nd English edn. Elsevier, Amsterdam

Lindeman RL (1942) The trophic-dynamic aspect of ecology. Ecology 23:399–418

McArdle BH, Anderson MJ (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology 82:290–297

Orlóci L (1967) An agglomerative method for classification of plant communities. J Ecol 55:193–205

Orlóci L (1978) Multivariate analysis in vegetation research, 2nd edn. Junk, The Hague

Pinel-Alloul B, Niyonsenga T, Legendre P (1995) Spatial and environmental components of freshwater zooplankton structure. Écoscience 2:1–19

Rao CR (1995) A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Qüestiió 19:23–63

Southwood TRE (1987) The concept and nature of the community. In: Gee JHR, Giller PS (eds) Organization of communities: past and present. Blackwell, Oxford, pp 3–27

ter Braak CJF (1983) Principal components biplots and alpha and beta diversity. Ecology 64:454–462

ter Braak CJF (1985) Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics 41:859–873

ter Braak CJF (1994) Canonical community ordination. I. Basic theory and linear methods. Écoscience 1:127–140

ter Braak CJF, Smilauer P (1998) CANOCO reference manual and user's guide to Canoco for Windows – software for canonical community ordination (version 4). Microcomputer Power, Ithaca, NY

Trueblood DD, Gallagher ED, Gould DM (1994) Three stages of seasonal succession on the Savin Hill Cove mudflat, Boston Harbor. Limnol Oceanogr 39:1440–1454

Wolda H (1981) Similarity indices, sample size and diversity. Oecologia 50:296–302

Species abundance paradox data
(3 sites, 3 species)

	Species 1	Species 2	Species 3
Site 1	0	1	1
Site 2	1	0	0
Site 3	0	4	8

	Transformations	
$D_{\text{Euclidean}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$		$\mathbf{D} = \begin{bmatrix} 0.0000 & 1.7321 & 7.6158 \\ 1.7321 & 0.0000 & 9.0000 \\ 7.6158 & 9.0000 & 0.0000 \end{bmatrix}$
$D_{\text{chord}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{2j}^2}} \right)^2}$	$y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}}$	$\mathbf{D} = \begin{bmatrix} 0.0000 & 1.4142 & 0.3204 \\ 1.4142 & 0.0000 & 1.4142 \\ 0.3204 & 1.4142 & 0.0000 \end{bmatrix}$
$D_{\chi^2\text{metric}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$	$y'_{ij} = \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}}$	$\mathbf{D} = \begin{bmatrix} 0.0000 & 1.0382 & 0.0930 \\ 1.0382 & 0.0000 & 1.0352 \\ 0.0930 & 1.0352 & 0.0000 \end{bmatrix}$
$D_{\chi^2\text{distance}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}/y_{++}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$	$y'_{ij} = \sqrt{y_{++}} \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}}$	$\mathbf{D} = \begin{bmatrix} 0.0000 & 4.0208 & 0.3600 \\ 4.0208 & 0.0000 & 4.0092 \\ 0.3600 & 4.0092 & 0.0000 \end{bmatrix}$
$D_{\text{species profiles}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$	$y'_{ij} = \frac{y_{ij}}{y_{i+}}$	$\mathbf{D} = \begin{bmatrix} 0.0000 & 1.2247 & 0.2357 \\ 1.2247 & 0.0000 & 1.2472 \\ 0.2357 & 1.2472 & 0.0000 \end{bmatrix}$
$D_{\text{Hellinger}}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2}$	$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}}$	$\mathbf{D} = \begin{bmatrix} 0.0000 & 1.4142 & 0.1697 \\ 1.4142 & 0.0000 & 1.4142 \\ 0.1697 & 1.4142 & 0.0000 \end{bmatrix}$

Fig. 2 (alternative). Species abundance paradox data, modified from Orlóci (1978). The paradox is that the Euclidean distance between sites 1 and 2, which have no species in common, is smaller than that between sites 1 and 3 which share species 2 and 3; this example shows that the Euclidean distance is not appropriate for species abundance data. With the other coefficients in the Table, the distance between sites 1 and 2 is larger than that between sites 1 and 3; furthermore, the distance between sites 1 and 2 is the same as between sites 2 and 3, or very nearly so.