7.7 Transformations for community composition data

In communities sampled over fairly homogeneous environmental conditions, e.g. short environmental gradients, the species composition data contain few zeros, and symmetric association coefficients, including the Euclidean distance D_1, can be used for clustering or ordination. Frequency histograms of individual species may, however, display asymmetric distributions because species tend to have exponential growth when conditions are favourable. This well-known fact has been embedded in the theory of species-abundance models; see He & Legendre (1996, 2002) for a synthetic view of these models. To reduce the asymmetry of the species distributions, a species abundance variable **y** may be transformed to **y**' by taking the square root or the fourth root (equivalent to taking the square root twice), or by using a log transformation:

$$y' = y^{0.5} \quad \text{or} \quad y' = y^{0.25} \quad \text{or} \quad y' = \log(y + c) \qquad \textbf{(7.65)}$$

where y is the species abundance and c is a constant. Usually, $c = 1$ in species abundance log transformations; in this way, an abundance $y = 0$ is transformed into $y' = \log(0 + 1) = 0$ for any logarithmic base. These transformations represent the series of exponents $\gamma = 0.5, 0.25$ and 0 of the Box-Cox transformation (eq. 1.15).

Another interesting transformation that reduces the asymmetry of heavily skewed abundance data is the one proposed by Anderson *et al*. (2006). The abundance data y_{ij} are transformed as follows to a logarithmic scale that makes allowance for zeros:

$$y'_{ij} = \log_{10}(y_{ij}) + 1 \qquad \text{when } y_{ij} > 0 \qquad \textbf{(7.66)}$$

$$\text{or } y'_{ij} = 0 \qquad \text{when } y_{ij} = 0.$$

Hence, for $y_{ij} = \{0, 1, 10, 100, 1000\}$, the transformed values y'_{ij} are $\{0, 1, 2, 3, 4\}$. Note that this is *not* the $\log(y_{ij} + 1)$ transformation. This transformation is available in the ***decostand()*** function of VEGAN (method = "log") where users can choose the base of the logarithm. Changing the base of logarithms in eq. 7.65 (right) produces a linear change among the y'_{ij} values, so it does not induce any change in the relationships among the transformed values. With eq. 7.66 on the contrary, the transformations produced by different bases of logarithms are not perfectly linearly related.

Community composition data sampled over variable environmental conditions, e.g. along long environmental gradients, typically contain many zero values because species are known to generally have unimodal distributions along environmental gradients (ter Braak & Prentice, 1988) and to be absent from sites far from their optimal living conditions. The proportion of zeros is greater when the environmental conditions are more variable across the sampling sites. For association coefficients, this situation generates the double-zero problem that was discussed in Subsection 7.2.2 and leads to the selection of an asymmetrical similarity or distance coefficient for clustering or ordination.

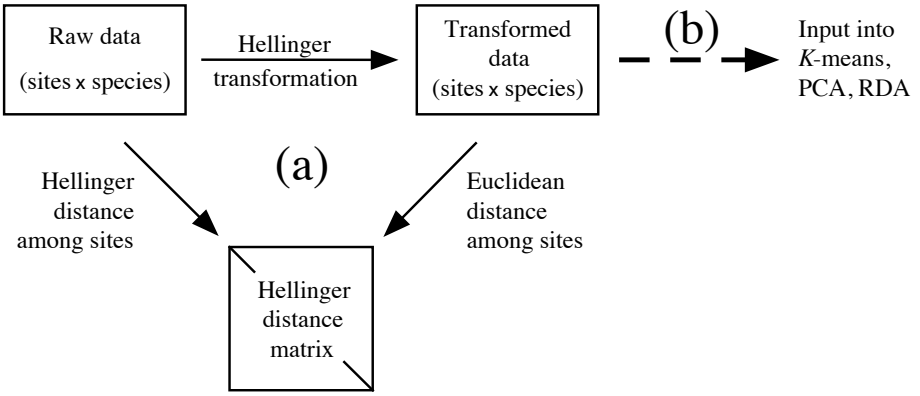

Figure 7.7 (a) Calculation of a distance matrix either directly from the raw data (left diagonal arrow) or through a two-step approach in which the raw data are transformed (horizontal arrow) before computation of the distance matrix (right diagonal arrow). The example shown here uses the Hellinger transformation to obtain the Hellinger distance matrix (D_{17}). The same approach can be used to obtain the chord (D_3), species profile (D_{18}), chi-square metric (D_{15}) and chi-square distance (D_{16}) matrices, as summarized in Fig. 7.8. (b) The transformed species data can also be used as input (dashed arrow) into linear methods of analysis, in particular PCA, RDA, and *K*-means partitioning. Modified from Legendre & Gallagher (2001).

An alternative method of computation for the asymmetrical distance coefficients D_3, D_{15}, D_{16}, D_{17} and D_{18} was proposed by Legendre & Gallagher (2001). The method consists of a transformation of the community composition data followed by the calculation of Euclidean distances (D_1) among sites. These two steps produce the distance function corresponding to the name of the transformation (Fig. 7.7). Data subjected to one of these transformations can also be used directly as input into linear methods of analysis that carry out computations in Euclidean space, such as *K*-means partitioning, PCA, and RDA (Sections 8.8, 9.1, 11.1). This approach is called transformation-based PCA (tb-PCA), transformation-based RDA (tb-RDA), and transformation-based *K*-means partitioning (tb-*K*-means).

1 — Transformation formulas

The following transformations, found in the vertical rectangle in the centre of Fig. 7.8, can be used to obtain the distance coefficients found on their left. The effect of these transformations is to remove the differences in total abundance (for abundance data) or total biomass (for biomass data) from the data, keeping the variations in relative species composition among sites. The chord and Hellinger transformations described below have been in use in community ecology and palaeoecology for a long time (e.g. Noy-Meir *et al.*, 1975; Prentice, 1980). Legendre & Gallagher (2001) showed

Species abundance paradox data ⇒
(3 sites, 3 species)

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

	Transformations ⇓	
$D_1(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$		$\mathbf{D}_1 = \begin{bmatrix} 0.0000 & 7.6158 & 9.0000 \\ 7.6158 & 0.0000 & 1.7321 \\ 9.0000 & 1.7321 & 0.0000 \end{bmatrix}$
$D_3(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2}} - \frac{y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{2j}^2}} \right)^2}$	$y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}}$	$\mathbf{D}_3 = \begin{bmatrix} 0.0000 & 0.3204 & 1.4142 \\ 0.3204 & 0.0000 & 1.4142 \\ 1.4142 & 1.4142 & 0.0000 \end{bmatrix}$
$D_{18}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$	$y'_{ij} = \frac{y_{ij}}{y_{i+}}$	$\mathbf{D}_{18} = \begin{bmatrix} 0.0000 & 0.2357 & 1.2472 \\ 0.2357 & 0.0000 & 1.2247 \\ 1.2472 & 1.2247 & 0.0000 \end{bmatrix}$
$D_{17}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \left[\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right]^2}$	$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}}$	$\mathbf{D}_{17} = \begin{bmatrix} 0.0000 & 0.1697 & 1.4142 \\ 0.1697 & 0.0000 & 1.4142 \\ 1.4142 & 1.4142 & 0.0000 \end{bmatrix}$
$D_{16}(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{y_{++}} \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$	$y'_{ij} = \sqrt{y_{++}} \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}}$	$\mathbf{D}_{16} = \begin{bmatrix} 0.0000 & 0.3600 & 4.0092 \\ 0.3600 & 0.0000 & 4.0208 \\ 4.0092 & 4.0208 & 0.0000 \end{bmatrix}$

Figure 7.8 Species abundance paradox data, modified from Orlóci (1978). The paradox is that the Euclidean distance between sites 2 and 3, which have no species in common, is smaller than that between sites 1 and 2 which share species 2 and 3. This results in an incorrect assessment of the ecological relationships among sites. With the other coefficients in this figure, which are asymmetrical, the distance between sites 2 and 3 is larger than that between sites 1 and 2, and the distance between sites 1 and 3 is the same as between sites 2 and 3, or very nearly so. Distance matrix $\mathbf{D}_{15}$ (not shown) is equal to $\mathbf{D}_{16}/\sqrt{y_{++}} = \mathbf{D}_{16}/\sqrt{15}$.

that these transformations were the first step towards the calculation of one of the asymmetrical distances that are appropriate for Q-mode analysis of community data. Only five of the coefficients discussed in this chapter can be computed by the two-step procedure described in Fig. 7.7, i.e. $D_3, D_{15}, D_{16}, D_{17}$ and D_{18}.

Chord transformation

1) *Chord transformation*. — The species abundances from each object (sampling unit) are transformed into a vector of length 1 using the following equation:

$$y'_{ij} = \frac{y_{ij}}{\sqrt{\sum_{j=1}^{p} y_{ij}^2}} \qquad \textbf{(7.67)}$$

where y_{ij} is the abundance of species j in object i. This equation, called the "chord transformation" in Legendre & Gallagher (2001), is available in the program CANOCO (Centring and standardisation for "samples": *Standardise by norm)* and in the ***decostand()*** function of VEGAN (method = "normalize"). If one computes the Euclidean distance (D_1) between two rows of the transformed data table, the resulting value is identical to the chord distance (D_3, eq. 7.35) computed between the rows of the original (untransformed) species abundance data table; this is how the chord distance can be computed through the two-step calculation shown in Fig. 7.7a. As a consequence, after a chord transformation, the community composition data are suitable for PCA or RDA, as well as other methods of analysis that preserve the Euclidean distance among the objects (Fig. 7.7b).

Species profile transformation

2) *Species profile transformation*. — The data can be transformed into profiles of relative species abundances through the following equation:

$$y'_{ij} = \frac{y_{ij}}{y_{i+}} \qquad \textbf{(7.68)}$$

This is a method of data standardisation that is often used prior to analysis, especially when the sampling units are not all of the same size. Data transformed in that way are called *compositional data*. In community ecology, the species assemblage is considered to represent a response of the community to environmental, historical, or other types of forcing; the variation of any single species has no clear interpretation. Compositional data are used because ecologists feel that the vectors of relative proportions of species can lead to meaningful interpretations. Relative abundances can be transformed into percentages by multiplying the values y'_{ij} by 100. Computing Euclidean distances among rows of a data table transformed in this way produces distances among species profiles (D_{18}, eq. 7.53). The transformation to relative abundance profiles is available in the ***decostand()*** function of VEGAN (method = "total"). Statistical criteria investigated by Legendre and Gallagher (2001) show that this is not the best transformation and that the Hellinger transformation (next paragraph) is often preferable.

Abundance data transformed into profiles by eq. 7.68 have the following property: centring the data by columns to means of 0 automatically centres the rows to means of 0. Make sure that the raw abundance data contain no row that sums to 0, though.

Hellinger transformation

3) *Hellinger transformation*. — A modification of the species profile transformation produces the Hellinger transformation:

$$y'_{ij} = \sqrt{\frac{y_{ij}}{y_{i+}}} \qquad \textbf{(7.69)}$$

Computing Euclidean distances among objects of a data table transformed in this way produces a matrix of Hellinger distances among sites (D_{17}, eq. 7.56; Fig. 7.7). The Hellinger distance has good statistical properties as assessed by the criteria of R^2 and monotonicity used by Legendre and Gallagher (2001) in their comparison of transformation methods. The Hellinger transformation is available in the ***decostand()*** function of VEGAN (method = "hellinger").

Chi-square distance transformation

4) *Chi-square distance transformation*. — A more complex modification of the species profile transformation is the chi-square distance transformation:

$$y'_{ij} = \sqrt{y_{++}}\frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \qquad \textbf{(7.70)}$$

where y_{ij} is a species presence or abundance value, y_{i+} is the sum of values over row (object) i, y_{+j} is the sum of values over column (species) j, and y_{++} is the sum of values over the whole data table. Euclidean distances computed among the rows of the transformed data table $[y'_{ij}]$ are equal to chi-square distances (D_{16}, eq. 7.55) among the rows of the original, untransformed data table. The chi-square distance transformation is available in the ***decostand()*** function of VEGAN (method = "chi.square").

The chi-square distance transformation equation reduces the value of an abundant species more than that of a rare species. Hence this transformation is interesting when one wants to give more weight to rare species; this is the case when the rare species are considered to be good indicators of special ecological conditions.

Chi-square metric transformation

5) *Chi-square metric transformation*. — The *chi-square metric* (D_{15}) only differs from the *chi-square distance* (D_{16}) by the constant $\sqrt{y_{++}}$ found in eq. 7.70. It can be obtained by the simplified transformation:

$$y'_{ij} = \frac{y_{ij}}{y_{i+}\sqrt{y_{+j}}} \qquad \textbf{(7.71)}$$

followed by calculation of the Euclidean distance. Data transformed using eq. 7.71 are smaller than the same data transformed using eq. 7.70 by a constant factor of $\sqrt{y_{++}}$.

Before applying the transformations described in the previous paragraphs, any of the standardizations investigated by Noy-Meir *et al*. (1975), Prentice (1980), and Faith *et al*. (1987) may be used if the study justifies it. These include species adjusted to equal maximum abundances or equal standard deviations, sites standardised to equal totals, or both. In particular, one may apply a square root or log transformation to the species abundances in order to reduce the asymmetry of the species distributions.

The chord and Hellinger transformations appear to be the best for general use. Legendre & Gallagher (2001) showed that the values of the corresponding distances are monotonically increasing across a simulated ecological gradient and are maximally related (R^2) to the spatial distances along the geographic gradient. Other asymmetrical distances, like D_{14}, that are useful for the analysis of community composition data cannot be obtained through the two-step process of a transformation followed by calculation of the Euclidean distance illustrated in Fig. 7.7. The chord and Hellinger transformations are closely related: chord-transformed abundance data are equal to squared abundance data that are then Hellinger-transformed.

The five transformations described above can be applied to presence-absence data. In that situation, the chord and Hellinger transformations produce identical results, and the corresponding distances, D_3 and D_{17}, are both equal to $\sqrt{2}\sqrt{1-\frac{a}{\sqrt{(a+b)\ (a+c)}}}$ where $\frac{a}{\sqrt{(a+b)\ (a+c)}}$ is the Ochiai similarity coefficient for binary data (S_{14}). Correspondence analysis, which preserves the chi-square distance, has long been used with species presence-absence data; hence the chi-square transformation can also be applied to this type of data.

2 — Numerical example

The modified Orlóci paradox data set was used in Subsection 7.4.1 to shows that the Euclidean distance function may produce misleading results when applied to assemblage composition data. Asymmetrical similarity and distance functions, which were specifically designed for the analysis of community composition data, do not have this drawback. Figure 7.8 (right-hand side) shows, for five distance functions, the distance matrices obtained for these data. From a community ecologist viewpoint, the identity of the species present at two sites is more important for assessment of the differences among these sites than their abundances. Following that conception, sites 1 and 2, which share two species, are more similar to each other than either of them is to site 3, which harbours a single species not found at sites 1 and 2.

Instead of that, Euclidean distances (D_1) show that sites 1 and 2 (D = 7.6158) are more dissimilar than sites 2 and 3 (D = 1.7321). This assessment would be considered incorrect by most community ecologists although the calculations are mathematically correct. In contrast, the four other distance matrices in Fig. 7.8 indicate that the two less dissimilar sites are 1 and 2, an answer that would be considered a correct

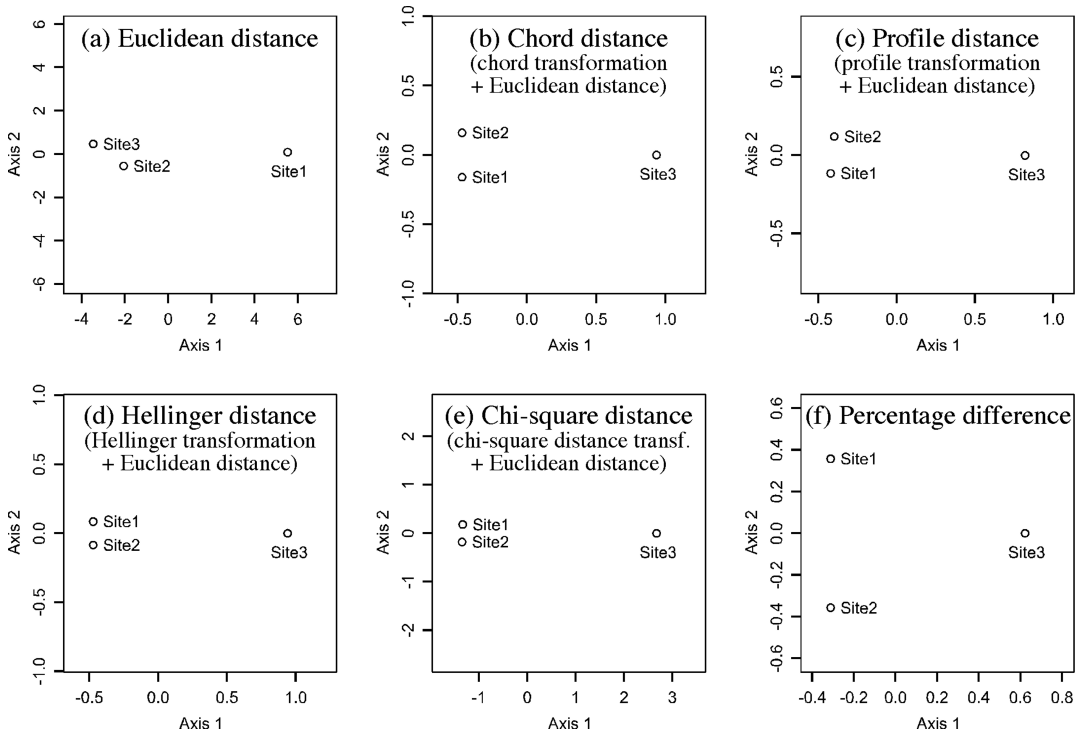

Figure 7.9 Principal coordinate ordination plots (PCoA, Section 9.3) of the distance matrices computed in Fig. 7.8: (a) D_1, (b) D_3, (c) D_{18}, (d) D_{17}, (e) D_{16}, and (f) a PCoA plot of the percentage difference (Steinhaus/Odum/Bray-Curtis) distance matrix (D_{14}) computed for the same data.

assessment of community similarity by most ecologists. Observe also that the chord and Hellinger distances produce a value of $\sqrt{2} = 1.4142$ between sites that have no species in common; this is the maximum value attainable by these distance functions, as noted in Subsection 7.4.1.

Figure 7.9 presents principal coordinate ordination plots (PCoA, Section 9.3) computed from the distance matrices in Fig. 7.8, plus a PCoA plot of the percentage difference matrix (D_{14}) computed for the same data. In the Euclidean distance ordination (Fig. 7.9a), sites 2 and 3 are the closest among the three sites, which may be seen as incorrect for the data under consideration. In all five other ordination plots (Fig. 9b-f), sites 1 and 2 are the closest. The plots also display the interesting property that the different asymmetrical distance functions deal with the differences among sites differently: sites 1 and 2 are the closest to each other in Fig. 7.9e and the farthest in Fig. 7.9f (percentage difference). The distance between sites 1 and 2 would be even larger if PCoA had been computed from square-rooted D_{14} values, which is recommended before PCoA to make percentage difference matrices Euclidean.