7

Ordination

(The 'art' of efficient reduction of dimensionality)

As mentioned in Section 2.1, any conventional data matrix has two alternative geometric representations: the objects are points in a space spanned by variables as axes and vice versa, the variables are points in a space with the objects as axes. Section 2.2 already introduced some simple graphical means designed to allow simple examination of these data structures. The methods described in Chapters 4-6 also operate in these multidimensional spaces, with the explicit purpose of arranging the points into groups or graph structures such that reduction of dimensionality is only indirectly present. It is left to the present chapter to introduce procedures whose primary objective is to replace original dimensions by a few artificial axes so as to represent data structure as efficiently and faithfully as possible. After Goodall (1954), these methods will be discussed under the common heading of *ordination*, although they do not form a mathematically homogeneous category (for example, 'scaling' refers to a subset of ordination procedures with the specific purpose of 'expanding' distance matrices backwards into a new artificial 'data space'). For the majority of methods, the objects studied belong to the same group, but this is not necessarily so. Discriminant analysis, as a special ordination technique, derives new axes in order to maximize separation among groups of objects defined *a priori*. In addition, variables may also be grouped into two classes on strict logical grounds before the computations, a separation considered important in reducing dimensionality through canonical correlation and canonical correspondence analysis. As seen, ordination is understood in a much wider sense than usual: *any technique that extracts artificial variables in order to reduce the dimensionality of the data is referred to as ordination*. These variables are termed differently for each technique, for example, component, factor, canonical variate and so on.

Whereas in cladistics the greatest intellectual demand is to understand many, potentially new or ambiguous terms, the methodology of ordination assumes knowledge of elementary matrix algebra. Without familiarity with the fundamentals, even a most simplified description of ordination algorithms will be difficult or even impossible to follow. Nevertheless, I will try to present an intuitive characterization (with the aid of graphic displays, assuming again that the reader is a visual type mentally) before giving the mathematical details of each technique.

Going deeply into the subject is inevitable for those wishing to make an appropriate choice of methods, thus avoiding the pitfall of applying the actually 'most popular' ordination procedure and program to any research problem. In a sharp contrast with clustering, some initial conditions must be satisfied by the data[1], otherwise the interpretation of ordination results can easily lead to false conclusions.

7.1 A fundamental ordination method: principal components analysis

Principal components analysis (generally abbreviated as PCA[2]) is as central a procedure of multivariate data exploration as the analysis of variance in conventional biometry. Its detailed discussion is prerequisite to the treatment of any other ordination procedure. The method has been rooted in the pioneering works by Pearson (1901) and – in particular – Hotelling (1933, 1936). PCA had remained only of theoretical significance for decades, but the development of high-speed computers abruptly interrupted the dream of this 'Sleeping Beauty' in the sixties.

The underlying principle of the method can be introduced in several ways (cf. Mirkin 1996); a graphical approach is perhaps the most appropriate here. Figure 7.1a depicts a very simple situation, since the original dimensionality of the point cloud is only two – chosen that way deliberately for didactic reasons. In reality, the number of dimensions to be explored is much higher of course. Observe that the two variables (axes x_1 and x_2) explain approximately the same portion of the total variance for the ten points (Equation 3.108). However, if a new axis is laid down so as to coincide with the longitudinal axis of the cloud (longer dotted line in the figure) then this axis will account for almost all of the variance (the meaningful variation), whereas the other new axis, perpendicular to the previous one, explains a tiny fraction only (the stochastic variation or noise in the data). These axes, drawn by hand for the time being, are called the *components*.

To sum up what has been said: without modifying the relative positions of the points the original coordinate system is replaced by a new one such that the first new axis encompasses the maximum variation, leaving as little as possible for the second axis. The logic is the same for any number of starting variables, because every subsequent axis is derived to explain the highest percentage of variance remained after determining the previous axes. At this point, we must admit that the number of components is not necessarily smaller than the number of original variables – rearranging the shares in total variance does not imply automatic reduction of dimensionality (the maximum number of components that can be obtained is discussed on p. 220). The real achievement here is that a few components will summarize most of the variation, and the majority of the new artificial variables will be negligible and can be discarded. This is what we mean by dimension reduction. The efficiency of this operation is of course *case dependent*, the stonger the linear correlations (Equation 3.70) between the variables, the fewer axes will be necessary to explain the meaningful linear variation in the data. If there is little or no correlation among the original variables, then PCA will not add anything new to the

1 Non-metric multidimensional scaling (Subsection 7.4.2) is a noted exception.

2 An alternative is Digby & Kempton (1987) who prefer the acronym PCP. Universality is not a requirement, of course, and it is more essential that authors use the abbreviations consistently in the whole book or paper.

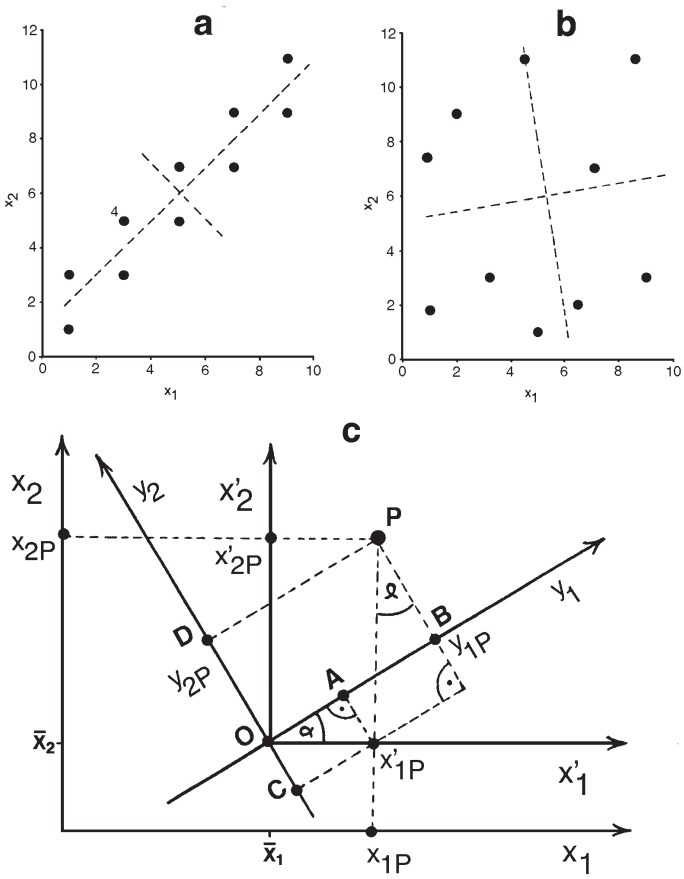

Figure 7.1. Graphical illustration of principal components analysis. **a**: variance-extraction is efficient for highly correlated variables; **b:** the components do not help much when the variables are linearly uncorrelated; **c:** diagram showing the calculation of coordinates for point P in the component space. (Note that y_{1P} and B refer to the same point, as well as y_{2P} and D.)

picture, the only thing that happens is that the coordinate system is shifted to the centroid (Fig. 7.1b). This is because there are no particular 'directions' in the point cloud to which components could be fitted efficiently. The success of PCA is therefore conditioned upon strong linear correlations of the variables, a criterion often satisfied by biological data. A byproduct of using PCA is the identification of groups of highly correlated variables, as we shall see below.

First, it is shown that the new coordinates of objects can be derived based on the original coordinates and the angles between components and original axes. What the reader needs is merely a high-school level background in planar geometry. In Figure 7.1c, the original variables are x_1 and x_2, whereas the components are denoted by y_1 and y_2. For the sake of clarity, only a single object is represented in the diagram, by point P, the others are removed. α is the angle between variables x_1 and component y_1. In the first step, the data are centered (Formula 2.2), that is, from each value the mean of the given variable is subtracted. As a result, the origin O (the intersection of axes x'_1 and x'_2) of the new coordinate system will be shifted to the centroid of the point cloud. Let the centered coordiantes of point P be x'_{1P} and x'_{2P}, that is

$$x'_{1P} = x_{1P} - \bar{x}_1, \text{ and } x'_{2P} = x_{2P} - \bar{x}_2 \qquad (7.1)$$

The coordinates for point P on the new axes are obtained using the line segments $\overline{OA}$ and $\overline{AB}$, as well as $\overline{OC}$ and $\overline{CD}$, following elementary trigonometry:

$$y_{1P} = \overline{OA} + \overline{AB} = \cos\alpha\ x'_{1P} + \sin\alpha\ x'_{2P} \quad (7.2a)$$

$$y_{2P} = -\overline{OC} + \overline{CD} = -\sin\alpha\ x'_{1P} + \cos\alpha\ x'_{2P} \quad (7.2b)$$

Since $\sin\alpha = \cos(90^o - \alpha)$, and $(90^o - \alpha)$ is the angle between the second variable and the first component, the above equations can be rewritten using the cosine function exclusively:

$$y_{1P} = \cos\alpha\ x'_{1P} + \cos(90^o - \alpha)\ x'_{2P} \quad (7.3a)$$

$$y_{2P} = -\cos(90^o - \alpha)\ x'_{1P} + \cos\alpha\ x'_{2P} \quad (7.3b)$$

In words, the coordinate of P on component y_1 is derived as the sum of P's coordinate on axis x'_1 multiplied by the cosine of the angle between x_1 and y_1 and its coordinate on axis x'_2 multiplied by the cosine of the angle between x_2 and y_1. That is, the centred data and the angles are required. The new coordinates are called the *component scores*. In matrix algebraic terms, the above equations can be rewritten as:

$$\mathbf{y} = \mathbf{V}'(\mathbf{x} - \bar{\mathbf{x}}), \text{ that is } \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\alpha & \cos(90-\alpha) \\ -\cos(90-\alpha) & \cos\alpha \end{bmatrix} \begin{bmatrix} x_{1P} - \bar{x}_1 \\ x_{2P} - \bar{x}_2 \end{bmatrix} \quad (7.4)$$

Matrix **V** is therefore a rotating matrix which moves the point P into a new coordinate system. The column h of the matrix contains the cosines of the angles between the original variables and component h ('direction cosines'), so that the matrix algebraic analogue 7.4 of 7.3 is achieved using the transpose of **V**. Matrix **V** satisfies the equation $\mathbf{V}'\mathbf{V} = \mathbf{I}$, the condition for the *orthonormality* of its columns (that is, the components are orthogonal to each other, see Appendix C).

The 'only' thing that remained is the calculation of the matrix of direction cosines. Matrix **V** in fact converts the original covariances and variances into 0 covariances (corresponding to zero correlations between components) and the 'rearranged' variance shares, according to the following equation:

$$\mathbf{V}'\,\mathbf{C}\,\mathbf{V} = \mathbf{L}, \quad (7.5)$$

in which **C** is the variance/covariance matrix of variables, and **L** is the variance-covariance matrix of the components. In the latter, because the covariances are zero, only the diagonal scores assume positive values, so that **L** is a diagonal matrix. These values will be denoted by λ_h. The column vector $\mathbf{v}_h$ and its associated variance λ_h satisfy the following matrix equation:

$$(\mathbf{C} - \lambda_h\mathbf{I})\ \mathbf{v}_h = \mathbf{0}. \quad (7.6)$$

To solve the equation we assume that

$$|\mathbf{C} - \lambda_h\mathbf{I}| = 0\,. \quad (7.7)$$

Expansion of the determinant yields several solutions (roots) for λ_h (see below), each associated with a vector $\mathbf{v}_h$, such that $\mathbf{v}_h'\mathbf{v}_h = \sum v_{ih}^2 = 1$ (the length of the vector is 1, otherwise we do not get direction cosines). Given this condition and the diagonal values of **L**, Equation 7.6 directly provides the column vectors $\mathbf{v}_h$ of matrix **V**. The λ values are called the *eigenvalues*, the **v** vectors are termed the *eigenvectors* of **C** (see Appendix C).

The above derivation is illustrated by a small numerical example. Consider the two variables depicted in Figure 7.1a, which provide the following coordinates for the ten points:

```
variable 1:   1 1 3 3 5 5 7 7 9   9
variable 2:   1 3 3 5 5 7 7 9 9 11
```

The variance-covariance matrix for the two variables is given by:

$$\mathbf{C} = \begin{bmatrix} 8.89 & 8.89 \\ 8.89 & 10.0 \end{bmatrix}$$

(the variance of variable 1 happens to be equal to the covariance of the two variables by accidence). First, we solve Equation 7.7 by writing

$$\mathbf{C} - \lambda\mathbf{I} = \begin{bmatrix} 8.89 & 8.89 \\ 8.89 & 10.0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} 8.89 - \lambda & 8.89 \\ 8.89 & 10.0 - \lambda \end{bmatrix}.$$

Since the determinant of a 2 × 2 matrix is

$$\begin{vmatrix} a & c \\ b & d \end{vmatrix} = ad - bc,$$

we obtain that

$$(8.89 - \lambda)(10.0 - \lambda) - 8.89^2 = \lambda^2 - 18.89\,\lambda + 9.87 = 0,$$

with the two roots $\lambda_1 = 18.35$ and $\lambda_2 = 0.54$.

At this point we should pay attention to an important observation. Whereas the variances of the two variables were 8.89 and 10.0, respectively, that is their share from the total variance is similar (47% compared to 53%), the new variances are considerably different from each other. The variance of 18.35 obtained for the first component accounts for 97% of the total, leaving a tiny fraction of 3% to the second component. Thus, in this case variance 'compression' is very efficient. The sum of the eigenvalues equals the total variance, which is the sum of diagonal values in **C** (the trace of the matrix). The relative importance of an eigenvalue λ_h is therefore $100 \times \lambda_h / \mathrm{tr}\{\mathbf{C}\}$ percent. Note also that the product of the eigenvalues (in this case: 9.9) equals the determinant of the variance/covariance matrix ($10 \times 8.89 - 8.89^2$), a quantity often called the *generalized variance.*

Having determined the eigenvalues, the eigenvectors are easily obtained. By omitting the details of calculation, we get the following rotating matrix for the two eigenvalues derived above, satisfying the condition of unit vector length:

$$\mathbf{V} = \begin{bmatrix} 0.685 & -0.729 \\ 0.729 & 0.685 \end{bmatrix} \text{ that is } \mathbf{V}' = \begin{bmatrix} 0.685 & 0.729 \\ -0.729 & 0.685 \end{bmatrix}$$

For example, the new coordinates of point 4 are obtained by Equations 7.3a-b, using the two means (5 and 6) of the variables as follows

$$y_{14} = 0.685\,(3 - 5) + 0.729\,(5 - 6) = -2.099$$
$$y_{24} = -0.729\,(3 - 5) + 0.685\,(5 - 6) = 0.773$$

The result is easily verified by examining Figure 7.1a. If one wishes to check its correctness by a commercial package as well, and finds that the signs of component scores differ from the above, then it does not mean that the calculations are wrong. The signs are absolutely arbitrary, the whole configuration may be reflected over the axes. Note also that the vectors of **V** are normalized in both directions (the sum of squared elements is 1). That is, not only the eigenvectors have unit length, but the sum of squared direction cosines for a given variable is also unity: $\sum v_{ih}^2 = 1$. In other words, the condition of orthonormality satisfies for the rows as well: $\mathbf{VV}' = \mathbf{I}$!

The above two-variable example serves the purpose of illustration of what is going on during calculations in PCA. The worked example can be expanded to more than two variables, but the calculations are better performed by the computer. A usual definition of PCA (e.g., Manly 1986) is that the components are *linear combinations* of the variables, as shown by Equations 7.3, given the conditions of orthonormality. In addition, PCA is also considered as a generalization of regression analysis using minimum sum of squares (Jongman et al. 1987). Component 1 was laid down in

Fig. 7.1a such that the sum of squared distances of the points from the line was minimized. The angle between the component and any original variable can be computed by iterations, thus circumventing the use of determinants and eigenvalues.

The possible number of components is an interesting issue in ordination theory. The solution of Equation 7.7 provides t positive eigenvalues with the following restrictions:

$$t \leq \min \{ n, m - 1 \} \tag{7.8}$$

where, as everywhere in this book, n is the number of variables and m is the number of points (objects). This inequality implies that no more than n components can be extracted from the data when there are more points than variables. The number of components can be lower than this when some variable(s) can be expressed as linear combinations of the others, i.e., some function describes their relationship (its simplest case is a unit correlation between two variables). If $m–1 < n$, then this inequality becomes decisive because, in some sense, the many variables 'overdefine' the small number of points. It is known from multidimensional geometry that we need a maximum of $m–1$ dimensions in order to depict the distance relationships for m objects (the distance for two points is measured correctly along the line, in one dimension; the Euclidean distances between three points are faithfully represented on the plane, in two dimensions, and so on...). The number of positive eigenvalues, t, is termed the *rank* of matrix **C**; it is in fact the background or inherent dimensionality of the data structure (Appendix C).

A natural question arises immediately in every user of principal components analysis: what is the number of potentially '*meaningful*' dimensions that need to be demonstrated, and how many dimensions explain only stochastic variations in the data and can therefore be ignored? There are several possibilities to examine this problem. If the egivenvalues are ranked, then the relative importance of dimensions becomes interpretable. The simplest graphical vehicle to illustrate this is the so-called *scree diagram* (Cattell 1966, see inset in Fig. 7.2). According to the original proposition, a breakpoint from which the eigenvalues start to decrease very slowly is identified in this bar diagram. Dimensions pertaining to the eigenvalues before this point are deemed to be 'important', contrary to the others. This subjective rule of thumb may work in some cases, but cannot be recommended as a general solution. The Kaiser criterion (Mardia et al. 1979) selects components that exceed the mean of all eigenvalues. The use of this criterion usually provides fewer 'meaningful' dimensions than the scree diagram. Statistical tests of the significance of PCA axes are also possible. More conventional formulations require the multivariate normality of the data; if this condition holds true then Bartlett's isotropy test may be used to find the breakpoint (Mardia et al. 1979). More advanced techniques based on bootstrap tests do not assume normality of data and thus have more general validity (Jackson 1993, Pillar 1999b).

Principal components analysis is illustrated first using the phytosoiological data of Table A1 (Appendix A). In this, eight sample plots are characterized in terms of percentage cover scores of 12 plant species (variables). The efficiency of variance extraction is clear from the results: the first two components (Fig. 7.2) account for 82% of the total variance (55% and 27%, respectively), leaving only 18% to the remaining five components (there are 7 positive eigenvalues; refer to the discussion on top of this page to see why). Plots taken in open grasslands are relatively close to one another in the scatter diagram, whereas the more closed community stands, represented by plots 7 and 8, are positioned far apart, illustrating pretty well the quantitative differences in species cover data. The scattergram agrees well with the hierarchical classification of objects which is superimposed onto the ordination via concentric lines drawn around the groups.

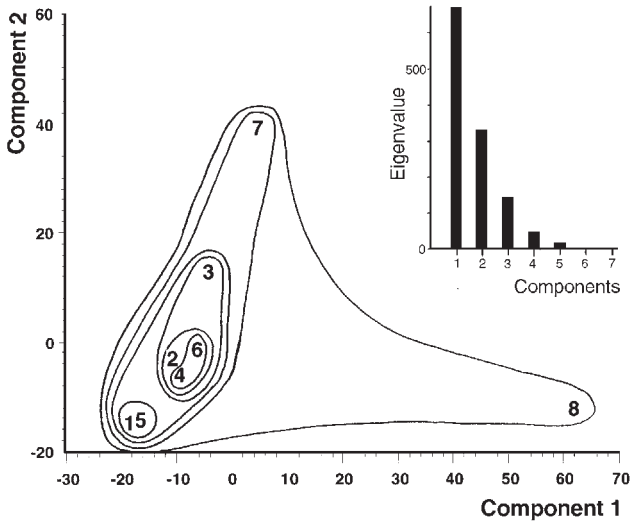

Figure 7.2. PCA ordination of phytosociological relevés of Table A1. The intersection of axes is moved to the lower left corner to improve clarity of the diagram (a convention to be followed throughout in this book). Contour lines represent different steps of a hierarchical classification from Euclidean distances. The scree diagram (inset) depicts the relative sizes of eigenvalues.

7.1.1 Component-covariance and correlation: ordination of variables

An advantage of PCA is that in addition to portraying the multidimensional scatter of objects via efficient variance extraction, the relationships among variables can also be evaluated thoroughly. Based on the component scores and the original data, Formula 3.69 could be used to yield the covariance between components and variables, allowing interpretation of components in the context of the actual study. The formula of covariance is not needed, however, because from Equation 7.6 we have that

$$\mathbf{Cv} = \lambda\mathbf{v}, \tag{7.9}$$

in which $\lambda_h v_{ih}$ corresponds with the *covariance* of variable i and component h. Standardizing this covariance by the standard deviation of the variable (s_i) and the component (the eigenvalue is its variance, so that the standard deviation is its square root, $\sqrt{\lambda_h}$) provides the so-called *component correlation* sought:

$$r_{ih} = \lambda_h v_{ih} \,/\, s_i\sqrt{\lambda_h} = v_{ih}\sqrt{\lambda_h} \,/\, s_i \tag{7.10}$$

(an alternative term is *component 'loading'*). The component covariances and correlations can be used to draw an ordination diagram of variables. In this, the axes are the components as above, and the points represent the variables. As yet, between-variable relationships were expressed in terms of covariances, so it is absolutely logical to measure component/variable relationships in the same way. That is, the coordinate of variable i on axis h will be given by $\lambda_h v_{ih}$, a value without theoretical upper limit. The use of correlation is also plausible, but its interpretation is more straightforward for the case of standardized PCA (Subsection 7.1.4). It is noted here that in case of correlation the coordinates will fall into the interval [–1,1]. Owing to the orthogonality of axes, the correlations between a given variable and the first two components are constrained to fall within the unit circle drawn around the origin (for all components, the point representing a variable will lay on the surface of the unit radius hypersphere,

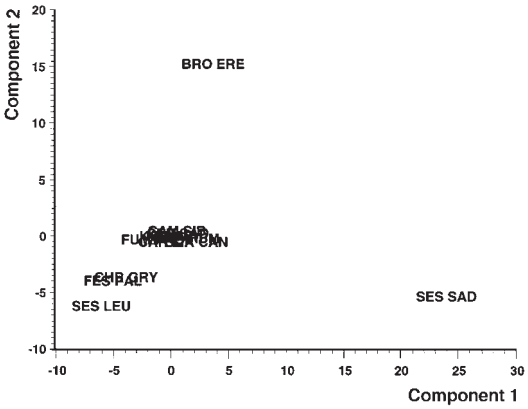

Figure 7.3. The ordination of species of Table A1 based on component covariances. The labels are centered on ordination positions. Several labels overlap around the origin, these indicate species with low variance in the data set.

that is $\sum_{h=1}^{t} r_{ih}^2 = 1$). Note, further, that in the ordination space r_{ih} is the cosine of the angle between component *h* and the vector that represents variable *i.*

The alternative PCAs, one based on component covariances and the other on correlations, of species in Table A1 deserve a comparison. In Fig. 7.3, the variables (species) are arranged according to component-covariances. As seen, the components are determined primarily by species that have a high cover value in at least one site, thus large variance in the data set: the first axis is determined by *Sesleria* and the second by *Bromus*. Three less varying species, namely *Seseli leucospermum, Festuca glauca* and *Chrysopogon gryllus*, represent a tendency opposed to the first two, whereas the species with even lower cover are concentrated around the origin. When correlation is used, these absolute cover differences diminish and – as a result – the points become more evenly scattered in the ordination space (Fig. 7.4). Component 1 is positively correlated with the species typical of closed grasslands, and negatively with species of open areas. This is the strongest contrast, the main trend, that can be deduced from this extremely simple data table. *Bromus erectus* apparently 'strives for' independence from the other species, because it is present everywhere with small or high cover. This behavior is

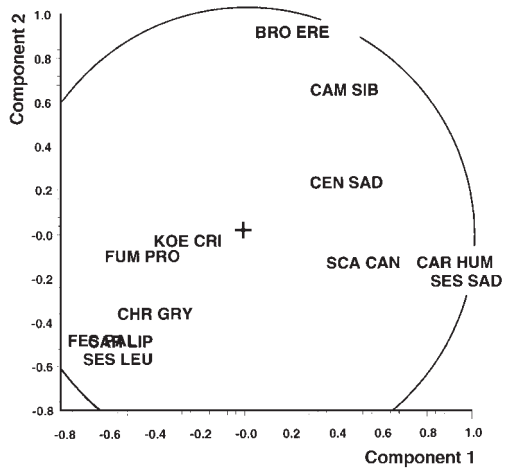

Figure 7.4. PCA ordination of species of Table A1 based on component correlations. Labels are centered on species positions. Compare this diagram with Fig. 7.3!

Table 7.1. Percentage contributions of the first five components to the variance of sites and species in the PCA of Table A1. The highest score in each row is set by boldface.

Sites	Components and their % contributions				
	1 (55%)	2 (27%)	3 (12%)	4 (4%)	5 (1.5%)
1	32.788	20.828	39.445	6.354	0.575
2	**49.861**	6.096	0.196	0.723	42.499
3	7.354	**63.758**	12.651	13.412	2.024
4	33.613	12.784	**38.315**	5.207	6.086
5	**41.815**	30.638	1.654	25.245	0.035
6	9.825	0.455	**82.130**	6.476	0.006
7	1.258	**92.687**	5.169	0.872	0
8	**96.230**	3.636	0.134	0	0
species					
BRO ERE	4.488	**83.167**	11.594	0.351	0.028
CAM SIB	19.279	**42.352**	10.104	3.185	18.178
CAR HUM	**85.061**	1.790	11.184	0.361	0.460
CAR LIP	**30.611**	23.965	3.248	8.452	17.107
CEN SAD	**19.884**	5.256	7073	6.169	13.545
CHR GRY	15.597	12.942	**48.084**	23.269	0.108
FES PAL	**39.173**	23.352	16.754	19.037	1.473
FUM PRO	20.446	0.968	**42.720**	1.465	26.912
KOE CRI	6.194	0.099	**67.650**	17.206	7.508
SCAN CAN	26.722	1.818	2.256	0.417	**63.037**
SES LEU	31.618	**32.224**	28.702	5.829	1.436
SES SAD	**94.276**	4.752	0.925	0.038	0.003

reflected by the high correlation with component 2. Species falling close to the origin *(Koeleria glauca, Scabiosa canescens, Centaurea sadleriana)* are the most independent from the other species, and their role is negligible in determining data structure. In general, one might say that the *closer a point of a variable to the unit circle, the more completely it is explained by the given two components* in terms of correlation structures.

7.1.2 Percentage contributions

When evaluating the role of a component in explaining the variance of an object or a variable, graphical display is not the only possibility; the relationships can also be expressed quantitatively. The percentage contribution of component h to the 'variance' of object k (more precisely, to the sum of its squared deviations from the origin) is given by the following formula:

$$z_{hk} = 100 \cdot y_{hk}^2 \,/\, \sum_{j=1}^{t} y_{jk}^2 \qquad (7.11)$$

where y_{hk} is the score of object k on component h. Using the percentages, one may determine how many components are responsible for the position of any point, components affecting a few or no objects can be identified, and objects with an 'average' behavior can be found. The proportion of variance of variable i accounted for by component h is obtained as:

$$w_{hi} = 100 \cdot v_{ih}^2 \, \lambda_h \,/\, \sum_{j=1}^{t} v_{ji}^2 \, \lambda_j \qquad (7.12)$$

If variable i is highly correlated with component h, then the percentage will also be high, leaving a small fraction to the remaining components. Variables with high component correlations

are the most important in the interpretation of PCA results. On the other hand, if the variance contributions are evenly distributed over components, then the variable usually has very little meaning, and can be discarded safely from any interpret

> It is left to the reader to compare the percentages presented in Table 7.1 and Figures 7.2-3. It turns out that components that were ignored thus far may prove to be important in explaining some sites and species. For example, most of the variation of *Scabiosa canescens* is accounted for by component 5, showing the individualistic behavior of the species (it can have small or medium cover in closed and open stands alike). Its overall effect is nevertheless negligible, because component 5 explains no more than 1.5% of the total variance.

7.1.3 Simultaneous ordination of objects and variables: the biplot

The comparative evaluation of the separate ordinations of objects and variables is cumbersome, especially for many points, raising the need for some combination of the two configurations. Gabriel (1971, 1981) proposed first a method to derive a joint display of objects and variables, incorporating all essential information on data structure in a single diagram, called the *biplot*. (In this term, the prefix *bi-* refers to the number of ordinations combined, rather than to the number of dimensions!) Since the coordinates of objects and variables are expressed on different scales, the variable scores are to be multiplied by an appropriate factor to allow superposition of the two ordinations. As a further aid to subsequent interpretation, arrows are directed from the origin towards the points representing variables, so that the relationships between components and variables become more apparent.

There are several ways of defining biplot coordinates. In the original proposition by Gabriel, the biplot is not merely a superposition of two configurations, but also a tool of an optimum reconstruction of original data from the components – whose success is always case dependent. In the Gabriel biplot, the coordinates of objects are determined as described above, whereas the coordinates of variables are provided by the respective values of the eigenvectors (i.e., the direction cosines), multiplied by an arbitrary scale factor. For this type of diagrams, ter Braak (1983) suggests the name *Euclidean biplot*, because in this case the interpoint distances in the ordination, multiplied by the scale factor, give the best approximation to the original distances of objects. The higher the coincidence between a component and a variable in the multidimensional space, the smaller the angle between the corresponding axis and arrow in the biplot diagram.

> The Euclidean biplot derived from the PCA of Table A1 is displayed in Fig. 7.5. Species with low cover, hence with low variance, are concentrated around the origin so that their labels overlap and are therefore unreadable. The relative positions of species are similar to those in the covariance-based ordination (Fig. 7.3). Consequently, *Bromus*, *Sesleria* and *Seseli* have the strongest impact upon the configuration.

The reconstruction of original data, i.e., approximation of species abundances in the above example, is achieved in the following way. By projecting point *j* onto the arrow (or its extension) of species *i,* one obtains the so-called *fitted score* of that site. The fitted score is an 'estimated' cover of species *i* in site *j*. When the fitted score is positive, the given species is expected to have a larger abundance in the site than the average, if negative, the abundance is lower than the average (recall that the data were centered before the analysis). The larger the variance explained by the two axes in question, the more efficient this approximation. (In any

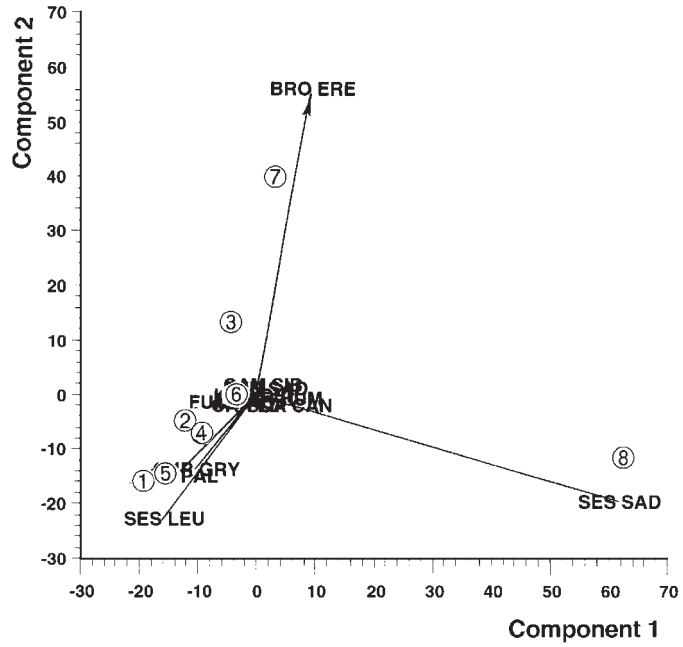

Figure 7.5. Euclidean biplot: the species are projected onto the ordination of sites (Fig. 7.2) according the the eigenvector scores (direction cosines) as coordinates multiplied by the arbitrary factor of 67.05.

case, the best fit is always obtained for the first two components, see Appendix C: *singular value decomposition* [expression C50, in which **US** stands for the coordinates of objects and **V**′ contains the eigenvectors]).

> To illustrate this, let us extend in the diagram of Fig. 7.5 the arrow pointing to *Bromus* beyond the origin, in a 'negative' or backward direction, and project each point to this line. Then, we see that in sites 1, 2, 4, 5 and 6 this species has a cover below the average (11.5) such that the fitted score approximates pretty well the original value. For site 8, this reconstruction is a bit less successful, because it is in positive domain although the original value is 11 only. Points 3 and 7 are far over the average, in good agreement with the abundance of *Bromus* in these two sites.

Another possibility is the so-called *Mahalanobis biplot,* for which the coordinates of objects are recalculated according to $u_{hi} = y_{hi} \,/\, \sqrt{\lambda_h \,\square\, (m-1)}$. As a result of this transformation, the variances on the components will be equalized, because the vectors of coordinates will have a unit length. (What we have derived is an element of the left matrix obtained from the singular decomposition of the data matrix, see Formula C50 in Appendix C). Interpoint distances for all components correspond to the Mahalanobis generalized distances of objects, that is, we have the best possible two dimensional approximation to these distances in the space of the first two components. In the biplot, the coordinates of variables are the component covariances (Formula 7.9), the length of arrows is proportional to the standard deviation of the variable, and the angles between any pair of arrows are proportional to the correlation between the corresponding variables. If we look at the formula of singular value decomposition in Appendix C, we realize that the data may be reconstructed from the ordination scores

(**U** contains the coordinates of objects, whereas **SV′** includes coordinates of variables in Formula C50).

A natural question may immediately arise in the reader: in preparing a biplot why do not we superimpose the ordination of objects which approximates their Euclidean distances (this one from the Euclidean biplot, Fig. 7.2) onto an ordination of variables based on component covariances (from the Mahalanobis biplot, Fig. 7.3)? It is of course a straightforward possibility, and the resulting diagram is a suggestive joint display of objects and variables (as an illustration, see Fig. 7.6c for a different version of PCA). Many authors do not consider this as a true biplot because it is not suitable to the reconstruction (more precisely, to the approximation) of original data. Nevertheless, Rohlf (in Marcus 1993) treats this as a reasonable alternative to the former biplot types, implicitly expressing the view that a biplot need not always be a tool of optimal data reconstruction. I share his views, because the other objective of joint displays, the facilitation of the mutual interpretation of two configurations is equally if not more important in biological data analysis. Often, the Euclidean and Rohlf biplots do not differ significantly because the ordinations of variables by eigenvectors and component covariances are similar, as are in the present example.

7.1.4 Standardized PCA

The previous subsections introduced the 'regular' version of principal components analysis which implies centring of the original data, hence its usual name, *centred PCA*. We understood from the sample results that the positions of the first components were determined mostly by variables with high variance, on account of variables with low or negligible variance; there was an unequal implicit weighting in the analysis. However, the user of this method may want to give equal weight to all variables, for example, when all species are to be considered equal in importance irrespective of the absolute cover values. Whenever the input variables are measured in different measurement scales, equalization becomes 'obligatory'; otherwise the results will reflect little more than our arbitrary choices of measurement scales and units. Equalization is achieved by standardizing the variables by Formula 2.4; in addition to centring each score is divided by the standard deviation of the corresponding variable. In effect, the multidimensional shape of the point cloud is changed – but our search for optimum directions will still make sense for the researcher.

This *standardized PCA* differs from the centred version in the following features:

- The analysis starts from the **R** correlation matrix of variables, rather than the matrix **C** of variances/covariances. Computing **R** implies the required standardization, so that the data need not be divided by standard deviation in advance.
- Standardization has the obvious consequence that all variables will have unit variance, therefore the total variance is n, just like the sum of eigenvalues. The percentage share of component h in the total variance is thus $100 \times \lambda_h/n$.
- Components with variance (eigenvalue) lower than 1 may be discarded safely from the interpretation of PCA results. A component conveys no information for us in this case, because its variance is lower than the variance of any standardized variable.

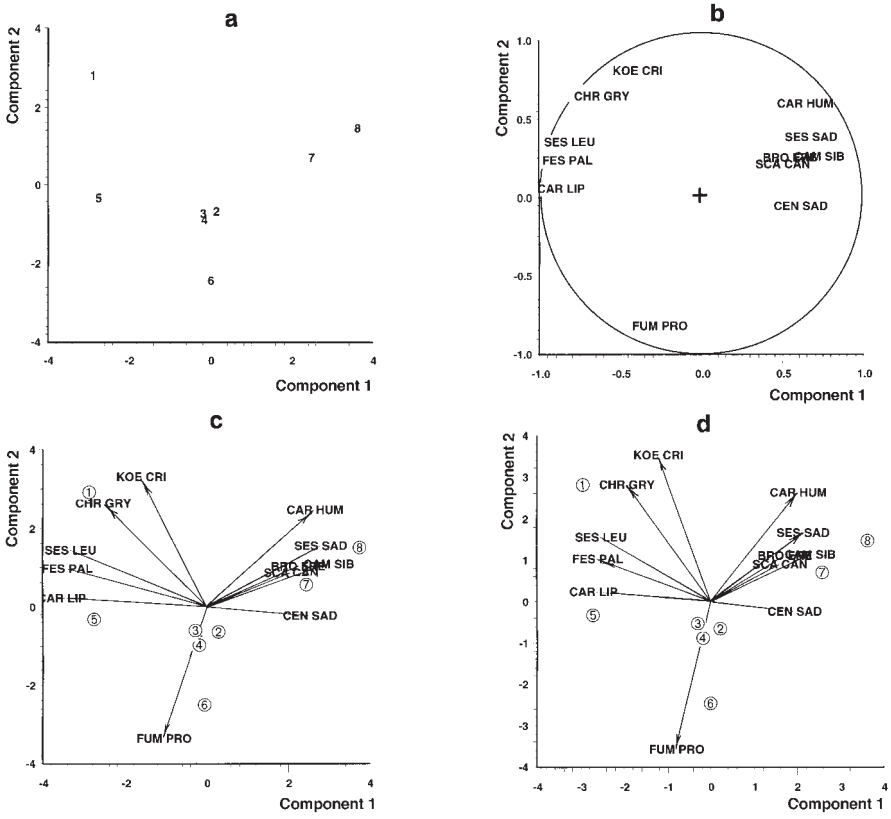

Figure 7.6. Standardized PCA of the phytosociological data in Table A1. **a:** ordination of sites, **b:** ordination of species based on component correlations, **c:** Rohlf biplot using the correlations, **d**: Euclidean biplot.

- The correlation of component h and variable i is obtained by the simplified formula $r_{ih} = v_{ih}\sqrt{\lambda_h}$, and in fact equals the component covariance. Therefore, it may be used to construct the Rohlf biplot.

The balanced effect of variables upon the result is shown by the standardized PCA of the data in Table A1 (Fig. 7.6). The first two components explain 'only' 63%, in contrast with 82% reached by centred PCA. The positions of sites also changed, of course; objects 7 and 8 fall closer to each other, because the discriminatory power of *Bromus* and *Sesleria*, the two variables mainly responsible for their separation, is diminished. On the contrary, sites 1 and 5 fall further apart, because their small absolute differences are exaggerated by standardization (Fig. 7.6a). Component 1 appears to reflect a contrast between open and closed stands of the grassland. Standardization influences the grouping of species as well; in accordance with the open versus closed trend, there are two groups of species (Fig. 7.6b). The separation of *Fumana* reflects its 'individualistic' behavior among the species. The slight differences between the Rohlf (Fig. 7.6c) and Euclidean (Fig. 7.6d) biplots are caused by the multiplier $\sqrt{\lambda_h}$ used in the former case (see the last item in the bulleted paragraphs above).

In a phytosociological study, it is up to the researcher to decide whether or not to use data standardization. In taxonomic studies, however, meaningful ordinations are obtained by standardized PCA only. This is the case for the *Iris* data set (Table A2) as well, because the length values are about three times larger than the width scores. The analysis shows that while the

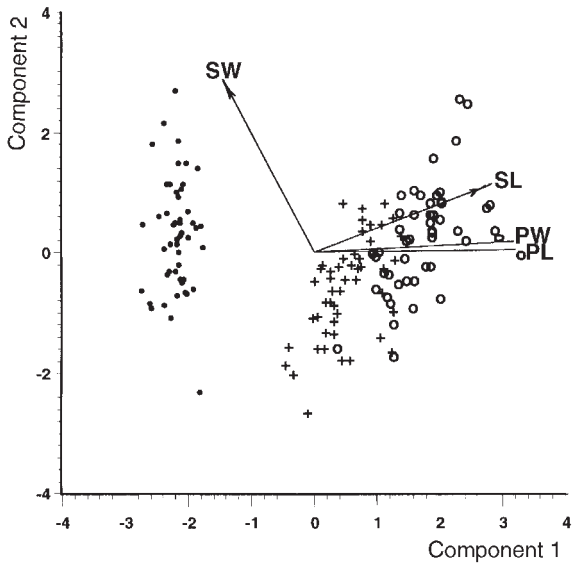

Figure 7.7. Rohlf biplot derived from the standardized PCA of *Iris* data of Table A2. Symbols:]: *I. setosa*, +: *I. versicolor,* o*: I. virginica*. Abbreviations: S: sepal, P: petal, L: length, W: width.

width and length of petals are almost 'perfectly' correlated, measurements for the petals are uncorrelated (the corresponding arrows are at right angle in Fig. 7.7, compare this with Fig. 2.3). The first component is practically an overall size component for the petals, whereas the second component can be identified with sepal width (we have seen from the data that the petals of *I. setosa* are much smaller than the other two species). These two components account for 73% and 23% of the total variance, respectively, that is, almost everything. While component 1 appears to separate the species well, the second is not suitable to this purpose (the separation of these taxa will be examined once more using canonical variates analysis in Subsection 7.2.1).

7.1.5 Non-centred PCA

Principal component analyses starting from covariance and correlation matrices agree in that the data are centred and, as a consequence, the components intersect in the centroid of the original point swarm. However, centring is not absolutely necessary; PCA can also start from the cross product matrix **K** of the variables (Expression 3.68). This version is termed *non-centred PCA* in the literature[3]. The short list below will summarize its basic features:

- Non-centring implies that the components are forced to go through the origin of the axes of the raw data space. As a result of this constraint, the components are usually not orthogonal, and their correlations may considerably differ from zero.
- The amount of variability explained by a given axis is measured by the *sum of squared deviations* of points from the origin on that axis, rather than by variance. For the sum of eigenvalues, we have the following relationship:

3 In fact, there is a fourth combination of data manipulations in PCA, when non-centring is combined with division by standard deviation. This version is perhaps the least important in our practice, and is not treated here. .

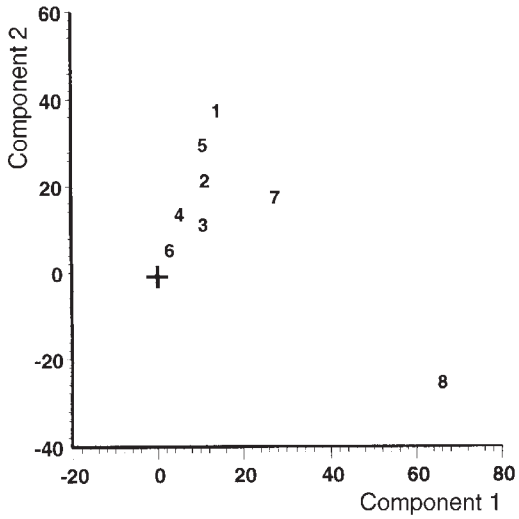

Figure 7.8. Non-centred PCA for the data of Table A1. The position of the origin is indicated by a + sign.

$$\sum_{h=1}^{t} \lambda_h = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}^2 = \text{tr}\,\{\mathbf{K}\}\,. \qquad (7.13)$$

Thus, the quantity $100\lambda_h / \text{tr}\{\mathbf{K}\}$ will be the percentage of the total sum of squares (the sum of diagonal values of **K**) falling to component *h*.

- Since the eigenvalues are interpreted in terms of sum of squares, the component correlations can only be obtained by direct calculations using Formula 3.70, that is, by comparing the original variables with the component scores. A biplot can also be drawn, but its interpretation is less obvious, mostly because the axes are not orthogonal.

When looking at the above list, one may ask somewhat precariously: is there any specific question that may be addressed by non-centred PCA only? As we have said, the essence of the approach is the maximization of sum of squared deviations from the origin on each axis. This fact may be utilized in explaining the ordination of ecological sampling units (quadrats, plots) as follows. The more dominant are some species on account of the others in a plot, that is, the lower the *diversity*, the larger the sum of squares pertaining to this plot (that is, sum of squares is inversely proportional to the Simpson diversity). As a reflection to this, in non-centred PCA ordinations the plots with low diversity (high deviations) will fall far apart from the origin, whereas those with high diversity (lower deviations) will concentrate near the origin. This gives an immediate possibility for interpretation in terms of diversity (Carleton 1980, ter Braak 1983, Digby & Kempton 1987). However, our conclusions are correct only if the data comprise percentage cover or, preferably, relative dominance scores (standardization by the sum of each object, Formula 2.20). Raw adundances cannot be used directly, because the total number of individuals can differ remarkably over plots. Another potential utility of non-centred PCA is its sensitivity to the grouping of objects. When there is a clear-cut group structure in the data, each component will have high scores for one group only, while the scores of the members of the other groups will be low (Pielou 1984). As a confirmation, it is always advisable to compare non-centred ordinations with distance-based classifications.

Let us examine the results of the non-centred PCA of Table A1. In the ordination of objects (Fig. 7.8), site 6 is the closest to the origin because it has the most even cover scores of species (there is one relative extreme, 12% cover for *Fumana procumbens*). Moving away

from the origin the within-plot diversity decreases because within-plot differences among species increase, which is most conspicious in sample plot 8.

7.1.6 The 'arch-effect' and its role in identifying underlying gradients

The conditions of applicability of different multivariate methods are often forgotten or, more so, misinterpreted by various authors. For example, some discussions of component analysis implicitly suggest that the data should follow multivariate normality and if this requirement is violated, the analysis should not be performed at all. In fact, however, the success of PCA does not depend on multivariate normality, the distribution of variables may 'even be even'! (It is a different matter, of course, if some statistical test of the number of meaningful components assumes normality). The interpretation of PCA is usually rendered more difficult by another problem, the potential existence of non-linear relationships among the variables. As an illustration of non-linearity, let us consider the following data matrix containing 7 objects (rows) described by 9 variables (columns):

$$\begin{matrix} 1\,3\,1\,0\,0\,0\,0\,0\,0 \\ 0\,1\,3\,1\,0\,0\,0\,0\,0 \\ 0\,0\,1\,3\,1\,0\,0\,0\,0 \\ 0\,0\,0\,1\,3\,1\,0\,0\,0 \\ 0\,0\,0\,0\,1\,3\,1\,0\,0 \\ 0\,0\,0\,0\,0\,1\,3\,1\,0 \\ 0\,0\,0\,0\,0\,0\,1\,3\,1 \end{matrix} \qquad (7.14)$$

The data comprise regular changes because the objects are gradually transformed into one another: when going down row by row in the matrix one variable disappears, a new one occurs, and the values of two variables are swapped in every step. We are tempted to say that there is in fact a single underlying gradient which influences the variables and, in turn, the objects. In ecology, for example, some environmental factors, such as elevation above sea level, may exhibit such a behavior. The variables (species) respond to this background gradient in different ways, illustrated by the so-called *response curves*. Fig. 7.9a shows an hypothetical and idealized case, an oversimplification of which is matrix 7.14. Anyone who expects that the PCA ordination of sampling units taken along the gradient at equal intervals will arrange the corresponding points along a line, thus reconstructing the background gradient, will be disappointed by looking at the PCA results. The standardized PCA of these data places the points into an *arch* (or *horseshoe*) in the space of the first two components (Fig. 7.9b). The relative importance of eigenvalues does not deacrease abruptly, which would indicate the presence of the single background factor. Instead, the eigenvalues diminish only gradually. We are tempted to say therefore that the method itself is responsible for the 'distortion' of an otherwise obvious and expected arrangement. In other words, PCA appears to produce an 'artefact'.

The explanation has long been known in the literature of ordination (Kendall, 1971, was the first to use the term horseshoe effect).When we examine the reasons behind this 'effect', we find that there is no distortion or artefact at all. Taking the gradient as a whole, the relationships among the nine species turn out to be far from being linear. For example, whereas the abundance of species 1 decreases, the abundance of species 2 first increases, then decreases and, finally, both reach the zero level. We find similar relationships for other species pairs as well. In addition, there are species pairs which cannot even be compared: while one is changing the other remains zero and vice versa. As a result of these complex relationships, the

species occurring at the beginning of the gradient are fully replaced by others by the fourth object. That is, the distance between objects 1 and 4 reaches the possible maximum, which cannot increase any further for objects 1 and 5 or the subsequent quadrats, because there is simply no possibility for further species replacement. As a matter of fact, this is represented by the ordination quite faithfully, because PCA 'attempts' to keep the distances from 1 to 4, 5, 6 and 7 to be the same. With some experience, the arch effect is easy to recognize in ordination scattergrams, raising the possibility of non-linear data structures. To sum up, the method does not introduce any distortion, the result is an obvious consequence of non-linearity of the data.

Contrary to some 'pessimistic' views, the presence of an arch in the arrangement of points does not spoil the interpretation of the PCA ordination. Nevertheless, many authors take the view that even though the species respond to a single background gradient in a non-linear fashion, the points in the ordination should fall approximately onto a line, thus facilitating the recognition of that gradient. They suggest automatic 'detrending' procedures to build in into the algorithm of ordinations. Phillips (1978) proposed a procedure which fits a parabola to the points in dimensions 1 and 2, and then straightens the parabola to provide new coordinates for points for dimension 1. (Actually, a third-order polinom expresses the relationship between axes 1 and 3, a fourth-order polinom between axes 1 and 4, and so on, hence the name, *polinomial ordination*.) Regression to a parabola and its straightening may prove spectacular in certain cases, thus giving a better impression of a single underlying gradient. The problem with automatic detrending is that the presence of a gradient, to which the variables respond

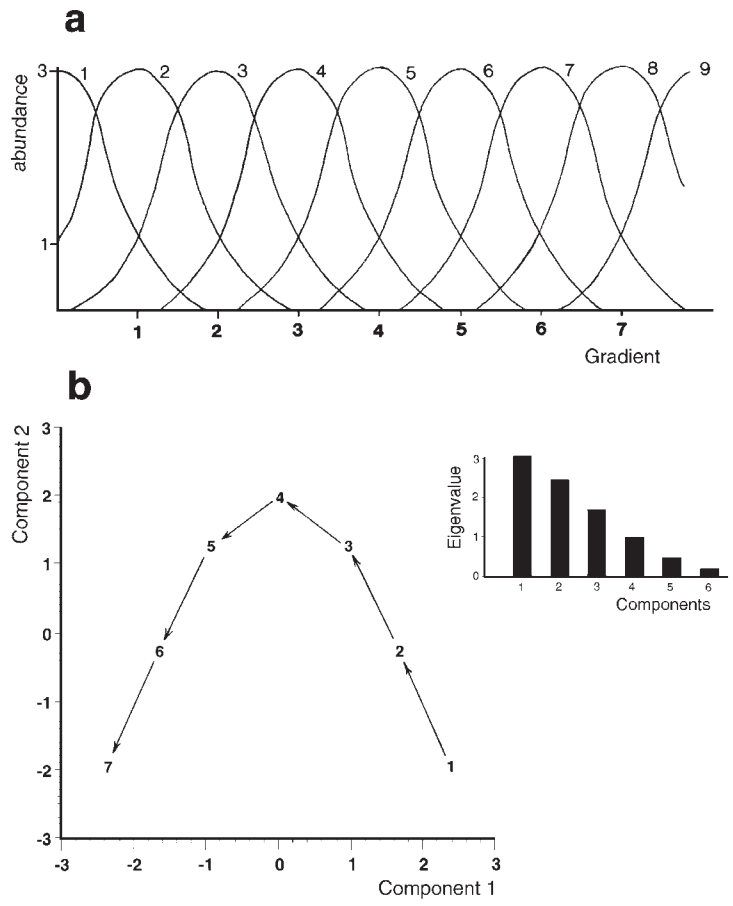

Figure 7.9. The arch effect. **a:** The response of the 9 variables of matrix 7.14 to the background gradient is non-linear, therefore their correlations are non-linear either. **b:** As a result, the seven objects observed along the gradient will be arranged along an arch (a horseshoe) in the PCA ordination. The bar diagram (inset) shows the gradual decrease of eigenvalues.

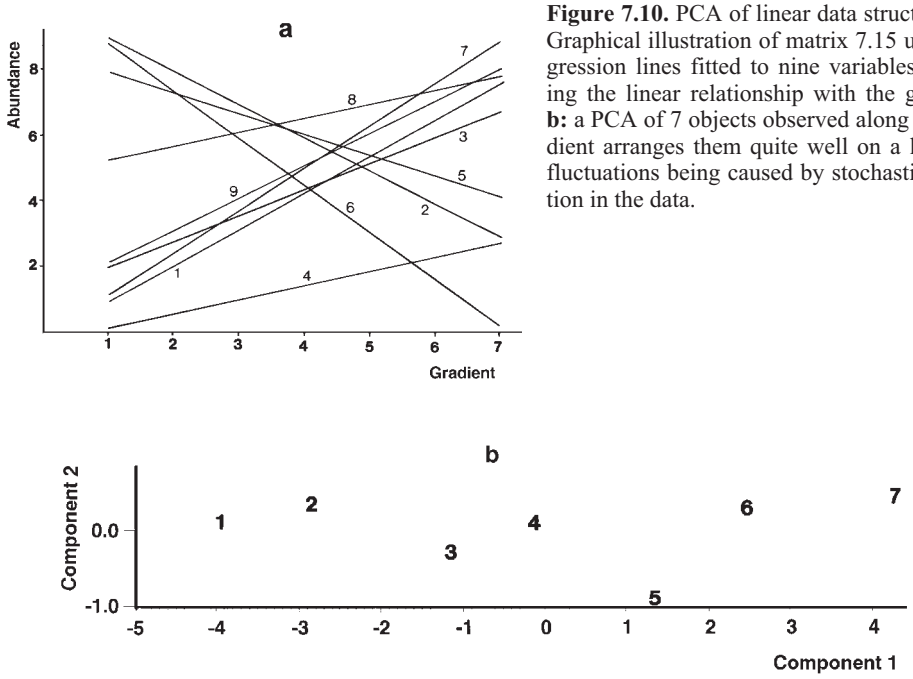

Figure 7.10. PCA of linear data structures. **a:** Graphical illustration of matrix 7.15 using regression lines fitted to nine variables, showing the linear relationship with the gradient. **b:** a PCA of 7 objects observed along the gradient arranges them quite well on a line, the fluctuations being caused by stochastic variation in the data.

according to an optimum curve, is presupposed in the data. We do not know therefore what would happen otherwise, without detrending. Logic dictates that a plain PCA should be performed first and then, if there are signs of non-linearity manifested by arched arrangements, we utilize detrending to clarify the picture. In many instances, the user will find that detrending does not add anything new to the interpretation of PCA (or other, see later) ordinations.

Arch effect is encountered in ecological ordinations whenever the speed of changes along the gradient (‘*species turnover’,* β*-diversity,* Whittaker 1967) is high. In such cases, an alternative to detrending is the so-called minimum path adjustment. It is discussed later (Subsection 7.4.1) because the procedure applies to cases when distances are measured among the objects.

The relationship between variables can be approximately linear for objects observed only in a short segment of the background gradient. Then, the points representing the objects will be arranged almost linearly. This is illustrated by the following 7 \ 9 data matrix:

$$\begin{array}{l} 1\,9\,2\,0\,8\,9\,1\,5\,2 \\ 2\,7\,3\,0\,8\,7\,2\,5\,3 \\ 3\,6\,3\,1\,7\,5\,4\,6\,4 \\ 4\,6\,4\,1\,6\,3\,5\,6\,5 \\ 5\,5\,4\,2\,6\,1\,7\,7\,6 \\ 6\,4\,6\,2\,5\,1\,8\,7\,7 \\ 8\,3\,7\,3\,4\,0\,9\,8\,8 \end{array} \qquad (7.15)$$

The nine variables (colums) change linearly over the gradient, and a line can be fitted to each of them (Fig. 7.10a). This implies that there are high positive and negative linear correlations among the variables, although they are diminished by some ‘noise’ arbitrarily introduced into the data. Standardized PCA performs ‘ideally’ in such cases. The effect of the underlying gra-

dient manifests itself in component 1 which explains 96% of the total variance. There is some noise, so that the points do not fit perfectly the regression line, but the deviation from linearity is a tiny fraction of the variance, merely 4% (Fig. 7.10b).

The components, particularly when there is no horseshoe in the configuration, are often identified *a posteriori* as external variables that were not included in the analysis. In ecology, such variables are usually environmental factors. The linear correlation calculated between components and external variables may help us interpret the results of PCA. Such a detection of ecological gradients is termed *indirect gradient analysis* by ter Braak & Prentice (1988), because the gradient is revealed by using the species data, rather than incorporating environmental variables. The alternative strategy, *direct* gradient analysis utilizes information conveyed by environmental variables (see Subsections 7.2.5 and 7.3.5).

7.1.7 Factor analysis

Some space needs to be devoted to *factor analysis* (FA), a technique closely related to PCA in its algorithmic implementations, yet radically different in its conceptual foundations. This short discussion is inevitable, because PCA and FA are far too often confounded in the biological literature. An obvious evidence of confusion is when PCA axes are directly interpreted and called as factors, thus leaving the reader in doubt as to the real nature of the ordination performed in that study.

The most essential difference between PCA and FA is that while the components are used to explain as high a percentage of the *total* variance as possible, the factors are extracted to account maximally for the *co*variances (the shared variances) of variables. Thus, they are responsible for the relationships among variables only. Factor analysis usually standardizes the data before computations, so that in effect the correlation structure is explained by the new axes. The portion of the total variance falling to the individual variables cannot be explained by the common factors. This is the *specific variance* of variables, caused by the *specific* or individual factors (there are *n* such factors). Since specific variances are ignored, the common factors explain lower percentages than the components for a given data set. To sum up: factor analysis attempts to reveal the correlation structure of variables, whereas the ordination of objects (observations) is of secondary importance and, very often, is not even illustrated. Therefore, the role of FA in biological data analysis is relatively small[4]. Without entering into the algorithmic details of factor analysis, it is useful to list some of its general features that make it very distinct from other ordination procedures disccussed in this book.

- The variance shared by the variables is explained by a pre-specified number of factors, *p*. Since *p* is defined arbitrarily, i.e., the model of FA is modified with much freedom by the investigator, many authors consider factor analysis as an ‘artistic activity’, rather than an objective multivariate statistical method, and therefore do not recommend it for data exploration (e.g., Kendall 1975, Chatfield & Collins 1980). Others (Jolliffe 1986) take the view that the question whether other ordination procedures are

4 Factor analysis is most popular in humanities and psychology. A few decades ago it was one of the most widely applied multivariate methods in biology as well.

better than FA is irrelevant, because FA can provide valuable information on data structures for the researcher.

- The factors, like components, are uncorrelated in the basic FA model. The correlation between factor *i* and variable *j* is denoted by a_{ij}, and is called the *factor loading* of that variable. These correlations provide the coordinates for displaying the ordination of variables (similarly to the component correlations in PCA). The orthogonality of axes is not a rule, however, and the factors can also be extracted such that they have non-zero correlations (Cattell 1978 mentions examples when this may be meaningful).
- The model assumes that the 1-s in the diagonal of the correlation matrix, the 'self-correlations', are cumulative results of the influence of common factors and the specific factors. In other words,

$$r_{jj} = 1 = \sum_{i=1}^{p} a_{ij}^2 + e_j \qquad (7.16)$$

  where the sum of *a* values indicates the effect of common factors and is called the *communality* of variable *j*, and the quantity e_j stands for the specific variance of the same variable. The larger the value of *p*, the higher the share of the communality from the unit correlation. Ultimately, when *p* equals the rank of the correlation matrix, the specific variance reduces to zero, so that FA becomes identical to PCA. Nevertheless, PCA and FA often produce similar ordinations of variables, suggesting that specific variances are relatively low.
- The best-known algorithm of FA is the *principal factor method*. This is implemented as an iterative procedure in which each step is a standardized PCA. At the outset, some estimation is made for the communalities and the iterations stop when the communalities do not change more than a pre-specified threshold.

7.2 Two groups of variables: canonical correlation analysis

In principal components analysis, all variables are treated assuming that they represent a single logical group, for example, when all variables are cover scores of species, or all are morphological characters. There are situations, however, when this practice is less acceptable beause the variables form two logically separate groups. Ecological sampling often leads to a data set in which the sampling units are described in terms of species abundances as well as some environmental measurements. Although these variables could be lumped together in a plain PCA, such an ordination would not be able to reveal the relationships between variable groups. Instead, the ecologist may be interested in evaluating the relationship between environmental and biological variables as groups. In other words, the question whether species data are predictable from the environmental variables and vice versa is considered. In systematic studies, the taxonomist may also want to evaluate the correspondence between two groups of variables, one describing the taxa in their larval stage, and the other referring to the adult stage. This kind of data exploration is facilitated by a derivative of PCA, the so-called *canonical correlation analysis* (CCA; in other sources, e.g., Jongman et al. 1987, the abbreviation COR is preferred).

The method was developed by Hotelling (1936). The main idea is that the linear combinations of original variables are searched for in the groups separately, with the constraint that these linear combinations are maximally correlated. To put it differently: CCA can be conceived as a double principal component analysis supplemented by finding the best fit of axes obtained for the two groups. The new axes extracted by CCA are the *canonical variates,* and the correlation between the variates, one coming from the first group of variables and the other from the second, is the *canonical correlation*. There are several pairs of such axes, they are determined from the data in order of importance. While PCA attempts to explain the total variance effectively, CCA maximizes the covariance between two groups of variables (Cooley & Lohnes 1971, Gittins 1979).

The graphical illustration of CCA helps understand the principle of the method (Fig. 7.11). Assume that both variable groups include two variables, abbreviated by x_1 and x_2, as well as by y_1 and y_2, respectively. If a PCA were applied to the two subsets of data sperately, then the first components would run through the points in the manner illustrated by thin lines (we can forget about the next component, for simplicity). However, the coordinates of objects on the component of the left diagram do not correlate maximally with the corresponding coordinates in the right diagram. Both axes have to be rotated a little bit (thick lines in Fig. 7.11) to attain the highest possible correlation between these axes.

The computations of CCA start from the correlation matrix of variables, which is subdivided to four submatrices according to the variable groups:

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{R}_{21} & \mathbf{R}_{22} \end{pmatrix} . \qquad (7.17)$$

$\mathbf{R}_{11}$ contains the correlations between the n_1 variables in the first group (*left domain*), $\mathbf{R}_{22}$ is a submatrix of correlations between the n_2 variables in the second group (*right domain*), whilst $\mathbf{R}_{21}$ and its transpose $\mathbf{R}_{12}$ summarize cross-correlations between the groups. In CCA, we wish to maximize the between-group correlations and minimize the within group correlations. Since their ratio does not exist in matrix algebra, we use the inverse of within-group submatrices to derive a correct formulation, given by the multiple $\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$. Then, the resulting matrix is subjected to eigenanalysis by solving the following characteristic equation:

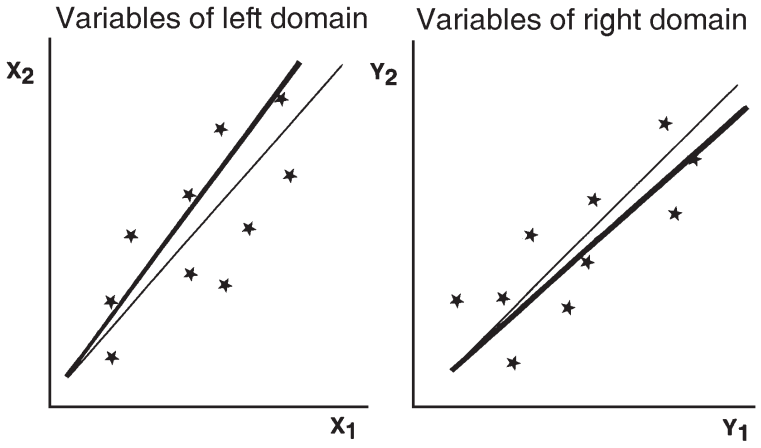

Figure 7.11. Illustration of canonical correlation analysis. The components obtained separately for the two variable groups (thin lines) very rarely coincide with the canonical variables (thick lines).

$$(\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12} - \lambda_j \mathbf{I})\ \mathbf{v}_j = \mathbf{0}\,. \tag{7.18}$$

The solution is not as simple as in case of PCA, because the matrix $\mathbf{R}_{22}^{-1}\mathbf{R}_{21}\mathbf{R}_{11}^{-1}\mathbf{R}_{12}$ is not symmetric. The eigenvectors $\mathbf{v}_j$ are computed such that their length is unity. Then, the following normalization provides the canonical weights for the second group of variables on canonical variate j:

$$\mathbf{c}_j = \frac{\mathbf{v}_j}{\left(\mathbf{v}'_j\ \mathbf{R}_{22}\ \mathbf{v}_j\right)^{1/2}}\,. \tag{7.19}$$

As a result of this transformation, the variance of the variates becomes 1. The weights for the first group of variables on the same canonical variate are obtained as:

$$\mathbf{b}_j = \frac{(\mathbf{R}_{11}^{-1}\ \mathbf{R}_{12}\ \mathbf{c}_j)}{\sqrt{\lambda_j}} \tag{7.20}$$

The canonical weights multiplied by the original data provide the coordinates of objects for the first (left) domain

$$\mathbf{T} = \mathbf{X}'\mathbf{B}, \tag{7.21}$$

and for the second (right) domain of variables

$$\mathbf{U} = \mathbf{Y}'\mathbf{C}\,. \tag{7.22}$$

In these matrix equations, $\mathbf{X}'$ is the upper portion of the data matrix with a size of $m \setminus n_1$ and $\mathbf{Y}'$, with a size of $m \setminus n_2$ is the lower portion of the data matrix. Recall that the objects are the rows and variables are the columns of the original data matrix. All variables are centred and standardized to unit variance beforehand. Matrices **B** and **C** contain the canonical weights as determined by Equations 7.19 and 7.20; their sizes are $n_1 \backslash q$ and $n_2 \backslash q$, respectively (the meaning of q is clarified right below).

7.2.1 Canonical correlations and their significance

The number of positive eigenvalues in CCA is $q = \min\{n_1, n_2\}$, provided that $m > n_1, n_2$. The square roots of the eigenvalues measure the *canonical correlations* between the variable groups:

$$|R_j| = \sqrt{\lambda_j}\,. \tag{7.23}$$

The are p different canonical correlations which can be arranged in descending order. The absolute sign indicates that the canonical correlations range between -1 and 1 like usual correlations, but the sign cannot be determined from the analysis. Therefore, convention dictates that the linear correlation between $\mathbf{b}_j$ and $\mathbf{c}_j$ is measured by the absolute value.

If the observations are independent and the condition of multivariate normality holds, then the difference of canonical correlations from zero can be tested statistically. Bartlett's Λ (cf. Cooley & Lohnes 1971) can be used to test the simultaneous significance of several canonical correlations such that the first 0, 1, 2, ..., k, ..., q canonical variates are removed:

$$\Lambda = \prod_{j=k+1}^{q} (1-\lambda_j)\,. \tag{7.24}$$

Bartlett has shown that Λ follows the χ^2-distribution after the following transformation:

$$X^2 = -[m - 1 - 0{,}5(n_1 + n_2 + 1)]\ \ln\ \Lambda \tag{7.25}$$

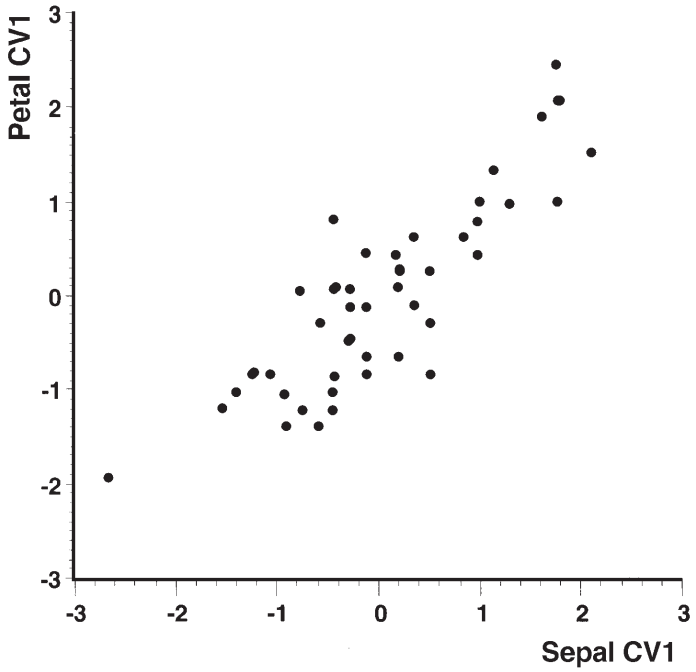

Figure 7.12. Canonical correlation analysis of 50 individuals of *Iris virginica.* The ordination depicted in this figure arranges the individuals along the first canonical variate obtained for the sepals (horizontal axis) and the first canonical variate for the petals (vertical axis).

at $(n_1 - k)\,(n_2 - k)$ degrees of freedom. For example, when the test indicates significance for $k = 0$ but not for $k = 1$, then we can say that only the first canonical variate is meaningful statistically.

There are potential dangers with the above significance test, as pointed out by Gittins (1979, 1985). He warned that caution is needed even if both conditions of the test are satisfied. A more appropriate procedure is the direct statistical analysis of the eigenvalues, but this assumes knowledge of the distribution of maximum eigenvalues that can be obtanined from the data.

Canonical correlation analysis allows several ordination diagrams to be made. The most common practice is to construct a two-dimensional configuration such that the horizontal axis is the first canonical variate from the left domain, whereas the vertical axis is the first variate from the right. If the canonical correlation (Formula 7.23) is high, then the points will fall close to a 'diagonal' line. Weaker inter-domain relationships are indicated by more scattered arrangements. Another interpretive tool is the ordination of objects for the two most important canonical variates for each group. This is recommended particularly if both of the first two canonical correlations are high and significant. The correlations among variables and canonical variates can also be demonstrated in these ordinations, in a similar manner to PCA biplots.

The CCA of the third *Iris* species of Table A2 (*I. virginica*) illustrates what has been said. This example is deliberately simple; the length measurements of the sepals comprise the left domain, whereas the petal variables constitute the right domain. The problem is to reveal the relationship between sepal and petal measurements, notwithsanding that these variables can be treated together logically. Fig. 7.12 is the CCA ordination of iris individuals on the first canonical variate from each domain. The fit of points to a 'diagonal' line is very close, indicating the relatively strong relationship between the two groups of variables ($R_1 = 0.86$, $\chi^2 = 75.9$, $p \ll 0.001$). Note that the second canonical variates are also significant ($R_2 = 0.47$, $\chi^2 = 12.0$, $p < 0.001$). The results are similar for *Iris versicolor,* although R_1 is 'only' 0.76. The

analysis of the third species, *Iris setosa,* provides a different result; both canonical correlations proved to be low (e.g., $R_1 = 0.32$), and none of the canonical variates was significant. A straightforward conclusion is that for this species the sepal data are not predictive as to the petal measurements and vice versa. Another interesting point is that the canonical correlation analysis of all species taken together would have confounded the differences between species. Actually, this joint CCA of the species yields *one* significant canonical corrrelation (cf. Podani 1994), reflecting that on the genus level the petal and sepal measurements are strongly correlated.

7.2.2 Correlation with the original variables

There are two kinds of correlation between canonical variates and the variables in CCA, as discussed below.

Within-group ('structure') correlations. Measuring the contribution of each variable to the canonical structure in the group is a useful interpretive aid in CCA. Variable *i* from the left domain, represented by vector $\mathbf{x}_i$, and canonical variate *j* derived for this domain have the following correlation:

$$\rho(\mathbf{x}_i, \mathbf{b}_j) = \sum_{k=1}^{n_1} r_{ik} b_{kj} \,. \tag{7.26}$$

Similarly, the correlation of variable *i* in the right domain, given by vector $\mathbf{y}_i$, and the canonical variate *j* extracted from this domain is obtained as

$$\rho(\mathbf{y}_i, \mathbf{c}_j) = \sum_{k=1}^{n_2} r_{ik} c_{kj} \,. \tag{7.27}$$

The r_{ik} values in both equations correspond to the usual linear correlations between the original variables. Formulae 7.26-27 can be used to identify variables that represent their own group most markedly when compared to the other domain.

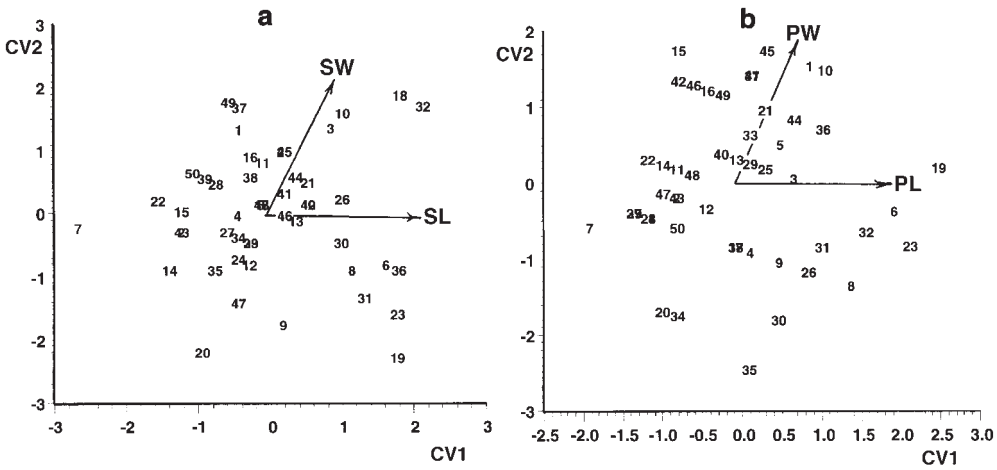

Figure 7.13. CCA of 50 individuals of *Iris virginica* based on **a:** sepal measurements, **b:** petal measurements. Arrows indicate the strength of correlations between original variables and canonical variates in a way analogous to the correlation-based Rohlf-biplot in PCA.

The CCA of *Iris virginica* data reveals that in both variable domains the length measurements are correlated most strongly with the first canonical variate. They are almost 'identical', because the structure correlations are close to 1.0. The second canonical variate in both groups, as expected, has high correlations with the width measurements (0.88 and 0.94). These are best illustrated using biplot-like diagrams prepared separately for the first two canonical variates in both groups. The coordinates of objects are obtained by Formulae 7.19-20, while the coordinates of variables are determined using equations 7.26-27. The latter are multiplied by an arbitrary scale factor for commensurability with the object scores (Fig. 7.13a-b). Comparison on strictly visual basis shows that there is considerable agreement between the two ordinations, although the large number of points and the overlaps confuse the picture. More possibilities for preparing CCA biplots are discussed exhaustively by ter Braak, 1990.

Between-group (interset) correlations. Another possibility to describe canonical results is in terms of the correlations between the variables of one domain and the canonical variates derived from the other (Gittins 1979). Formally, variable *i* of the left domain has the following correlation with canonical variate *j* from the right domain:

$$\rho(\mathbf{x}_i, \mathbf{c}_j) = \sum_{k=1}^{n_2} r_{ik} c_{kj} \tag{7.28}$$

Analogous to this is the calculation of the correlation between variable *i* of the right domain and canonical variate *j* from the left domain:

$$\rho(\mathbf{y}_i, \mathbf{b}_j) = \sum_{k=1}^{n_1} r_{ik} b_{kj} \tag{7.29}$$

In the above two formulae, the r_{ik} values are the correlations from the $\mathbf{R}_{12}$ submatrix. The square of a between-group correlation is the proportion of the variance of the variable which is explained by the canonical variate from the other group. These correlations are not suitable to graphical display.

The evaluation of *Iris virginica* measurements indicates that the variable most predictable by a canonical variate from the other group is the length of sepals. As much as 75% of the variance of sepal length is accounted for by the canonical variate from the petal variables. The same percentages for the width measures do not exceed 18%.

7.2.3 Variance and redundancy

The possibilities for evaluating CCA results are not yet exhausted. We can also calculate the proportion of *variance* of a given group explained by its own canonical variate. This is in fact the average of the squared within-group correlations. For the left domain, the variance proportion explained by canonical variate *j* is thus:

$$100 \sum_{i=1}^{n_1} \rho(\mathbf{x}_i, \mathbf{b}_j)^2 \,/\, n_1 \tag{7.30}$$

The percentages for the right set are derived in a similar way:

$$100 \sum_{i=1}^{n_2} \rho(\mathbf{y}_i, \mathbf{c}_j)^2 \,/\, n_2 \tag{7.31}$$

For the *Iris virginica* data, these variance proportions are 61% and 39% for the left set (sepal measurements) and 55% and 45% for the right set (petals). The figures illustrate appropriately the contrast between canonical variates and components. In the PCA of sepal data,

component 1 explains 74% of the total variance, leaving 26% for the second. This is, of course, a much more efficient variance extraction than 61% plus 39%. As emphasized above, however, this is not the primary objective of CCA.

The proportion of variance of one domain, explained by a canonical variate from the other is called the *redundancy*, a between-group analogue of variance (Gittins 1979). It is determined as the average of squared between-group correlations. The variance of the left domain is explained by the canonical variate *j* from the right group to the extent given by:

$$100 \cdot \sum_{i=1}^{n_1} \rho(\mathbf{x}_i, \mathbf{c}_j)^2 \,/\, n_1 , \tag{7.32}$$

whereas the redundancy in the other way is computed as:

$$100 \cdot \sum_{i=1}^{n_2} \rho(\mathbf{y}_i, \mathbf{b}_j)^2 \,/\, n_2 . \tag{7.33}$$

The sum of redundancy values over *j* is the *total redundancy* of one domain with respect to the canonical variates of the other. This is the proportion of total variance which is explained by the other group of variables. Total redundancy is an asymmetric measure, being usually different for the left and the right domains.

This is the case in the *Iris* example. The amount of variance of the two sepal variables explained by the canonical variates of the petals is 55%, whereas this percentage in the reverse direction is merely 51.1%.

7.2.4 Notes on the applicability of CCA

Many authors agree that the interpretation of CCA results is more problematic than any other ordination in multivariate analysis. Bock (1975) points out that the canonical variates are results of a compromise; given the condition of orthogonality the between-group covariance is maximized such that within-group variance is minimized (see what I have said under Equation 7.31). As a consequence, the canonical variates and principal components rarely, if ever, agree. This incongruence can go so far that a canonical variate explains a small fraction of the total variance of the group, thus having no interpretive value (Rohlf, in: Legendre & Legendre 1983, p. 330). In such cases, the canonical correlation expresses a relationship which has no meaning (cf. Pimentel 1979). This problem is circumvented if PCA is performed for one group of the variables, and then the component scores are involved in CCA (e.g., Digby & Kempton 1987, p. 82, or Ludwig & Reynolds 1988, p. 299). Alternatively, both groups are analyzed by PCA only (Williams & Lance 1968, Shaukat & Uddin 1989). Another advantage of these separate PCAs is that potential singularity problems with the $\mathbf{R}_{11}$ or $\mathbf{R}_{22}$ matrices are automatically solved (when the number objects is smaller than the number of variables, then these matrices cannot be inverted). As an ultimate solution, standard CCA can be performed on original variables and PCA scores as well, and the results thus obtained are compared. In this book, this comparative approach is not illustrated.

7.2.5 Redundancy analysis

In the CCA of *Iris* variables, the relationship between the two groups was assumed to be symmetric; neither direction of comparison could be favoured. In many ecological investigations, however, no such symmetry is plausible becasue species respond to environmental variables, which is not so in the other way! There is no symmetry in predictability, contrary to what I suggested briefly in the introduction to CCA for mere didactic reasons. In ecological studies, it were illogical to determine canonical variates such that linear combinations of variables are found in both groups of variables and their correlations maximized (as done in CCA). This would be hard to interpret. For species and environmental variables, it is more reasonable to extract ordination axes such that they reflect the environmental variables, because they influence the performance values of species (cover, biomass etc.). Thus, let the ordination axes be linear combinations of environmental variables which, at the same time, explain as much as possible of the variance in the species-based ordination of objects (sites, sample plots). In other words, the objects are ordinated as in PCA with the requirement that the components maximally interpret the environmental variables as well. The method of *redundancy analysis* (RDA), developed by Rao (1964) will do it for us. RDA was introduced to ecological data analysis by ter Braak & Prentice (1988), as a case of direct gradient analysis. Since the axes are not fitted freely to the objects, but are also determined by the environmental variables, ter Braak & Prentice suggested the term *constrained* ordination (RDA is thus a *constrained PCA*). It has relatively few applications, mostly because the constrained strategy of canonical correspondence analysis (Subsection 7.3.5) has become more popular in ecologcial data analysis (Birks et al. 1996).

> The strategy of RDA is in fact a canonical correlation analysis in which correlations between the species are neglected. The interspecific correlations convey no directly useful information to us, contrary to the within-group correlations of environmental variables, and the between-group correlations of environmental variables and species. If the species data are summarized in submatrix **X**, then $\mathbf{R}_{11}$ is to be replaced in Equation 7.18 by the identity matrix **I** (its diagonal is filled up with 1-s, and all other values are zero).

7.3 Correspondence analysis

We have seen that the biplots in PCA or CCA reveal the relationships between objects and variables quite efficiently. For both methods, however, the ordinations of objects and variables are obtained separately, and only afterwards are they superimposed to each other using a rescaling procedure or some less elegant ‘tricks’. One might feel that this is an unwise and cumbersome strategy and seek a procedure that finds the optimum fit of the two ordinations directly and simultaneously. There is an ordination method, namely *correspondence analysis*, that is designed to fulfil this need; the term correspondence referring to the requirement that the positions of objects and variables in their joint ordination should mutually correspond to one another. Many variants of the method have long been used in various fields of science and humanities and, as a result of this diversity, they have been named very differently. For example, it can be shown that reciprocal averaging, dual scaling, contingency table analysis and other procedures (cf. Legendre & Legendre 1983) are in fact derivatives of a method proposed and popularized under the term *l’analyse des correspondances* by the French school

(Benzécri et al. 1973). Its translation to English, correspondence-analysis (COA[6]) is now considered as a collective term referring to all algorithmic variants of the original method.

7.3.1 An intuitive example and reciprocal averaging

The essence of COA is most easily understood if we consider the following simple one-dimensional ecological example. Assume that the effect of an ecological background gradient, such as a moisture gradient, is to be revealed indirectly on the basis of species data derived from sample plots (phytosociological relevés or quadrats). The species occurring in the plots can be arranged according to their moisture requirement into a point system ranging from 0 to 10. (These values or *weights* are determined empirically by earlier investigations.) For each plot, the abundance of each species is multiplied by its weight, and then these products are summed over all species which, in turn, is divided by the total abundance in the plot to compensate for excessive abundance differences in the sample. Thus, the more dominant are the drought-tolerant species in a plot, the lower its score will be and, conversely, high plot score will refer to the dominance of species with high moisture requirement. Based on these scores, the plots can be arranged along an ordination axis which may be a good approximation to the actual humidity gradient that characterizes the sites. This is the essence of the *weighted averaging* approach proposed by Whittaker (1967). However, we need not stop at this point, because the positions of plots can now be used to refine the 0-10 scale of species further. The refined new value is obtained by weighted averaging again: for each species the coordinate of every plot on the ordination axis is multiplied by the abundance of the species, the scores are added and then divided by the total abundance of the species. Afterwards, standardization is necessary to rescale the new values so that they have unit variance and zero mean. Using the new weights thus obtained, an improved ordination of plots is derived. From this, new weights are calculated for the species, and so on, the calculations being continued until the changes of weights between two subsequent steps do not exceed a prespecified threshold value. The final result of these iterations is a one-dimensional ordination of plots and species in which the species positions correspond to the quadrat positions and vice versa, as perfectly as possible. This iterative strategy has been known as *reciprocal averaging* (RA, Hill 1973) in the literature of numerical ecology.

The calculations involved in RA are summarized as follows. Let the rows of the data matrix $\mathbf{X}_{n,m}$ be species (variables) and the columns be plots (objects). The coordinates of plots along the first axis are determined as follows:

$$b_j = \frac{1}{\lambda^{1-\alpha}} \sum_{i=1}^{n} a_i \frac{x_{ij}}{u_j} \tag{7.34}$$

where $u_j = \sum_{i=1}^{n} x_{ij} = x_{.j}$ is the total with respect to plot j and $1/\lambda^{1-\alpha}$ is the rescaling parameter mentioned above with $\alpha = 0.5$ (we shall see other possibilities later on). Furthermore, a_i is the weight of species i on the same axis. Consequently, the coordinates of objects are derived from the weighted relative contributions of variables. The standardization by the total of each object is implied in the algorithm of RA. To derive coordinates for the variables, we use a similar equation:

6 Other authors prefer the abbreviation CA, which is often used for *cluster analysis* in the literature of multivariate analysis. These acronyms are therefore without any universal meaning and their use may only be consistent within the same publication. I shall stick to COA in this book to make a clear distinction from cluster analysis.

$$a_i = \frac{1}{\lambda^{\alpha}} \sum_{j=1}^{m} b_j \frac{x_{ij}}{t_i} \qquad (7.35)$$

in which $t_i = \sum_{j=1}^{m} x_{ij} = x_{i.}$. In other words, the position of a variable on the axis is determined from the weighted relative contributions of objects. In this, division by the total for the variable is implied.

The mutual relationship between the so-called *transitional formulae* 7.34-35 is thus fairly obvious: one equation is solved by using parameters determined from the other. Therefore, the solution can only be iterative. It has been shown that the iterations converge into a stable result irrespective of the starting weights of variables and that these starting values influence only the speed of convergence. The significance of these findings is that the method applies even though we have no information on the environmental requirements of the species at all – contrary to the simple method of weighted averaging –, and we can even start from a random sequence of objects. Equations 7.34-35 have several solutions, each corresponding to an ordination axis. When the solution for the first axis is stable, then its effect is removed and an orthogonal second axis is determined, and so on. The computational steps of RA are described in detail by Hill (1973) and Pimentel (1979).

7.3.2 Computational steps of correspondence analysis

Whereas the algorithm of RA demonstrates the objective of the analysis in a didactic way, the extraction of all axes and associated 'rescaling factors' is much more efficient via eigenanalysis. The reader less familiar with details of linear algebra can even skip this subsection. The matrix algebraic forms of the transitional equations are expressed as :

$$\mathbf{B} = \mathbf{U}^{-1}\,\mathbf{X}'\,\mathbf{A}\,\mathbf{R}^{-1}, \qquad (7.36)$$

$$\mathbf{A} = \mathbf{T}^{-1}\,\mathbf{X}\,\mathbf{B}\,\mathbf{R}^{-1}, \qquad (7.37)$$

where $\mathbf{U}^{-1}$, $\mathbf{T}^{-1}$ and $\mathbf{R}^{-1}$ are diagonal matrices with elements $1/u_j$, $1/t_i$ and $1/\sqrt{\lambda}$, respectively. Replacing 7.37 into 7.36 we obtain:

$$\mathbf{B} = \mathbf{U}^{-1}\,\mathbf{X}'\,\mathbf{T}^{-1}\,\mathbf{X}\,\mathbf{B}\,\mathbf{R}^{-2}\,. \qquad (7.38)$$

Since $\mathbf{U}^{-1} = \mathbf{U}^{-1/2}\,\mathbf{U}^{-1/2}$ and $\mathbf{T}^{-1} = \mathbf{T}^{-1/2}\,\mathbf{T}^{-1/2}$, the above formula can be rewritten as

$$\mathbf{B} = \mathbf{U}^{-1/2}\,\mathbf{U}^{-1/2}\,\mathbf{X}'\,\mathbf{T}^{-1/2}\,\mathbf{T}^{-1/2}\,\mathbf{X}\,\mathbf{B}\,\mathbf{R}^{-2} \qquad (7.39)$$

which, after some rearrangement, takes the following form:

$$\mathbf{R}^2\,\mathbf{U}^{1/2}\,\mathbf{B} = (\mathbf{T}^{-1/2}\,\mathbf{X}\,\mathbf{U}^{-1/2})'\,(\mathbf{T}^{-1/2}\,\mathbf{X}\,\mathbf{U}^{-1/2})\,(\mathbf{U}^{1/2}\,\mathbf{B}) \qquad (7.40)$$

In this, $\mathbf{U}^{-1/2}$ and $\mathbf{T}^{-1/2}$ are diagonal matrices with elements $1/\aleph u_j$ and $1/\aleph\, t_i$, respectively; $\mathbf{U}^{1/2}$ contains the $\sqrt{u_j}$ values in its diagonal. If we introduce the following symbols: $\mathbf{Z} = \mathbf{T}^{-1/2}\,\mathbf{X}\,\mathbf{U}^{-1/2}$, $\mathbf{V} = \mathbf{U}^{1/2}\,\mathbf{B}$ and $\Lambda = \mathbf{R}^2$, then Equation 7.40 simplifies to a form that has been known from the description of PCA (cf. Equation 7.6):

$$(\mathbf{Z}'\mathbf{Z} - \lambda\mathbf{I})\,\mathbf{v} = \mathbf{0}\,. \qquad (7.41)$$

That is, the λ-s of matrix Λ and the **v** vectors summarized in matrix **V** are the eigenvalues and eigenvectors, respectively, of the $\mathbf{Z}'\mathbf{Z}$ cross-product matrix. After having completed the eigenanalysis. the relationship $\mathbf{v} = \mathbf{U}^{1/2}\mathbf{b}$ can be used to derive coordinates of the objects:

$$\mathbf{B} = \mathbf{U}^{-1/2}\,\mathbf{V}\,. \qquad (7.42)$$

Then, the coordinates of variables, given by matrix **A**, are easily obtained using the 7.35 transitional equation.

7.3.3 Notes on the implementation of COA and the interpretation of its results

If one wishes to write a program for COA and then to evaluate the results correctly, then the following considerations should be kept in mind:

1) The **X** data matrix should be standardized by the *grand total* of all values of the matrix ($x_{..}$), that is, each value is modified according to:

$$x'_{ij} = x_{ij} / x_{..} \tag{7.43}$$

As a result of this operation, the grand total of new values will be 1. Obviously, such a standardization is meaningful only if the variables are of the same type (e.g., presence/absence or abundance of species) and there are no negative data. COA is not suitable to the analysis of non-commensurable variables, and to variables expressed on different measurement scales. The analysis treats the data matrix as being a large contingency table in which the rows and the columns represent logically comparable entities such that the scores themselves are frequencies or other data that can be interpreted as frequencies (such as percentage cover).

2) There is a *trivial solution* for the transitional formulae ($a_i = 1$, $b_j = 1$), which corresponds to $\lambda = 1$. This eigenvalue refers to the centroid of rows and the columns and, being present in all cases, has no practical significance. We can get rid of this superfluous dimension by centring performed prior to the analysis:

$$y_{ij} = x_{ij} - x_{i.}x_{.j} / x_{..} \tag{7.44}$$

Centring implies that from each value the expectation obtained from the row and colum totals is subtracted. This operation is well-known from conventional biometry in the calculation of the χ^2 statistics based on $2\backslash 2$ contingency tables. This 'coincidence' underlines the view that COA has in fact been designed to the evaluation of contingency tables.

The above operations are included simultaneously if the starting **Z** matrix is calculated using the following formula:

$$z_{ij} = \frac{x_{ij}x_{..} - x_{i.}x_{.j}}{x_{..}[x_{i.}x_{.j}]^{1/2}}. \tag{7.45}$$

Since the eigenanalysis is performed on **Z**′**Z**, the computations will be faster if the data are prepared for input such that number of columns is not larger than the number of rows (in this case will **Z**′**Z** be the smallest). This is absolutely 'legal', because the objects and variables are treated symmetrically by COA when α=0.5 (see below), obeying the principle of attribute duality (Subsection 2.1). No analogous flexibility is present in PCA or other ordination methods, however. When interpreting the results, one has to make sure what the rows and the columns are, of course.

3) The non-trivial eigenvalues are usually smaller than 1. Their square root is the *canonical correlation*

$$R_i = \sqrt{\lambda_i} \tag{7.46}$$

which expresses the mutual agreement or *correspondence* of object and variable coordinates. In the example, the canonical correlation measures the reliability of species to get the ordina-

tion of quadrats and vice versa: the efficiency of quadrats to derive an ordination of species on axis *i.* The larger R_i, the higher the correspondence between the two orderings. In the transitional formula with $\alpha = 0.5$, the reciprocal value of canonical correlation was in fact used.

The sum of eigenvalues is the χ^2 calculated for the data matrix taken as an $n \backslash m$ contingency table. Because all values were divided by the grand total previously, for the original data we obtain that

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \sum_{i=1}^{t} \lambda_i = x_{..} \sum_{i=1}^{t} \lambda_i \quad . \tag{7.47}$$

Regarding the number of positive eigenvalues, *t*, the explanation given for PCA needs to be somewhat rephrased: *t* is the number of orthogonal dimensions necessary to explain the deviations between actual and expected data values. Therefore, if every value in the matrix equals its expectation (that is, $\chi^2 = 0$), then all eigenvalues are zero. The size of eigenvalues allows preliminary conclusions to be made on how far the data matrix deviates from an unstructured, random data set.

4) The coordinates for rows and columns can be displayed in separate scattergrams, and their explanation is as usual. At the same time, however, the coordinates can be used to prepare a COA biplot as well, whose interpretation is not the same as that of PCA biplots. The difference is emphasized such that the simultaneous display of row and column coordinates in COA is termed the *joint plot* (cf. Oksanen 1987). Whereas in PCA biplots the direction and the relative length of arrows have interpretive value, in a COA joint plot the relative closeness of points representing objects and variables may also be informative, and the arrows may be omitted. However, the evaluation and interpretation of joint plots depend greatly on the control parameter α and the magnitude of eigenvalues. In the algorithm described above, the variables (rows of the data matrix) and the objects (columns) were treated symmetrically, because α was set to 0.5. This parameter, however, may be freely changed within the interval [0,1] (cf. ter Braak 1985), so that there is an infinite number of (slightly) different ordination results for the same set of data! Of these, two other settings merit particular attention, especially in ecological ordinations with species as variables and quadrats as objects.

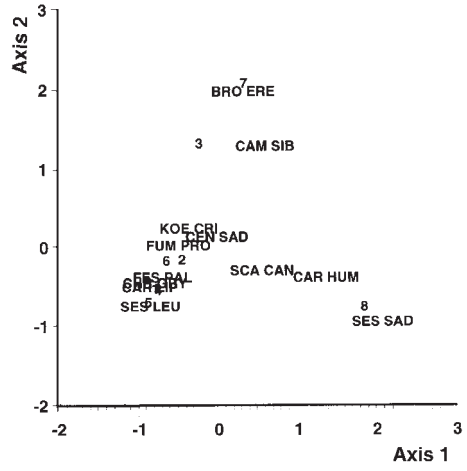

Figure 7.14. Correspondence analysis of Table A1 with $\alpha = 0.5$. Compare the result with the PCA diagrams (Figs 7.2-6 and 7.8).

If $\alpha = 1$, then the coordinates of quadrats (now columns) are obtained as the weighted average of species coordinates on the same axis. In the joint plot, species *i* will be the closest to those quadrats in which it has the highest proportion. That is, its position in the ordination space is an "estimate" of its optimum locality. It can easily happen that several species have their optima outside the set of quadrats examined, therefore the species scores usually have a wider range than the quadrat scores. If there is a quadrat in which only one species is present, then this quadrat will coincide with that species in the joint plot. Many COA ordination programs (such as **DECORANA**, Hill 1979b) offer only this option to the user. In case of $\alpha = 0$, we find the opposite situation: the species (row) coordinates are obtained as the weighted average of quadrat coordinates (without rescaling, because $1/\lambda^{\alpha} = 1$). Then, the points representing species fall usually close to the origin, and the quadrat coordinates have the wider range. If there is a species which appears in a single quadrat only, then its position in the joint plot will coincide with the position of that quadrat. In fact, the setting with $\alpha = 0.5$ provides the average of the previous two configurations. When the eigenvalues pertaining to the axes portrayed are close to 1 (the data matrix is strongly structured, χ^2 is high), the change of the a scaling parameter causes negligible differences in the resulting joint plots. In these cases, the relative positions of rows and columns in the graph convey meaningful information. For low eigenvalues – for less structured data matrices – the nearness of points is not interpretable, and the angles and directions remain meaningful. These are illustrated by examples in the next subsection (Figs 7.16a-c).

5) When one considers standardizations 7.43-44 and the matrix equation of COA (7.41), the statement that COA is a special version of PCA will not be too surprising (Greenacre & Vrba 1984, for example, introduces COA exactly in this sense). Whereas centred PCA maintains the Euclidean distances of objects in the ordination, COA preserves the so-called χ^2-distances

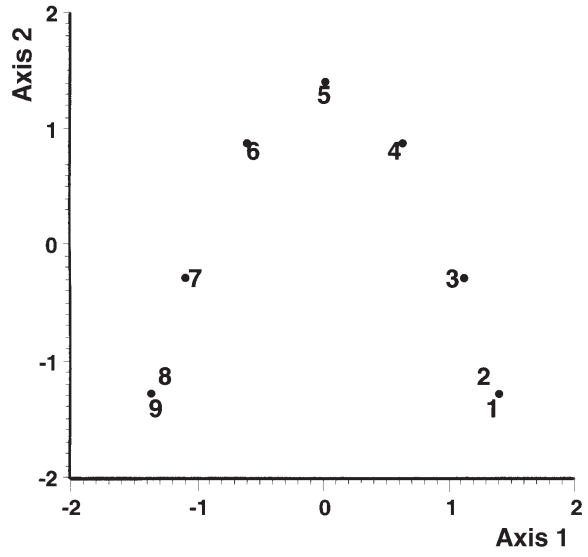

Figure 7.15. The result of COA with $\alpha = 0.5$ for a strongly non-linear data structure. The second axis is apparently a quadratic function of the first. The first two eigenvalues are high (0.92 and 0.72), so that the modification of α does not influence significantly the relative positions of points.

(Formula 3.67). More precisely, for $\alpha = 1$ in the ordination of columns the squared distance between points *j* and *k* for all axes will be proportional to the quantity

$$CHISQD_{jk} = \sum_{i=1}^{n} \frac{(x_{ij} / x_{.j} - x_{ik} / x_{.k})^2}{x_{i.}} . \qquad (7.48)$$

(This is in fact the squared Euclidean distance calculated from double standardized data.) $CHISQD_{jk} = 0$, when the two objects contain the same variables in identical proportions (one object is obtained from the other using a *q* arbitrary non-negative number as a multiplying factor, i.e., the second is *q* times the first). Alternatively, when $\alpha = 0$, the ordination of rows will preserve the chi-square distances (the equation for rows is similar to Formula 7.48). In tjis case, $CHISQD_{hi} = 0$, if variables *h* and *i* have always the same proportion in the objects (e.g., the abundance of species *h* is *q* times higher than the abundance of species *i* in every quadrat).

It is now due time to show the performance of the method using a simple actual example. Consider again the data in Table A1, now conceived as a 12\8 contingency table. The present arrangement is favourable as far as computing speed is concerned, because the number of columns is smaller than the number of rows. (If, however, 30 species describe, say, several hun-

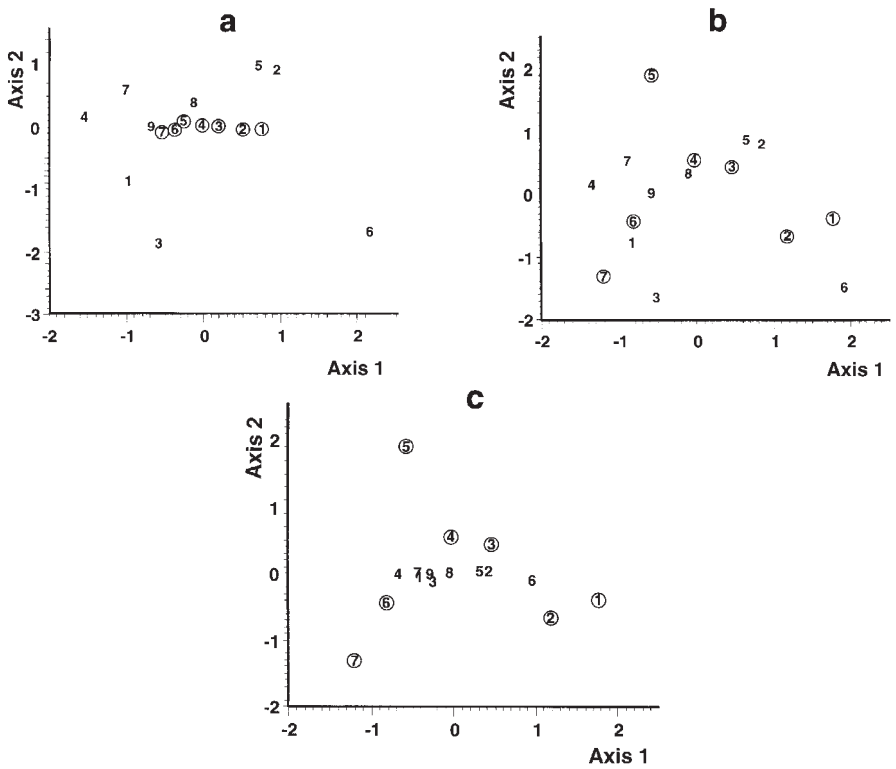

Figure 7.16. Correspondence analysis of matrix 7.15 (p. 232) for three different values of the scaling factor (**a:** $\alpha = 1$, **b:** $\alpha = 0.5$, **c**: $\alpha = 0$). In **a**, the coordinates of rows are obtained as the weighted average of column coordinates, contrary to case **c** in which this is in the opposite way, whereas **b** represents a perfect compromise between the previous two.

dred quadrats, then the species should be the columns to reduce computing time.) The joint plot obtained by a symmetrically weighted COA is displayed in Figure 7.14. If compared to the PCA results, we do not see substantial differences, and all what appear is caused by standardization with column and row totals. The results of asymmetric analyses (i.e., with $\alpha \surd 0.5$) need not be shown, because they do not differ much from the ordination in Fig. 7.17, owing to the relatively high first eigenvalue ($\lambda_1 = 0.7$). Since the reader has gained some experience already with arched arrangements in ordinations, a horseshoe-like trend is recognized easily in this result. Let us therefore see how this "effect" can be treated in correspondence analysis.

7.3.4 The "arch effect" and data linearity in correspondence analysis

Since both PCA and COA can be traced back to the same eigenanalysis problem, the arch effect is also known from COA, as shown by the previous example. In the joint plot of the COA of matrix 7.14 (Fig. 7.15), as expected, a conspicious double arch can be observed.

The reason is known already to us: the data structure is non-linear. As a result, the interpoint distances are maintained most faithfully by the analysis if the points are arranged along a horseshoe. At the same time, the "correspondence" between variables and objects is perfectly visualized in the ordination. The interpretation of the result is thus partly if at all influenced by the presence of the arch (Greenacre 1984). Many authors nevertheless consider its "removal" as an essential task. Hill (1979b) and Hill & Gauch (1980) developed the so-called "*detrended correspondence analysis*" (DCA) which operates by splitting the first axis into arbitrary segments, and by shifting these segments to obtain the best fit of points to a line. Since the detrending algorithm of DCA has been still popular in vegetation science and in ecology, in general, I emphasize again what I have said already in the discussion of PCA: the uncritical use of detrending cannot be recommended. At best, DCA may prove useful if applied parallel to standard COA so that we can appreciate the differences. DCA will not add much to what we can conclude from standard COA, only some "aesthetic" improvement of the joint plots can be achieved. A potential danger in the exclusive use of DCA is its "black box" behaviour, which may even cause loss of ecologically meaningful information (Pielou 1984). The question of "detrending versus not detrending" has been subject of hard methodological debates (see Gauch 1982, Kenkel & Orlóci 1986, Minchin 1987, Wartenberg et al. 1987, Peet et al. 1988, Oksanen 1988, Knox 1989). Reyment (1991) concluded the dispute and left the question open by saying that the "proof of the pudding is in the eating", thus emphasizing that everything is case-dependent. Jongman et al. (1987) and ter Braak & Prentice (1988) appear to prefer the method of polynomial regression (as proposed by Phillips, 1978, for PCA) against detrending and it is included as an option in ter Braak's (1988) **CANOCO** program.

Let us now examine how the method of COA performs for variables that are linearly related to one another. The starting data, as for PCA, is matrix 7.15 in which the columns are the variables (their approximately linear relationships are obvious).

The COA evaluation shows that the data values are not as unexpected as were in case of matrix 7.14, because the first eigenvalue is merely 0.19, and the second is as little as 0.003 (as far as their proportions are concerned, there is no big difference from the PCA result, the first eigenvalue is much more important than the second). A consequence of the low eigenvalue is that the value of α will considerably influence the relative positions of points in the joint plot (Figs 7.16a-c). If the row (quadrat) coordinates are obtained as weighted averages of column (species) coordinates (Fig. 7.16a), then the seven quadrats will closely fit the first axis, showing that the linear arrangement is well-detected by COA. The relationships in terms of χ^2-distances are maintained for the quadrats in this diagram. The species are arranged around the quadrats, showing that the species optima fall outside the range of quadrats included in the analysis. As ter Braak & Prentice (1988) pointed out, COA assumes a *unimodal response of species* to a background gradient, contrary to PCA, and its lack causes the low eigenvalue.

(Nevertheless, the unimodal response, as we have seen above, will not exclude the possibility of an arch.) In the opposite situation (Fig. 7.16c), the coordinates of columns (species) are derived as weighted averages of quadrat coordinates, so that the species will fall close to the origin, arranged approximately linearly. Their distances will be proportional to their χ^2-distances. Symmetric COA (Fig. 7.16b) is an "average" of the previous two configurations. In this case, the closeness of points is not suggestive of whether a species characterizes certain quadrats unequivocally. The directions and relative distances from the origin, however, do have a definite meaning.

7.3.5 Canonical correspondence analysis

Like PCA, COA also has its constrained form, known as *canonical correspondence analysis* (CCOA, ter Braak 1986, 1987). The method has been widely used in ecological data analysis, and serious journals can easily reject manuscripts devoted to evaluating the relationships between species and environment (e.g., in direct gradient analysis, ter Braak & Prentice 1988) if CCOA happens to be 'forgotten' by the author. It is therefore inevitable to provide at least a brief summary on its theoretical foundations and the interpretation of its results.

As in RDA, the ordination of objects (sites, quadrats) is not exclusively based upon the species data since the axes are also influenced in some way by the environmental variables (as we shall see below, there are two possibilities to exert this influence). As usual in ordinations, the axes must explain as much of the total variance as possible, but owing to the constraining effect of external variables the variance accounted for by CCOA axes is more or less lower than in COA ordinations. This 'sacrifice' has to be made in order to be able to interpret ordination axes more directly and to reveal species/environment relationships within the ordination of objects.

The theory behind CCOA is best understood if we consider the reciprocal averaging algorithm (Subsection 7.3.1) modified in several steps (Jongman et al. 1987). Assume that data matrix $\mathbf{X} = \{x_{ij}\}$ includes the species as rows and the objects as columns. Assume further that matrix **Z** contains the environmental data, with variables as rows and objects as columns, the latter presented in exactly the same order as in **X**. Note that the variables are centred. Let the number of environmental variables be q. Then, the species and object coordinates on the first axis are determined using the following iterative algorithm:

1. Generate an arbitrary set of coordinates for objects (vector **b**) such that all scores are different. Let this coordinate be denoted by b_j for object j.

2. Species coordinates (vector **a**) are obtained as weighted averages from the object coordinates, that is for species i we calculate $a_i = \sum_{j=1}^{m} b_j \frac{x_{ij}}{t_i}$, in which t_i is the row total for species i.

3. The new species scores are now used to derive new object scores (vector **b***) by weighted averaging, that is $b_j^* = \sum_{i=1}^{n} a_i \frac{x_{ij}}{u_j}$, where u_j is the total of species scores in object j (column totals). (These totals are written into the diagonal matrix **N.**) This step implies that **b*** contains *weighted average scores* of objects.

4. The coordinates of objects (taken as 'dependent' variables) are weighted by u_j and then fitted to the environmental variables (now considered as 'independent' variables) using *multiple regression*. The matrix equation of this weighted least squares regression is $\mathbf{c} = (\mathbf{ZNZ}')^{-1}\mathbf{ZNb}^*$. The regression coefficients (more precisely, *canonical coefficients*) thus obtained will be used to derive the fitted scores of objects. For object *j,* this score is calculated as

$$b_j = \sum_{h=1}^{q} c_h \, z_{jh} \qquad (7.49)$$

and is written into vector **b** (recall that z_{jh} is the centred score of environmental variable h in site). Equation 7.49 implies that the scores are now *linear combinations* of the environmental variables (see Appendix C, for more on linear combinations).

5. The coordinates obtained in the previous step are standardized (rescaled): from the weighted scores the mean is subtracted and the result is divided by the standard deviation.

6. If the difference between the scores just derived and those computed in the previous iteration does not exceed a pre-specified threshold, then the analysis stops. Otherwise return to step 2.

The main difference from RA is the multiple regression applied in step 4. The iterations converge into a stable solution regardless of the initial scores; only the number of iterations may vary. When the analysis is completed, the standard deviation of points corresponds to the eigenvalue. When the first axis is determined, its effect can be removed and a second CCOA axis orthogonal to the first is extracted using the same algorithm, and so on.

In the interpretation of CCOA results, the following considerations apply:

- First of all – as mentioned already – there are two sets of object coordinates when the analysis is finished: vectors **b*** and **b** obtained in the last iteration steps 3 and 4, respectively, for each ordination axis. The first set may be abbreviated as WA, the second as LC scores (Palmer 1993). The WA scores represent the direction of variation in the species by sites matrix **X** constrained to be maximally correlated with the LC scores. The LC scores, on the other hand, are the best fit of species to the environmental variables. The ordinations based on these two sets of scores may differ considerably, and care is needed when examining computer outputs[7]. As McCune (1997) points out based on his experiments with simulated data sets, the LC scores are very sensitive to even moderately noisy environmental data. Since environmental data are usually noisy anyway, the sensitivieness is even more critical when the measured variables are in fact irrelevant. For the exploration of community structure, therefore, McCune suggests the use of WA scores, unless there is a good reason to assume that the environmental variables are almost noiseless.

7 A good check is to do an analysis such that two sites differ only in the proportion of constituting species (all scores in site 1 are constant times the scores in site 2), and this is not so with the environmental data. For WA axes, these two sites should fall into the same position.

- The WA scores can be rescaled, as in correspondence analysis, using Equation 7.34, (the explanation given on p. 243 applies here as well).
- The two sets of scores are analogous to the two sets of CV scores obtained in COR, and could be displayed in the same way (see Fig. 7.12). Instead of this, however, the linear correlation between **b*** and **b** is used to measure the strength of the relationship between species and the environmental variables. This is often called the *species/environment correlation*. Its value is often misleadingly high, suggesting that the eigenvalues should also be considered in evaluating the importance of ordination axes. Axes with low eigenvalues can produce high correlation, but such dimensions account for a small portion of the variance, thus having little interpretive value.
- The species and object coordinates can be simultaneously illustrated in a joint plot, as in COA. Objects in which a given species has high proportions will be close to the position of that species, depending on the value of α, as we have seen already. In the same diagram, the environmental variables are also displayed. Those measured on the interval or ratio scale are represented by arrows (as in PCA). The interpretation of relative species positions and arrow directions is as follows: the species can be projected onto each arrow to derive a species sequence which reflects approximately the response of species to that environmental variable. In this way, species positively or negatively associated with the environmental variable can be identified, as exemplified below. The length of arrows is proportional to the correlation of the variable with the axes and, of course, the long arrows are the most important in interpreting the results. Nominal variables may also be included in the analysis, they are represented by several binary variables whose number is one less than the number of states of the nominal variable. Usually, these states appear as points in the display (Jongman et al. 1987, p. 142). The coordinate for a state of the variable is obtained as the weighted average of object coordinates in which this character state appeared.
- The arch 'effect' is usually less conspicious in CCOA than in COA. Occurrence of arched arrangements may be explained by the presence of too many, potentially irrelevant environmental variables (ter Braak & Prentice 1988), whose removal will solve the 'problem'. (Do not forget that inclusion or exclusion of any variable remains arbitrary and one may always be tempted to play with the selections until the 'best' configuration results.) If the number of environmental variables is low and they are highly correlated with the axes, then the arch may completely disappear.
- CCOA is not merely an exploratory approach. A permutation test may be built into the procedure to test whether the species/environmental correlation is significant (ter Braak & Šmilauer 1998).

The method is illustrated using sample data taken from ter Braak (1988, Table A4). The matrix comprises 30 species and 20 sites, the latter also characterized by three environmental variables: the depth of A1 horizon in the soil, soil humidity and the quantity of fertilizer. The analysis reveals how these environmental factors influence the species composition of the dune vegetation.

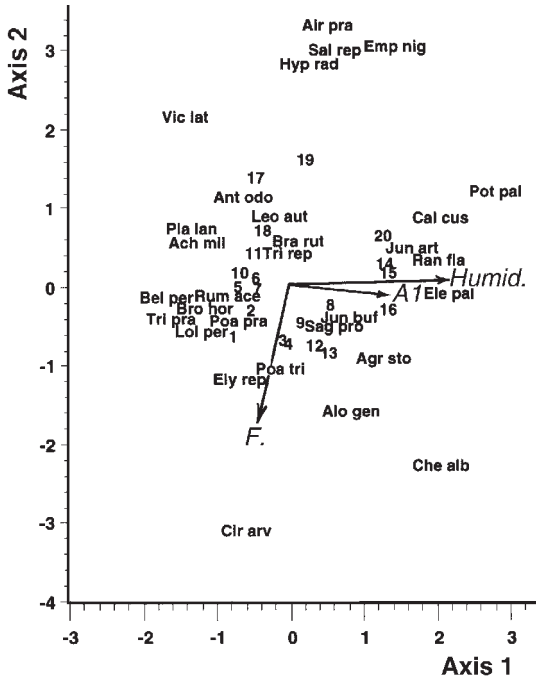

Figure 7.17. CCOA of the dune vegetation data (Table A4) using $\alpha = 1$. The diagram displays three kinds of points, sample site WA scores (1-20), species (labeled by abbreviated names) and environmental variables (depth of soil layer A1, humidity and fertilizer). Therefore, the term *triplot* appears most appropriate (Šmilauer 1992).

The site scores are weighted averages of species coordinates (Fig. 7.17). As a result, several species are positioned outside the range of site scores. Of the three environmental variables, the A1 horizon and humidity appear to be responsible for the first canonical axis (with correlations of 0.56 and 0.90, respectively), whereas the second axis is most correlated with the amounf of fertilizer ($r = –0.79$). These relationships are demonstrated by the length and direction of arrows in the triplot. The eigenvalues pertaining to the first two dimensions are 0.42 and 0.23, a little smaller than the eigenvalues from the plain COA of the same species \ sites matrix (0.53 and 0.40, respectively). It was expected because the current axes are constrained to be maximally correlated with the linear combination of environmental variables, and the pure and constrained axes very rarely, if ever coincide. Nevertheless, the difference is small, as shown by the high species/environment correlations (0.925 on axis 1, 0.816 on axis 2). In order to intepret species positions and arrow, consider the humidity variable. If the arrow is extended in both directions and species are projected onto this line, then we get the species ordering according to humidity requirement: on the right side species of more humid habitats appear, as opposed to the left side where dry-tolerants are concentrated. The higher the relative importance of the corresponding eigenvalues (25% and 15%, in this case), the more faithfully represented are the 'true' humidity requirements by this ordering.

7.4 Multidimensional scaling

A common property of ordination methods discussed thus far is that access to the raw data matrix is inevitable throughout the computations. In certain situations in biology, sampling or other observations may directly produce distance or dissimilarity matrices, as mentioned in the chapter on cladistics (Subsection 6.2), with Sarich's immunological distances as good examples. The question arises: if evolutionary relationships can be reconstructed from distances,

then is efficient dimension-reduction also possible from distance matrices? The answer to this poetical question is of course yes; the methods of *multidimensional scaling* (abbreviated as MDS) have been designed to produce an ordination of objects from their distances or dissimilarities.

The topic of MDS considered alone is very complicated, so our attention must be concentrated upon two major issues, the metric versus non-metric approaches. Metric MDS is closely related to PCA in its linear algebraic algorithm. The non-metric MDS, on the other hand, gives a chance to discuss an alternative solution to the ordination problem which is not associated with eigenanalysis. In many cases, especially if the input dissimilarities do not obey the metric axioms, non-metric MDS offers the sole possibility for ordination.

7.4.1 Metric multidimensional scaling alias principal coordinates analysis

This method has originally been suggested by Torgerson (1952) but gained wide popularity thanks to Gower's efforts (1966). He proposed the name *principal coordinates analysis,* abbreviated here as PCoA (other names may also persist, for example, Digby & Kempton [1987] prefer the acronym PCO). PCoA is metric because the resulting ordination preserves the metric distances among the objects, as does PCA. The fundamental requirement is that as many ordination axes are extracted as necessary to maintain the input metric distances in the output ordination. Therefore, the main condition of its applicability is that the input distances satisfy the metric axioms (Subsection 3.1.1) although – as explained later – minor violations of these axioms do not hinder the interpretability of PCoA results. A typical illustrative example of PCoA is not biological but, nonetheless, owing to its didactic clarity and its connections to everyday life, many books start with this example (e.g., Manly 1986). Charts of road distances between large towns are often printed on the back cover of road maps. Based on such semimatrices, PCoA is able to reconstruct the relative positions of towns, that is, the map itself. The success of this reconstruction depends solely on the crookedness of the roads. When they are not entirely straight, and usually they are not, then the PCoA ordination in the first two dimensions may only approximate the true map positions, and further dimensions are required to explain the twists and turns of the roads. In the hypothetical situation with all roads being straight, the inherent dimensionality of the distance matrix (its rank, see Appendix C) is only 2, so that the map of towns can be portrayed by PCoA in the plane without distortion. We shall see that in fact PCoA will generate 'data' (the coordinates) from the distances in such a way that these data reproduce perfectly the input matrix.

> Consider the road distances between ten European cities (Table A7). Principal coordinates analysis of this matrix produces a fairly faithful reconstruction of the map (Fig. 7.18) and the approximate inter-city distances can be read from the scale on the axes. There is only a minor problem with the arbitrary directions; the point configuration needs to be rotated and reflected to find correspondence with the four cardinal points. The efficiency of the two axes shown will be discussed later in due time.

Principal coordinates analysis is performed in two main steps. The first one is an ingenious trick; the matrix of distances is used to produce another symmetric matrix viewed as a cross-products matrix that would result from the 'data' were are going to determine. This cross-products matrix is analogous to the variance/covariance or correlation matricesin PCA

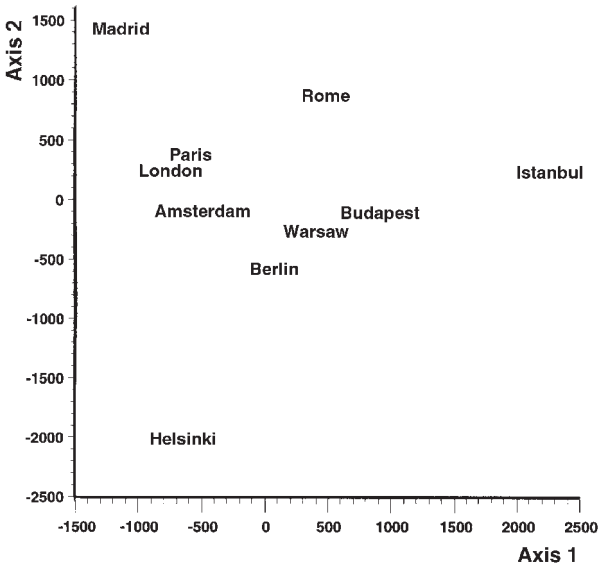

Figure 7.18. Reconstructing the relative positions of ten European cities from their road distance matrix (A6), using principal coordinates analysis.

or the **Z′Z** matrix in COA. In the next step, this matrix is analyzed for eigenvalues and eigenvectors which are in turn used to calculate the coordinates themselves.

The **A** cross products matrix of size $m \setminus m$ could be obtained from the coordinates written into matrix $\mathbf{X}_{n,m}$, if they were known, using the formula:

$$a_{jk} = \sum_{i=1}^{n} x_{ij} x_{ik} \tag{7.50}$$

or, in matrix algebraic terms:

$$\mathbf{A} = \mathbf{X'X}. \tag{7.51}$$

The relative positions of points do not change if these 'coordinates' are centred. An advantage of centring is that a trivial eigenvalue (as in COA) is automatically removed. That is,

$$\sum_{j=1}^{m} x_{ij} = 0 \text{ , for all variables } i \text{ .} \tag{7.52}$$

As a consequence of centring, all row and column totals in **A** are zero:

$$\sum_{j=1}^{m} a_{jk} = \sum_{k=1}^{m} a_{jk} = 0 \text{ .} \tag{7.53}$$

Now suppose that the initial, *squared* distances, d_{jk}^2, are expressed from the coordinates sought using the well-known formula,

$$d_{jk}^2 = \sum_{i=1}^{n} (x_{ij} - x_{ik})^2 \text{ ,} \tag{7.54}$$

which is equivalent to writing:

$$d_{jk}^2 = \sum_{i=1}^{n} [x_{ij}^2 + x_{ik}^2 - 2\, x_{ij}\, x_{ik}\,] = \sum_{i} x_{ij}^2 + \sum_{i} x_{ik}^2 - 2 \sum_{i} x_{ij}\, x_{ik} \text{ .} \tag{7.55}$$

From Formula 7.50, this expression can be rewritten as:

$$d_{jk}^2 = a_{jj} + a_{kk} - 2a_{jk} \,. \tag{7.56}$$

Then, expressing a_{jk} from 7.56 we obtain:

$$a_{jk} = \frac{1}{2}[-d_{jk}^2 + a_{jj} + a_{kk}] \tag{7.57}$$

the right side of which, after substitutions not detailed here (see, for example, Pielou 1984, p. 184) can be rewritten perfectly in terms of squared distances:

$$a_{jk} = -\frac{1}{2}[\, d_{jk}^2 - d_{j.}^2 - d_{.k}^2 + d_{..}^2 \,], \tag{7.58}$$

where

$$d_{j.}^2 = \frac{1}{m} \sum_{k=1}^{m} d_{jk}^2 \tag{7.59}$$

is the mean of squared sistances from object to all the other objects, and

$$d_{..}^2 = \frac{1}{m^2} \sum_{j} \sum_{k} d_{jk}^2 \tag{7.60}$$

is the grand mean of all squared distances, including zeros in the diagonal.

The above derivation demonstrates that PCoA starts from matrix **A** computed by Equation 7.58. After determining its eigenvalues and eigenvectors in the familar way:

$$(\mathbf{A} - \lambda\mathbf{I})\,\mathbf{v} = \mathbf{0} \tag{7.61}$$

such that the eigenvectors are of unit length and the eigenvalues are arranged in descending order, then our interest turns towards the spectral decomposition theorem (Appendix C). This states that the symmetric matrix **A** may be expressed as a product of three matrices according to the following equation:

$$\mathbf{A} = \mathbf{V}\,\Lambda\,\mathbf{V}' = (\mathbf{V}\,\Lambda^{1/2})\,(\Lambda^{1/2}\,\mathbf{V}'), \tag{7.62}$$

in which Λ is a diagonal matrix containing the eigenvalues. Using Equations 7.51 and 7.62, the coordinates are obtained as:

$$\mathbf{X} = \Lambda^{1/2}\,\mathbf{V}' = [\,\sqrt{\lambda_1}\,\mathbf{v}_1,\ \sqrt{\lambda_2}\,\mathbf{v}_2,\ ...,\ \sqrt{\lambda_m}\,\mathbf{v}_m\,]\,. \tag{7.63}$$

For a deeper understanding of the principles of the method and a more thorough interpretation of its results, one should consider the following points:

- The centred PCA of an $n \backslash m$ data matrix, starting from the covariances of n variables, and the PCoA using the squared Euclidean distances of the m objects will produce identical ordinations (only the directions, i.e., the signs may differ). This is not surprising to us, since both analyses rely upon the eigenanalysis of symmetric matrices. A COA of the same data, such that object coordinates are weighted averages of species scores, and a PCoA based on a matrix of χ^2-distances of objects reveal the same distance structure for all dimensions, but there may be slight changes if the first two dimensions are viewed (cf. Digby & Kempton 1987).
- If the starting matrix is Euclidean, then the maximum number of positive eigenvalues is m–1, and the mth one is zero. In this case, the diagonal of matrix **A** contains the squared distances of points from the centroid. The sum of these values, tr { **A** }, is the total sum of squares for the points, which may also be expressed using the pairwise distances (Equation 3.106). This amounts to the sum of eigenvalues:

$$\text{tr}\,\{\mathbf{A}\} = \sum_{j=1}^{m} a_{jj} = \sum_{j} \sum_{k} d_{jk}^2 \,/\, 2m = \sum_{k=1}^{m-1} \lambda_k \quad . \tag{7.64}$$

Accordingly, the first t dimensions will account for

$$100 \times \sum_{k=1}^{t} \lambda_k \,/\, \sum_{k=1}^{m-1} \lambda_k \tag{7.65}$$

percent of the entire distance structure. A two-dimensional PCoA diagram explaining no more than 20-30% of the total sum of squares can be misleading as to the nearness of certain point pairs. Those points falling close to each other in two dimensions may fall far apart if the other axes are also considered. A good check of it is the minimum spanning tree superimposed on the ordination (Subsection 5.4.3), as will be discussed in detail in Chapter 9.

- Negative eigenvalues indicate that the starting matrix was not Euclidean. Some very small negative eigenvalues may be ignored without risking the interpretability of dimensions with very high positive eigenvalues. However, large negative eigenvalues may cause some headache to us because the dissimilarity structure can only be represented in the Euclidean space with distortion, and the PCoA results are in doubt. A potential solution in such cases is offered by the method of non-metric multidimensional scaling (next subsection).

The two dimensions portrayed in Figure 7.18 account for 46.4% and 38.3% of the total sum of squares, respectively. Scaling the 10\10 matrix onto the plane has therefore a 84.7% success. The remaining portion of the variance is caused by the deviation of roads from straight airline distances. These are explained by three small eigenvalues (the associated percentages being 8.8, 5.8 and 0.7), the others are zero so the matrix has a rank of 5. The efficiency of variance extraction is even higher for the immunological matrix (A5), because in this case we have 63.1% and 22.7% (plus four small eigenvalues). The diagram for this second example is not shown. It is sufficient to say that axis 1 explains the separation of monkey from the remaining taxa, whereas on the second axis the cat is separated from the other six species. They form a group around the origin of axes 1-2, and their differences are manifested only in the subsequent dimensions.

The arch effect and its (flexible) shortest path adjustment. The PCoA result obtained from the Euclidean distances computed for the rows of matrix 7.14, as the reader may guess, will exhibit the same arch effect (Fig. 7.19a) as in the centred PCA ordination. It further clarifies the point made in Subsection 7.1.6 that the arched arrangement reflects the 'attempt' to preserve interpoint distances in the ordination as faithfully as possible. When the distance taken from the first object along the gradient reaches the possible maximum, then it cannot increase any longer. This observation led to the development of a correction algorithm by Williamson (1978) and Clymo (1980). They proposed to recalculate the distances between objects (e.g., sites) which have no characters (species) in common. The distance is increased such that the distance increments along the gradient are proportional to the previous changes. As the authors suggest, a sequence of objects is to be found between the two objects in question such that the neighbouring objects in the sequence agree in the presence of at least one species and the sum of pairwise distances along the sequence is the minimum ('*shortest path*'). The sum of these distances will then provide the adjusted, i.e., increased distance between the endpoints of the sequence.

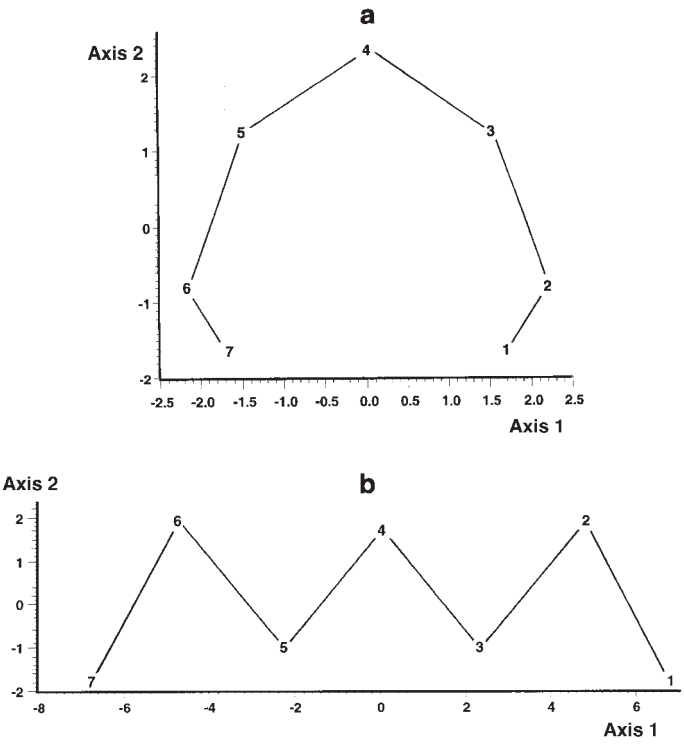

Figure 7.19a: If data structure is non-linear (matrix 7.14), then PCoA is not free from the arch effect either. **b:** Increasing the distances along the gradient by shortest path adjustment gives a fairly good approximation to the underlying gradient.

For example, in matrix 7.14 objects (rows) 1 and 4 attain the maximum distance of $EU_{14} = 4.69$, and the distance is the same for the pairs 1-5, 1-6 and 1-7. The shortest path between objects 1 and 4 is represented by the sequence 1-2-4, with associated distances 3.16 and 4.47 so that their sum, 7.63, will become the adjusted value. The modified distances $EU'_{15} = 8.94$, $EU'_{16} = 12.1$ and $EU'_{17} = 13.4$ are obtained analogously. The efficiency of these operations is best appreciated by comparing the PCoA results obtained for matrix 7.14 with and without adjustment (Figure 7.19). The analysis of the modified matrix, however, produced negative eigenvalues as well (the list of eigenvalues is 148.37; 18.76; 1.53; 0.13; 0.00; –2.26 and –8.52) indicating the non-Euclidean property of adjusted distance matrices. The first eigenvalue differs by magnitude from the others, and the point positions on axis 1 reflect pretty well the arangement along the gradient. In absolute terms, λ_2 is only twice higher than λ_7 suggesting that the ordination along the second and all subsequent axes is distorted and uninterpretable.

Bradfield & Kenkel's (1987) *flexible path adjustment* goes a little further. In addition to distances for which there were no common species in the corresponding objects, the distances for which only k common species were found are also recalculated. The value of k may be modified (k=1, 2, 3, and so on). This adjustment may prove useful in the analysis of gradients with extremely high β-diversity (speices turnover). The occurrence of negative eigenvalues for adjusted matrices requires future studies, however. Another 'disadvantage' of these adjustments is that the original data are required which is, in general, not so with principal coordinates analysis.

7.4.2 Non-metric multidimensional scaling

All methods discussed so far in this chapter generate ordinations by preserving the metric information in the data directly or indirectly. PCA, COA and PCoA assume the existence of linear relationships among variables, and the violation of this condition may lead to some problems in interpreting the final results. Minor deviations from linearity are tolerated practically by all methods (robustness), but strong non-linearity will burden the result with the 'arch effect'. The method of *non-metric multidimensional scaling"* (NMDS), however, is free from any assumptions on linearity and can be based on any symmetric distance or dissimilarity matrices.

The essence of this approach is that *differences* between the distance values themselves, which actually convey the metric information, are completely neglected and only the *rank order* of distances is considered[8]. The objective is to arrange the points in a prespecified number of dimensions (usually two) such that the rank order of ordination distances is as close to the rank order of starting distances as possible (Shepard 1962, Kruskal 1964). In other words, the ordering relation between the starting inter-object distances or dissimilarities (d_{jk}) and the ordination distances (δ_{jk}) should be monotonous. Since the final result is presented in form of metric coordinates, the method would be better called 'ordinal scaling', as Gordon (1981) suggests. Anyway, it turns out now very clearly that the term ordination, preferred by most biologists, is less precise mathematically, because most ordinations imply a lot more than simple preservation of ordinal relationships between points.

The best known algorithm of non-metric MDS (Kruskal 1964) is iterative. An initial configuration of points is specified by the user (randomly or arbitrarily generated, or derived from another ordination) and this is refined and improved through the iteration steps until the point where no substantial improvement is possible. Each iteration step consists of two parts, as detailed below:

1) The rank order of original distances (or dissimilarities) and the rank order of ordination distances are contrasted using the technique of *monotone regression* (Kruskal 1964). It is not a direct comparison of two rank orders, as in rank correlation, because the only thing examined here is how much the ordination distances should be modified (decreased or increased) in order to reach monotonicity with the input measures. Figure 7.20 helps understand this concept. Suppose we have only four objects to ordinate, so that there are six distances. Construct a coordinate system in which the original distances are measured on the vertical axis and the ordination distances on the horizontal one. Each point in this diagram corresponds to a pair of objects. The two diagrams in the figure exemplify contrasting situations. In Fig. 7.20a, only minor shifts (illustrated by arrows) are needed to reach monotonicity, whereas in diagram b the deviations are much larger. Apparently, the first solution is better than the second. The difference can be expressed quantitatively as well, using the *stress function* proposed by Kruskal, which operates on squared deviations:

8 The loss of information is inevitable here. The situation is analogous to the conversion of interval scale variables to the ordinal type.

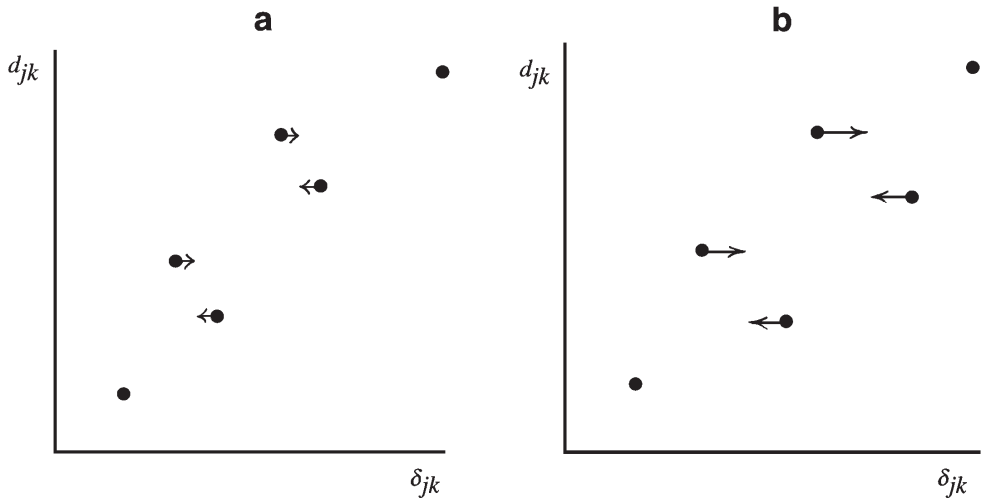

Figure 7.20. Comparison of the original (d_{jk}) and ordination (δ_{jk}) distances using a Shepard diagram, in two contrasting situations (see text).

$$ST = \left[\frac{\sum_{j<k} (\delta_{jk} - \hat{\delta}_{jk})^2}{\sum_{j<k} \delta_{jk}^2} \right]^{1/2} . \qquad (7.66)$$

Indded, this is the Euclidean distance normalized to the interval [0,1]; otherwise the sum of squares would be unbounded, rendering the regression results difficult to compare. The value of $\hat{\delta}_{jk}$ is the amount of change necessary to modify δ_{jk} to achieve monotonicity. That is, $\hat{\delta}_{jk}$ is the average of two or more distances (Figures 7.20a and b show two such averages). If $ST = 0$, the order of ordination distances perfectly fits the order of original distances and no changes need to be implemented. It is implied, of course, that the distance values can be radically different, because there is a remarkable freedom to modify the distances such that the ordering remains intact. The stress function informs us quantitatively how efficient the ordination is in preserving the ordering relations among the input distance values.

2. When the decrease of *ST* is smaller than a prespecified threshold ε (say, 0.001) between two iteration steps, then the analysis stops and the actual configuration is taken to be final because no significant improvement is possible any longer. Otherwise, the computations continue by shifting point positions to allow further reduction in the stress. The method of *steepest descent* is used for this purpose. It is essentially the computation of the partial derivative of the stress function for each coordinate to find the 'direction' of moves which leads to the maximum possible reduction of the stress. The algorithmic details are presented in Kruskal (1964) and Brambilla & Salzano (1981).

After shifting the point positions, another monotone regression is performed, and so on. When the iterations stop, the final configuration is rescaled to have a zero centroid and unit sum of squared distances of points from the centroid. This is recommended to facilitate comparison with other ordinations. Actually, the final configuration can be rotated to any degree and compressed or dilated by an arbitrary scale factor; NMDS has no restrictions in this regard contrary to other ordination procedures.

In addition to the ordination of points, the graphical comparison of the original and ordination distances via the so-called *Shepard diagrams* is also an integral part of the result. The diagrams in Figure 7.20 used to demonstrate stress are examples. In a Shepard diagram, the worse the fit of ordination distances to the original values (the stress is high), the more scattered are the points. On the contrary, when the stress is very low the points are concentrated along a diagonal line (see also Figure 7.21b).

The user of NMDS may consider the following points in preparing the analysis and in evaluating the results:

- The starting dissimilarity matrix may be non-metric; the axioms can be violated drastically. The method can be programmed to tolerate missing distance values as well. If PCoA produces some highly negative eigenvalues, thus questioning the interpretation of axes with positive eigenvalues of similar size, then NMDS remains the only plausible ordination method. In a comparative survey, Kenkel & Orlóci (1986) found that NMDS + chord distance were relatively efficient to reveal two-dimensional background gradients (='*coenoplanes*'). Minchin (1987) reached similar conclusions. Other authors (Gauch et al. 1981, Digby & Kempton 1987) emphasize the limitations and drawbacks of NMDS, relying upon the following points in their argumentation.
- The number of dimensions is determined in advance by the investigator, so there is some resemblance to factor analysis where the number of factors is defined *a priori*. We may immediately start with two dimensions, because this is easily illustrated in papers and theses. Choosing more dimensions is also a good start, because a four-dimensional solution may be the starting point for a 3-dimensional NMDS which in turn can be optimized for the final two dimensions. Contrary to metric ordinations, however, a k-1 dimensional ordination is not merely the omission of the kth dimension! The relationship between *ST* and the number of dimensions may be displayed graphically in order to determine the 'optimum' number of dimensions. Obviously, *ST* decreases over increasing values of k and, as a rule of thumb, we can keep the dimensionality where no great reduction of stress is achieved – but deciding whether a change is great or not is always arbitrary. In most of the cases, however, NMDS is used to end up with two dimensions, no matter how large the stress is.
- There is no general rule as to the optimal value of *ST* either. Many authors consider a stress of $ST = 0.05$ as being very good, although our judgement must depend on the number of points and dimensions. Often, values between 0.1 and 0.2 are still acceptable but again, there are no general rules. Permutation and bootstrap tests may offer a solution in the future.

- As with other iterative procedures (RA being an exception, but see also Chapter 8 for other examples), the final result of the computations depends on the starting configuration. The analysis does not necessarily converge into the same, the best solution. It may 'hit' the optimum (the *global* optimum) but from other starts the analysis may be trapped in very poor *local* optima. This problem is circumvented by performing the analysis from different random configurations and then maintaining the result that provides the lowest stress value (Shepard 1980). Two-dimensional ordinations obtained by metric methods may also be used efficiently to get closer to the optimum result. Other possibilities are discussed by Groenen (1992).
- Another practical question concerns the magnitude of input distances. Kruskal (1977) reports on a study in which the distances were categorized into three groups (large, medium and small distances) and then the effect of omitting each group upon the results was examined. It turned out that small distances had little influence, while the omission of large distances modified the ordination most signficantly. Uncertainties associated with small distances therefore do not appear to be critical in NMDS.
- Although linearity is not assumed at all, for gradients with high β-diversity the order of distances can only be preserved by forcing an arched arrangement of points upon the ordination. The method of shortest path adjustment described in the previous Subsection may improve the results greatly (e.g., Podani 1994).
- When attempting to interpret the axes one should bear in mind that – contrary to PCA and PCoA axes – the correlation between NMDS axes may be different from zero! As mentioned above, the entire configuration may be rotated which led some authors to propose performing a PCA from the NMDS coordinates. Identification of dimensions with external variables is therefore more problematic than in case of metric methods.

The NMDS of the phytosociological data of Table A1 was performed several times from different random configurations, for two dimensions, and based on the Euclidean distances of relevés calculated from the raw data. (In this regard, the example is not demonstrative of a typical application with no raw data being available.) The result giving the minimum stress was chosen for illustration (Figure 7.21a). $ST = 0.006$, indicating that the two-dimensional arrangement almost completely preserves the order of input distances. The good fit of points to a line in the Shepard diagram (Fig. 7.21b) is a confirmation of the above findings. A major diffrence from a logically comparable metric ordination (centred PCA, Fig. 7.2) is that the distances are more balanced in the NMDS case: large distances are diminished and small ones are increased.[9] Thus, the points are scattered more evenly in the non-metric ordination, a striking manifestation of monotone regression and of the attempt of depicting the relationships in two dimensions only. Otherwise, the NMDS result does not add anything new to the PCA ordination. Yet, it is a useful result because application of a completely different ordination criterion confirmed our earlier findings.

9 The results of standardized PCA are obviously incompatible with the NMDS ordination if the distances are calculated from raw data. Such a comparison would involve the change of two factors: the ordination procedure and the data type. If two or more things are modified simultaneously, a single comparison cannot reveal the causes of the difference. More details will be given in Chapter 9 ('complex comparisons').

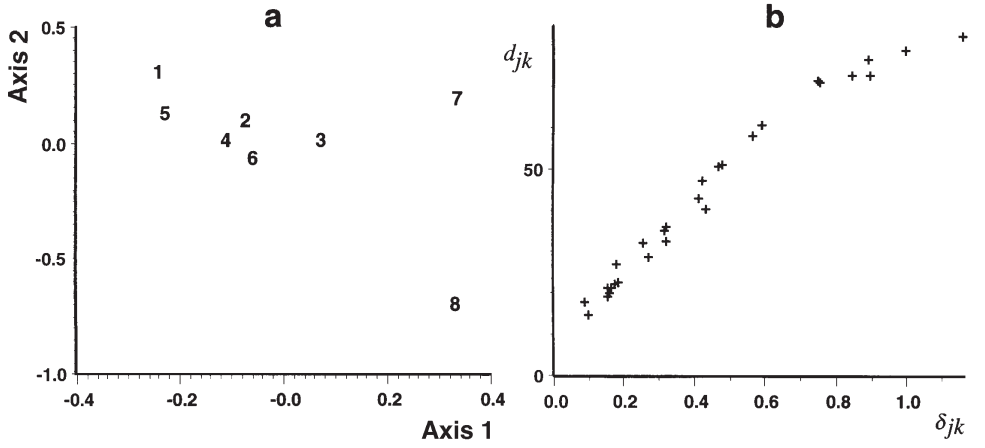

Figure 7.21. NMDS ordination of phytosociological relevés (Table A1) (**a**) and the associated Shepard diagram (**b**).

Other methods of non-metric MDS. The scaling algorithm developed by Kruskal is only one possibility of arranging the points in a non-metric manner. There is a well-known modification of the standard algorithm by Sibson (1972), called *local* NMDS (abbreviated as LNMDS). In this, no attempt is made to preserve the order of all distances. Instead of this rigorous condition, a weaker requirement is adopted: the rank order of distances (dissimilarities) of each object from all the others is to be preserved maximally in the ordination. Thus, for the distance d_{jk} it is immaterial how d_{lm} is positioned in the global ranking, but its relation to d_{jm} and d_{jl} does have meaning. In other words, the LNMDS algorithm examines the relationships for each object separately, hence the term 'local'. Prentice (1977) suggests that LNMDS may be more appropriate to reveal ecological gradients than NMDS, because a given distance difference may not be equally important ecologically at the beginning and the end of the gradient. For LNMDS, the Shepard diagram would be meaningless, of course.

Reproducing distance orderings is just one objective in non-metric scaling. There is another group of methods, '*continuity analysis'* or *'parametric mapping' (*Shepard & Carroll 1966, Noy-Meir 1974) which utilizes a completely different criterion. A minimum number of new dimensions are determined in which the variables (e.g., species) provide a function to which the objects have the best fit. There is no assumption as to the properties of the functions to be used. In this case, as always in regression, the sum of squared deviations from the expectations is minimized. Continuity in this case refers to the goodness (or smoothness) of fit of points to a multidimensional surface.

7.5 Separating groups: canonical variates analysis

All ordination procedures discussed thus far treat the objects as a single group, whereas CCA, RDA and CCOA distinguish between two groups of variables. A possibility not yet exploited is that the objects have an *a priori* classification into $k \geq 2$ groups based on some external criterion or decision rule not derived from the data directly. Then, we have the following ordination problem: linearly uncorrelated axes are to be determined so as to best explain the *separation* of these groups, neglecting within-group tendencies. In other words, between group variances are to be maximized and within-group varainces minimized by the axes, rather than minimizing the total variance as usual in PCA (Mardia et al. 1979). If the same data set is evaluated in both ways then, depending on the grouping of objects, the ordination axes may differ considerably, as illustrated by the deliberately contrasting cases of Figure 7.22. The ordination procedure that maximizes group separation has been known as discriminant analysis or *canonical variates analysis* (CVA). The label CVA usually refers to applications in which reduction of dimensionality is the primary objective. The terms linear discriminant function analysis (LDFA) and multigroup discriminant analysis (MDA), however, appear to be confined to cases when determining the most discriminative variables and assigning new objects into one of the existing groups are attempted, and ordination diagrams are not even made. In this book, the ordination objective is emphasized and the abbreviation CVA will be used. For more traditionally oriented readers, it is noted that CVA has close connections to the multivariate analysis of variance (MANOVA) which is not discussed here.

CVA is based on cross-products matrices computed between variables (Equation 3.86 based on centred data, or Equation 3.69 without division by m–1). These arrays are called *dis-*

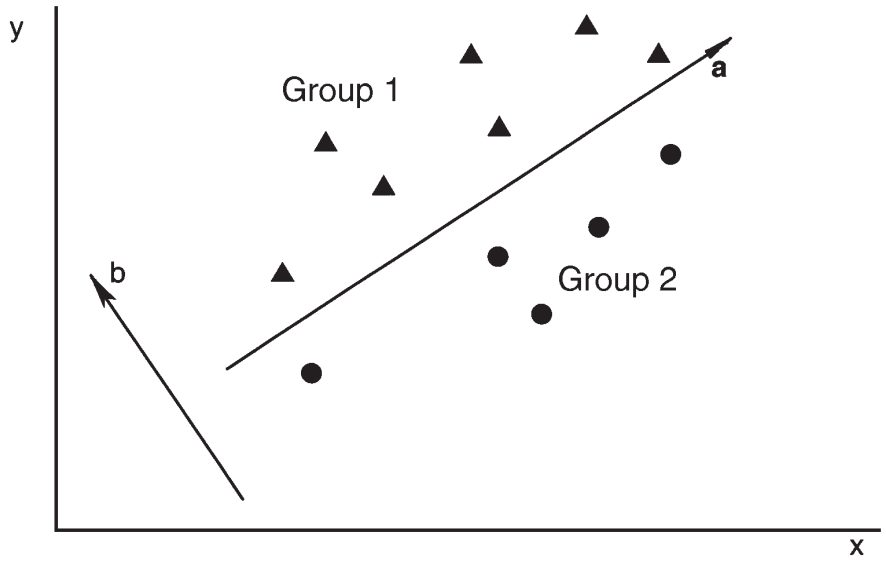

Figure 7.22. Comparison of the underlying ideas in PCA and CVA by an artificial example with two original dimensions. Component 1 (**a**) coincides with the main trend of variation in the entire sample, whereas canonical variate 1 (**b**, there is only one in this case) explains the optimum separation of the two groups.

persion matrices, all having a size of $n \backslash n$. Matrix **T** refers to the entire set of objects, regardless of the group membership of the objects (*total* dispersion). On the other hand, matrices $\mathbf{W}_1$, $\mathbf{W}_2$,...,$\mathbf{W}_k$ are determined separately for each of the k groups, such that centring is done using within-group averages. These *within-group* dispersion matrices may be summed, resulting in the *pooled within group* dispersion matrix, $\mathbf{W} = \sum_{i=1}^{k} \mathbf{W}_i$. The portion of cross products which explains differences between groups is simply obtained by subtraction:

$$\mathbf{A} = \mathbf{T} - \mathbf{W}. \tag{7.67}$$

At this point, relying upon knowledge on analysis of variance from conventional biometry, one might suggest that some sort of the ratio of between- and within-group dispersion should be maximized. However, division of one matrix by another is not allowed (the operation **'A/W'** does not exist in matrix algebra, Appendix C), but premultiplication of **A** by the inverse of **W** will provide the desired result to express proportionality of the two variance components. The matrix thus obtained is subjected to eigenanalysis according to the following equation:

$$(\mathbf{W}^{-1}\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}\,. \tag{7.68}$$

The analogy with PCA becomes now obvious, with the difference that the new axes coincide with maximum separation, rather than with maximum total variance, as already mentioned. Further difference from PCA is that although matrices **A** and **W** are both symmetric, the product $\mathbf{W}^{-1}\mathbf{A}$ is not, so that the resulting eigenvectors, although linearly uncorrelated, will not be orthogonal to one another. As a consequence, it is not straightforward to illustrate the results in an orthogonal coordinate system notwithstanding that the distortion may be diminished by an appropriate transformation (Equation 7.70).

The $\mathbf{v}_j$ eigenvectors are extracted in the usual manner: each is normalized to unit length. The coordinates of objects on the canonical axes do not derive directly from the eigenvectors, however. First, the canonical variates or *weights*, abbreviated as $\mathbf{c}_j$ are to be determined. The first proponents of CVA (Cooley & Lohnes 1971) advocated the use of the following transformation of eigenvectors:

$$\mathbf{c}_j = \frac{\mathbf{v}_j}{\left(\mathbf{v}'_j \dfrac{\mathbf{T}}{m-1} \mathbf{v}_j\right)^{1/2}} \tag{7.69}$$

in which m is the number of objects (as usual), and $\mathbf{T}/(m–1)$ is the total variance/covariance matrix (note that the denominator in the above equation is a scalar). As a result, the total variance will be the same on each of the canonical axes and the within-group dispersion is unequally reflected by the axes (Fig. 7.23a). Consequently, the scatter of groups becomes far too elongated in the canonical space, thus overemphasizing less important dimensions. If the eigenvectors are normalized using the pooled within-group variance/covariance matrix $\mathbf{W}/(m–k)$, as suggested by Mardia et al. (1979):

$$\mathbf{c}_j = \frac{\mathbf{v}_j}{\left(\mathbf{v}'_j \frac{\mathbf{W}}{m-k} \mathbf{v}_j \right)^{1/2}} \tag{7.70}$$

then the variance will no longer be the same on the canonical axes, and group separation becomes more pronounced. Furthermore, the contribution of each axis to within-group variance will be identical and the point scatter of each group becomes spherical (Fig. 7.23b). The transformation, called the *spherizing* appears to be a good choice because canonical variates are expected to maximize between group dispersion, rather than within-group varainces. The lack of orthogonality implies much less distortion for spherical scatters than elongated ones.

Having determined the canonical variates, the coordinate of object *s* on axis *j* is obtained using the centred data according to the equation:

$$e_{js} = \sum_{i=1}^{n} c_{ij}(x_{is} - \bar{x}_i), \tag{7.71}$$

where $\bar{x}_i$ is the mean of variable *i* in the entire data matrix. Centring implies that the centroid of ordination scores will coincide with the intersection of canonical axes.

When evaluating CVA results, the following points deserve particular attention:

- The eigenvalues obey the following relationship:

  $$\lambda_j = \frac{\mathbf{v}'_j \mathbf{A} \mathbf{v}_j}{\mathbf{v}'_j \mathbf{W} \mathbf{v}_j}. \tag{7.72}$$

  The eigenvalues are not influenced by rescaling canonical variates according to Equations 7.69-70. The ratio

  $$\frac{\lambda_j}{\mathrm{tr}(\mathbf{W}^{-1}\mathbf{A})} = \frac{\lambda_j}{\sum_{i=1}^{q} \lambda_i} = \frac{\lambda_j}{T^2} \tag{7.73}$$

  reflects the proportion of total between-group variance falling onto canonical variate *j* (Mardia et al. 1979). q is the number of canonical variates, as explained in the next paragraph. The denominator, i.e., the sum of eigenvalues (being equal to the trace of the matrix product) is generally known as Hotelling's T^2 in the literature. This quantity is used to test the statistical significance of between-group differences (nevertheless, there is another option proposed by Bartlett). The meaning of T^2 is further clarified in terms of the generalized distance (Equation 3.96): T^2 is the weighted average of the generalized distances between group centroids and the grand mean. Weighting implies that the larger the group, the more influential is its distance in contributing to T^2.

- The number of linearly uncorrelated, yet not necessarily orthogonal canonical axes is $q = \min\{k–1, n\}$. That is, when the number of groups is smaller than the number of variables, k–1 axes are sufficient to explain the relationships for k groups. Therefore, for a two-dimensional scattergram to be made we need a minimum of three groups. It is also obvious that q cannot be larger than the number of original variables. Note also that the number of variables cannot exceed the number of objects, since in this case

matrix **W** could not be inverted (Appendix C). The larger the number of objects in comparison with the number of variables, the more efficient is the ordination.

- If two strict conditions, the multivariate within-group normality of variables and the homogeneity of within-group variances/covariances (homoscedascity) are satisfied, then the canonical variates can be tested for significance. Bartlett's test (Cooley & Lohnes 1971) is suitable to determine whether the *q-p* canonical variates, remaining after removal of the first *p* variates, contribute significantly to group separation. The statistic is calculated as follows:

$$X^2 = -\left(m - 1 - \frac{n+k}{2}\right) \ln \left(\prod_{j=p+1}^{q} \frac{1}{1+\lambda_j} \right)$$

(7.74)

which follows the χ^2 -distribution with $(n-p)(k-p-1)$ degrees of freedom. Therefore, if we wish to ascertain the significance of canonical variate 1, then the above statistic is to be calculated for $p = 0$, and the result compared with the entries of the table of χ^2 values. When the statistic exceeds the threshold at a given probability level, the variate indicates significant group separation. If the threshold is not reached, then all subsequent axes will also be non-significant, so that there is no need to test for the other *p* values.

- The above formula includes Wilk's Λ:

$$\Lambda = \prod_{j=1}^{q} \frac{1}{1+\lambda_j} = \frac{|\mathbf{W}|}{|\mathbf{T}|}$$

(7.75)

whose value ranges from 0 (= maximum group separation) to 1 (= group centroids are indistinguishable statistically).

The reader can ask the question now: given a small Λ or a significant result for Bartlett's test, how can we identify pairs of groups that statistically differ? This situation is analogous to the *a posteriori* comparisons in univariate analysis of variance when the least significant difference (LSD) is calculated. It is not the objective of this book to provide insight into statistical hypothesis testing, yet it is noted that selecting pairs of signfiicantly different groups raises many theoretical and practical difficulties. A method that works in the case of two groups (e.g., F-test) cannot simply be extended to more groups, because selecting pairs on the same criterion would lead to the accumulation of Type I errors, and therefore to false conclusions (Bonferroni problem). The book returns to this issue, in a different context, in Chapter 9.

- CVA is a special case of CCA (Bartlett 1938, Cooley & Lohnes 1971: 249, ter Braak & Prentice 1988). To see this, let us introduce *k* binary group membership variables such that $g_{hi} = 1$ if object *h* belongs to group *i* and $g_{hi} = 0$ otherwise. If the left domain variables are the original descriptors in the data set and the right domain is composed of these indicator variables, CCA will provide results identical to those obtained by CVA. The canonical variates are linear combinations of the original variables which maximally correlate with the linear combinations of group membership variables. The *j*th canonical correlation from CVA is obtained as:

$$R_j = \left(\frac{\lambda_j}{1+\lambda_j} \right)^{1/2} \tag{7.76}$$

Its absolute value equals the canonical correlation calculated by CCA in the above mentioned manner (Equation 7.23). The stronger the separation of groups along an axis, the higher the value of coefficient 7.76.

- If **R** is the correlation matrix of *n* variables, then the correlations between canonical variate *j* and the original variables (*structure coefficients* or *loadings")* are calculated acording to*:*

  $$\mathbf{s}_j = \mathbf{R}\mathbf{c}_j \tag{7.77}$$

  where $\mathbf{c}_j$ is obtained from Equation 7.69. These correlations are not affected by the normalization of eigenvectors (Equations 7.69 and 7.70). The above relationship loses its validity if $\mathbf{c}_j$ is calculated using Formula 7.70. In the alternative CCA, these correlations correspond to the correlations of original (left domain) variables with their own canonical variate (Equation 7.26). As a consequence of this, in the CVA ordination the relative point positions are the same as in the CCA ordination of the same points based on the left set of variables.

  Correlations obtained by Equation 7.77 can be used to select the original characters best discriminating among the groups. Needless to say that these characters have high interpretive value. The CVA scores of objects and the correlations of characters with canonical variates can be used to construct biplots as well. In a biplot, the coordinates may be arbitrary in a sense that object and variable positions are superimposed after rescaling. In this case, the direction of arrows and their relative lengths are meaningful only. Dillon & Goldstein (1984) proposed to multiply the correlations with the corresponding univariate F-ratios to express the differences among variables more faithfully.

  Computer program printouts from CVA often begin with a list of univariate F-ratios. The F_i value is the between-group variance of variable *i* divided by the within-group variance of the same variable, so its magnitude gives useful preliminary information on the variables with high discriminatory power. It is most likely that variables with high F-ratios will have the highest correlations with canonical variates.

- The percentage *share* of canonical variate *j* from the variance of the correlation matrix (tr {**R**} = *n*) is obtained as follows:

  $$100 \square \quad \frac{\sum_{i=1}^{n} s_{ij}^2}{n} \tag{7.78}$$

  In the rare event of $(k-1) \mid n$, the variance tr {**R**} is entirely accounted for by the canonical variates, so that the cumulative percentage reaches 100%. In general, however, the cumulative percentage for all variates remains below 100%.

Table 7.2. A summary of the CVA of *Iris* data (Table A2). The canonical variates are normalized according to Equation 7.70; the other values of the table are not affected by normalization.

Variable	*F* - ratio	Correlation with variate 1	Correlation with variate 2	Canonical variate 1	Canonical variate 2	Communality
Sepal length	118	–0.791	0.206	0.723	–0.107	0.668
Sepal width	48	0.521	0.765	0.157	0.224	0.857
Petal length	1180	–0.985	0.046	–0.212	–0.834	0.973
Petal width	960	–0.973	0.221	–0.285	–0.274	0.996
Canonical correlation				0.985	0.475	
Eigenvalue				31.83	0.29	
Between-grop variance (%)				99.09	0.91	
Contribution to correlation				70.36	16.97	

- The *communality* of variable *i* is derived by the formula:

$$h_i = \sum_{j=1}^{q} s_{ij}^2 \qquad (7.79)$$

It is of interpretive value only if $(k–1) < n$, otherwise all communalities are 1 and therefore irrelevant. The variance of variables with low communality is not explained even by the entire set of canonical variates, and these variables usually convey negligible information on group separation. On the contrary, variables with communalities close to 1 are particularly important in this regard.

- In the canonical space, it is useful to show the position of the centroids of groups as well. If the coordinates of objects are derived by Equation 7.69, then the point scatter will be approximately spherical for each group and the *isodensity circle* of the given group can be drawn around the centroid. Its radius is obtained as $r = \sqrt{4\,\chi^2_{2,\alpha}} \,/\, 2$ (Giri 1977, see also Dillon & Goldstein 1984). At the significance level of $\alpha = 0.05$, which is most commonly adopted in biology, the radius is 2.45 units. The isodensity circle is *expected* to contain 95% of the members of the group, considered now as a statistical population. As seen, the radius is independent of the number of objects analyzed, so all groups will have the same isodensity circle. There is another circle, with variable radius, which may also be drawn into the ordination diagram. This is the *confidence circle* which is expected to contain the 'true' group mean with a probability of $100(1–\alpha)\%$. Its radius is $r = \left(\frac{\chi^2_{2,\alpha}}{m_i} \right)^{1/2}$, where m_i is the number of elements in group *i* (Mardia et al. 1979). Needless to say that both types of circles are meaningful only if the conditions of multivariate normality, the homogeneity of variances/covariances and random sampling are satisfied.

The most straightforward illustration of CVA is via by the *Iris* data set, because these scores were used originally by Fischer (1936) when introducing the methodology of

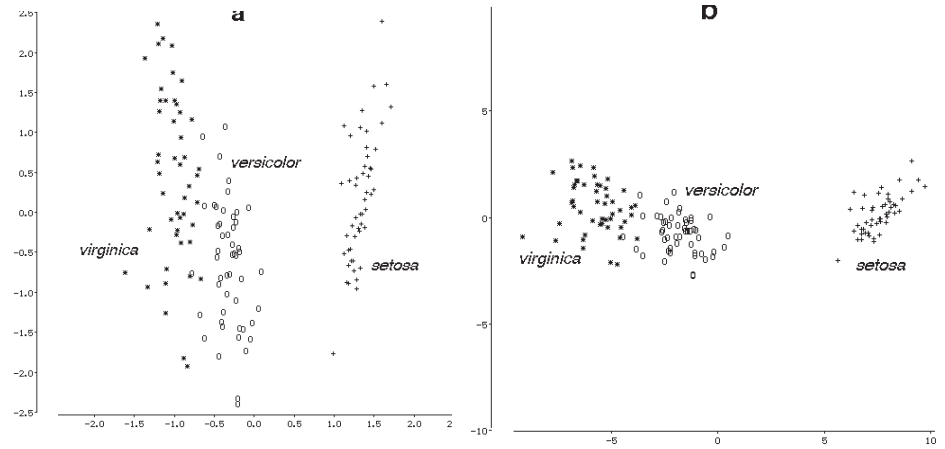

Figure 7.23. The CVA ordination of three *Iris* species (Table A2) using two methods of normalization **a:** normalization according to the total dispersion (Equation 7.69); **b**: normalization with the pooled within-group dispersion (variance/covariance) matrix (Equation 7.70). Symbols: + *Iris setosa,* O: *Iris versicolor*, *: *Iris virginica.*

discriminant analysis. There are three *a priori* groups, the three species, so that CVA can be used to evaluate their separation based on the four floral measurements. The difference between normalizations (7.69) and (7.70) is striking in Figures 7.23a and b (if the scale is the same on both axes, of course).

Iris setosa separates clearly from the other two species, irrespective of the method of normalization. This finding confirms our previous results obtained by PCA (Fig. 7.7) and fuzzy *c*-means clustering (Fig. 4.9). For *I. versicolor* and *virginica*, the separation is stronger on CV1 than on any PCA axis. Correspondingly, the first canonical correlation is very high and the between-group variance is almost completely explained by CV1 (Table 7.2). The second

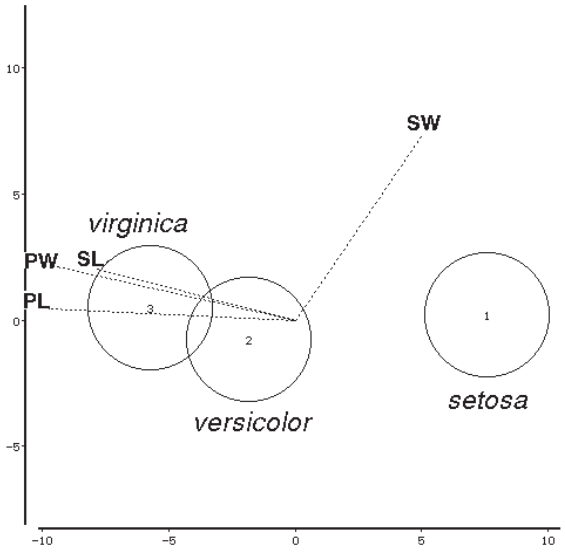

Figure 7.24. The CVA biplot of *Iris* data. The species centroids are: 1: *I. setosa*, 2: *I. versicolor*, and 3: *I. virginica*. Isodensity circles calculated for the two canonical variates are drawn around the centroids. The discriminative power of each variable is ascertained if the corresponding arrow is elongated, and the circles are projected onto the line thus obtained. Compare the results with the PCA biplot (Fig. 7.7)!

canonical correlation may also appear high by itself, but its associated between-group variance is less than 1%! Assuming that all conditions of a significance test are met, the results of Bartlett's test are worth examining. For both axes, we obtain that X^2 is 545.2 (the critical χ^2 value at d.f. = 8 and $\alpha = 0.05$ is 15.5). It is therefore very unlikely that the three species (groups) are derived from the same *statistical* population and, in other words, there are significant differences between the species. After omitting axis 1, we obtain that $X^2 = 37.2$, which is still significant (d.f. = 3, $\alpha = 0.05$, critical $\chi^2 = 7.85$), showing that axis 2, although much weaker than axis 1, is also meaningful as far as species separation is concerned. These findings are confirmed by the isodensity circles (Fig. 7.24): *I. setosa* is entirely separated, whereas the other two species overlap on axis 1. On axis 2, *I. versicolor* appears to be discriminated weakly. The confidence circles are omitted from the figure because their radii are very small, indicating clear-cut separation of centroids as a further confirmation of the results of Bartlett's test.

In view of the CVA results, the evaluation of original measurements provides useful additional information for the taxonomist. The three species are best separated by the petal measurements (Table 7.2 and Fig. 7.24). They have very high correlations with CV1 and their F-ratios are also high. The length of sepals is less discriminative, and the separation of species is the weakest for sepal width. These are also apparent from the variable/variate correlations. The communalities agree well with the above interpretation.

7.6 Morphometric ordination

Different analyses of *Iris* floral characters have already touched upon a special field of biological data exploration, widely known as *morphometry*. The primary objective of this approach is the evaluation of the variability of shape and size of biological objects, with emphasis often placed on the separation of inherent size and shape components of biological forms. Admittedly, the methods of reducing dimensionality discussed thus far may be applied to these cases with considerable success. In fact, two decades ago ordination methods were almost exclusive in multivariate morphometric research, as demonstrated vividly in the classical monograph by Blackith & Reyment (1971). Recently, however, many highly specialized morphometric techniques have become available, allowing a much more sophisticated analysis and a more exhaustive biological interpretation of results (Rohlf & Marcus 1993). These methods – in addition to their primary goals – can be applied to data exploration in taxonomy and evolutionary biology, and most certainly deserve at least a short section in this book. Somewhat analogously to the current 'upheaval' in molecular cladistics, the revolution in morphometrics has resulted in a highly diversified and elaborated subject which is not easy to follow in some places. Therefore, in agreement with the title of this chapter, the present discussion will centre around the relationships between ordination and the new methodology only. Ample references will be given for the information of those wishing to get a much deeper insight into this rapidly developing area.

In the *Iris* studies – exemplified earlier in this chapter –, perhaps it escaped our attention that we were concerned with *distance* values: the variables were length measurements between particular points (apex, base, lateral extremes) determined unambiguously on the sepals and petals. This is the case in many other morphometric studies: the characters are provided by distances measured between points identified according to some obvious feature of shape

(e.g., endpoints or crossings of certain structures). These points may be referred to as *landmark*s.[11] Interpoint distances, however, do not convey sufficient information for a faithful reproduction of the original shape from the data. For a more representative and exhaustive elaboration of the entire shape of the biological objects, new and more sophisticated data types need to be introduced. "Sophisticated" does not always mean that the required methods are radically different from those we already know from the previous sections. Rather, it refers to substantial differences in data types, data collection and processing, whereas the analytical methods themselves are not always new for us. It is emphasized that the new data types – notwithstanding their advantages – do not render the 'traditional' distance-based morphometric approach entirely obsolete, as pointed out by Reyment (1990) and Marcus (1990, 1993).

7.6.1 Contour analysis

A most striking feature of the shape of organisms is their *outline* or *contour line.* Having expressed outlines in numerical form, their analysis becomes straightforward. Rohlf (1990a) provides a review of methods designed to *fit functions* to entire outlines (closed contours) and to boundary curves drawn between two landmarks (open countours). The parameters of the functions thus obtained are in turn taken as input data to conventional multivariate procedures. This approach completely neglects all characteristics within the outline and is thus restricted to objects that are extremely poor in interior features, such as shells of ostracods and some bivalves.

Our attention will be focused upon shapes characterized by closed contours, because in morphometric analysis these are more important – and more common – than the open curves. For each shape, a landmark is identified such that its biological meaning is exactly the same for all other shapes; that is, all these points are homologous. Starting from the pivot landmark, several points are located systematically along the curve such that the last point coincides with the first. The shapes are then described according to the length values pertaining to radii drawn from the centroid (or other central point) to the outline. Finding another true landmark on each object could also be very useful, because two landmarks are required to place the object in an orthogonal coordinate system unambiguously so as to allow their meaningful comparison. In this case, the *x, y* coordinates of points selected at regular intervals along the curve are used as shape descriptors. The mathematical tools for analysing shapes based on radius lengths and coordinates are summarized briefly as follows:

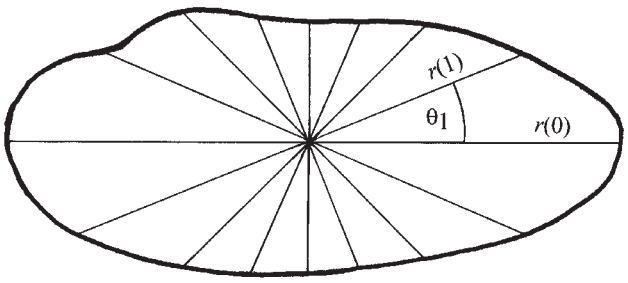

Figure 7.25. Radii drawn at equal angles to characterize the contour line of an *Unio pictorum* shell.

11 More precisely, a *landmark* is to be consistent with biological homology, and differences between extreme points are not always homologous. In such cases, reference is made to *pseudo-landmarks* (Rohlf & Marcus 1993). For a refined classification of landmark types, see Bookstein (1991).

- Some relatively simple[12] contour lines are described most easily by the *radius function* (Scott 1980, Lohmann & Schweitzer 1990). The reference line is the radius drawn from the centroid to the pivot landmark. From this, further radii are taken at equal angles and drawn to the contour line. The number of such radii is, say, p (Fig. 7.25). The radius function describes the relationship between the angle of rotation and the length of the corresponding radius in form of the pairs of values $[r, \theta]$. The shapes are fairly well approximated by p length scores, especially if p is high. No formula is determined explicitly for the radius function, however. Instead, the length data are summarized in a $p \setminus m$ matrix *(m* is the number of outlines) which is in turn subjected to standardized PCA. This is a peculiar application of PCA, because the correlation matrix is calculated among the objects, rather than the variables. More details of this so-called *eigenshape analysis* are discussed in Lohmann & Schweitzer (1990). Of the several resulting PCA scatter diagrams, the most interesting is perhaps the ordination of objects to which the component correlations are superimposed. The procedure is illustrated below by the eigenshape analysis of some European and Asian *Unio* (Mollusca, Bivalvia) shells.

  Four species are included in the study: three specimens collected from different localities will represent *U. pictorum* as well as *U. crassus,* whereas a single specimen represents each of *U. tumidus* and *U. elongatulus* (Table A8). Eigenshape analysis of radius data yielded a very high first eigenvalue, accounting for 97% of the total variance. This exceptionally large percentage is a reflection of the high overall similarity of contour lines. The smallest correlation was measured between *U. pictorum* and *U. crassus* (*COR* = 0.926), whereas the highest value was between *U. pictorum* and *U. tumidus* (*COR* = 0.99). The large first eigenvalue is a general size component, and on axis 1 all the eight shapes have high scores (between 0.971 and 0.994), so that this unipolar component has no interpretive value and is not illustrated. Thus, although they explain a small portion of the variance, the second and third components get into focus (Fig. 7.26a).

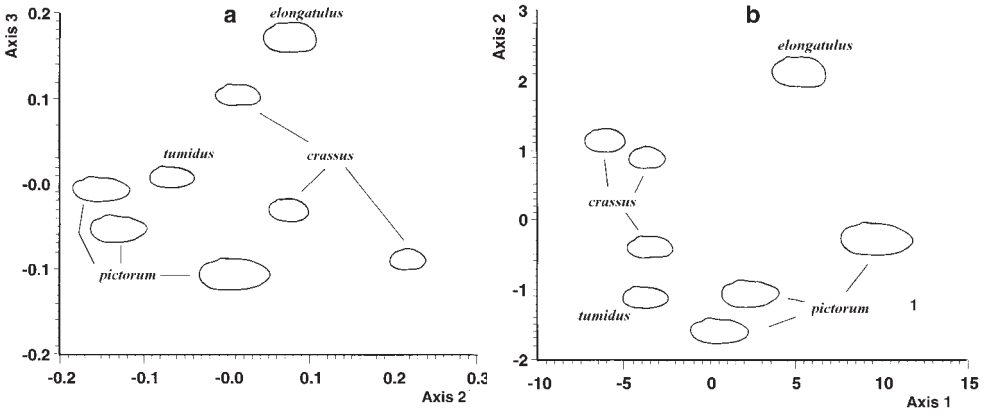

Figure 7.26. Principal component analysis of *Unio* contour lines, **a:** based on correlations applied formally to compare *Unio* individuals directly (shape components 2 and 3), **b:** based on correlations between radii (components 1 and 2). Note that in figure **b,** the scale radically differs on the two axes!

12 The meaning of 'relative simplicity' becomes more clear later on, in the discussion of the third procedure.

One may raise the possibility of a 'conventional' standardized PCA with the radius lengths as variables and the shell specimens as objects. In this case, the first eigenvalue is still high, although smaller than in the above eigenshape analysis (91.3%). The second component accounts for a mere 5.2% (Fig. 7.26b).

The data analyst is faced with the dilemma: which result should be considered superior in an ordination study of *Unio* shapes? An obvious disadvantage of eigenshape analysis is that in our case the depicted point scatter accounts for only 2% of the total variance, so that a large portion of total variation falls to a general size component. The first two axes of standardized PCA, on the other hand, explain 96.5% such that the individuals are reasonably evenly dispersed in the ordination space. None of the ordinations appear to support high within-population homogeneity, however, suggesting that outlines by themselves are insufficient for an un- ambiguous discrimination among species.

- The radii as defined above can also be subjected to Fourier analysis (harmonic analysis). This approach utilizes the mathematical law that – accounding to Fourier, the great French mathematician – any 'curve' can be reproduced according to a set of simple wave functions (harmonic functions, Rohlf 1990a). The length of a radius $r(\theta)$, which is at an angle of θ to the reference line, may be approximated by the following series:

$$r(\theta) = a_0 + \sum_{i=1}^{k} a_i \cos i\theta + b_i \sin i\theta, \qquad (7.80)$$

where k is the number of harmonics calculated ($k < p/2$), and

$$a_0 = \sqrt{\frac{2}{p}} \sum_{j=1}^{p} r_j, \qquad a_i = \sqrt{\frac{2}{p}} \sum_{j=1}^{p} r_j \cos i\theta_j, \qquad b_i = \sqrt{\frac{2}{p}} \sum_{j=1}^{p} r_j \sin i\theta_j. \qquad (7.81a\text{-}c)$$

Fourier-analysis estimates for k harmonic functions the parameters a_i and b_i, which in turn can be used as an abstract and indirect representation of the shape. The quantity $h_i = a_i^2 + b_i^2$ is the *harmonic amplitude*, the relative'contribution' of the ith function to the contour line. The Fourier coefficients derived for several objects, assuming that the reference radius is always directed towards a homologous landmark, will provide a raw data matrix, a starting point for further multivariate analysis. If there is no homologous landmark, only the harmonic amplitudes can be considered for further analysis, but this implies information loss. The harmonic functions themselves rarely have biological significance, but they are appropriate for descriptive and, consequently, for ordination purposes (Rohlf 1993a).

In the evaluation of *Unio* valves, the standard PCA of Fourier coefficients produces a somewhat surprising result, as far as the relative magnitude of eigenvalues (λ_1= 32%, λ_2= 23% and λ_3 = 16%) is concerned. The scatter diagram (Fig. 7.27) does not differ that much from the previous results, yet it seems more readily interpretable than those. *U. pictorum* separates from the other species on the first axis, and *U. crassus* forms a relatively compact group on the opposite end of this axis. However, the separation of *U. elongatulus* is less emphasized than in Figure 7.26b. *U. tumidus* takes an intermediate position between *pictorum* and *crassus* in all ordinations.

- A potential 'drawback' of the above two procedures is that the definition of the barycenter (centroid) is completely arbitrary biologically, and the use of another, more logical reference basis could easily provide diverging results. And to more com-

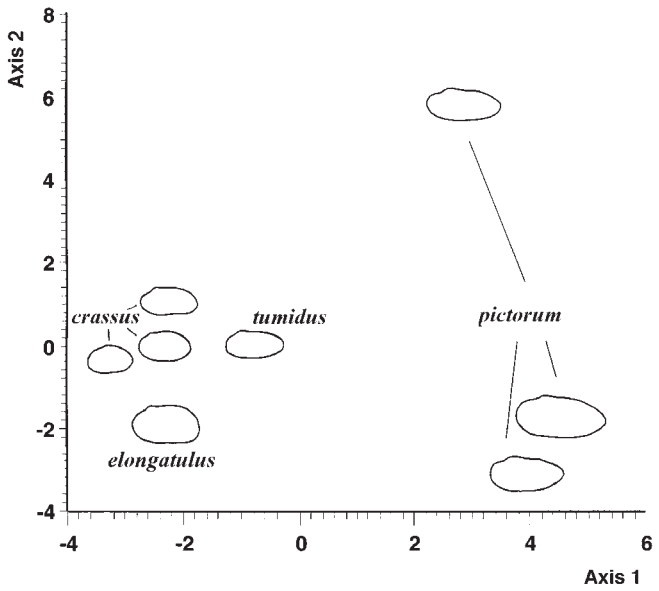

Figure 7.27. Standardized PCA of *Unio* valves based on Fourier coefficients fitted to the data of Table A8.

plicated shapes, in which the same radius intersects with the outline twice or more times, eigenshape and Fourier analyses are not suitable. The simplest, and perhaps the best-known resolution of these problems is offered by the shape function proposed by Zahn & Roskies (1972):

$$\varphi^*(t) = \varphi(t) - t \quad . \tag{7.82}$$

In this, t is the distance from a starting landmark (zero point) along the outline which is normalized to have a total length of 2π radians. $\varphi(t)$ is the angle between the tangent vectors drawn at point 0 and at the point with distance t from the zero point, again in radians. The function yields zero for any point along a circle, so that for other shapes the divergence from the circle, as the reference basis, is measured. Systematic 'sampling' of the outline is guaranteed by taking equal short intervals alng the contour, rather than by taking radii at equal angles. It is advisable to use at least 100 such segments (Reyment 1991). The values of the function $\varphi^*(t)$ are collected in a data matrix which is then subjected to eigenshape analysis (Lohmann 1983, Lohmann & Schweitzer 1990). The alternative discussed in the previous paragraph is also available here: the values are evaluated by Fourier analysis and subsequently by ordination procedures (Rohlf 1993a).

The calculation of the Zahn-Roskies shape fuction is illustrated on the example of the square (Fig. 7.28a). Starting from an arbitrary apex, 12 points are determined at equal intervals. Since the outline of the square is normalized to 2π radians, this interval equals $2\pi/12 = 0.52$. The reference line is the tangent drawn to point 1. The function $\varphi(t)$ is monotonically increasing (Fig. 7.28b), whereas the function $\varphi^*(t)$ oscillates around a line representing the circle (Fig. 7.28c), thus demonstrating the regular change of the difference between the circle and the square. It is easy to see that the value of $\varphi^*(t)$ is unaffected by the rotation of the square. The method, however, is not free from some problems. In addition to finding the homologous landmark, it is essential whether the contour lines are evaluated clockwise or counter-clockwise (in the case of the square, the direction does not matter, though).

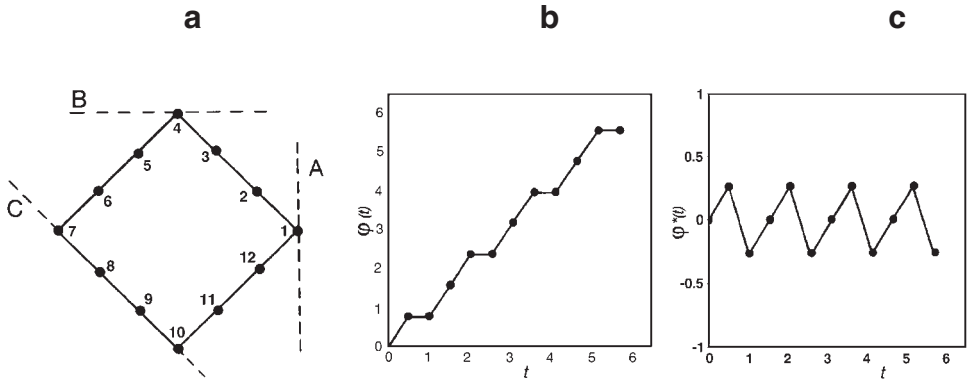

Figure 7.28. Determining the Zahn-Roskies shape function for the square. **a:** Twelve points selected on the outline. A is the tangent vector at point 1, B is the tangent vector at point 4, and C is the tangent vector at points 8 and 9. **b**: The function $\varphi(t)$,. **c:** The function $\varphi^*(t)$ for 12 points.

- A more general approach to the evaluation of contour lines is based on the so-called *elliptic Fourier-analysis* which starts from positions of points along the outline transformed into Cartesian coordinates. The 'independent variable' is the same as before: the distance along the diagonal normalized to the interval of $[0,2\pi]$. The function sought describes the simultaneous change of the coordinates (i.e., Δx and Δy) according to a set of superimposed harmonics (Kuhl & Giardina 1982). For each harmonic function, four Fourier-coefficients are derived (two for the horizontal, two for the vertical coordinates) and two additional constants are also required. The long and complicated formulae are not shown here, the reader may find them in Rohlf (1990a, 1993a). A relative advantage of the method is its independence from the direction of measurement and even from the position of the reference landmark (at least in its computerized realization in program **EFA**, Rohlf & Ferson 1992). The points need not be spaced evenly along the outline, and the algorithm applies to very complicated contours: crossings of the outline are allowed! Some illustrative applications to biological ordination are found in Rohlf & Archie (1984) and Ferson et al. (1985).

7.6.2 The use of landmarks in ordination

The analysis of contour lines has two fundamental disadvantages. The first, already mentioned problem is that features falling within the outline are completely neglected. Further criticism concerns the biological interpretation of the change of shape which is almost impossible in contour analysis (Bookstein 1991), because most measurement points are arbirarily placed. A solution is that we restrict ourselves to (biologically meaningful) landmarks selected over the entire shape; the data obtained this way are appropriate to conventional multivariate analyses and at the same time they allow more sophisticated investigations by the revolutionary methods of 'geometric morphometry'. In all cases, the objects are placed into a rectangular coordinate system and the landmark positions are expressed as vertical and horizontal coordinates.

- Comparison of shapes is possible through the direct comparison of coordinates. Two distant landmarks are chosen and the line segment between them is taken as the reference basis ('*baseline'*). Then, all shapes are placed into a coordinate system with their baseline coinciding the x axis, between the values of –0.5 and 0.5. Such a standardization produces Bookstein's (1991) *shape-coordinates*. If we have *p* landmarks, then the input for multivariate analysis incorporates 2(*p*–2) values for each object. An example for the application of Bookstein coordinates through discriminant and cluster analyses of mole cranial features is found in Loy et al. (1993), showing well the relative merits of this approach.
- Coordinates represent the direct input to another group of morphometric algorithms, the *superposition* methods as well. The task is the rotation and rescaling of one object to achieve the best fit of its homologous landmarks to the other object to which it is being compared.[13] The distance between two objects is conveniently defined as the sum of squared differences between homologous landmarks. This sum can be calculated for every pair of objects, resulting in an $m \backslash m$ distance matrix, a starting point for conventional multivariate analyses (e.g., Chapman 1990, Sanfilippo & Riedel 1990). The superposition methods have several variants; the reader is referred to Rohlf & Slice (1990), Chapman (1990) and Rolhf (1990b) for an evaluation of their relative merits and technical details. As the latter author concludes, superposition methods are most suitable to shapes characterized by relatively few landmarks, or to cases with differences distributed approximately randomly over the landmarks.
- Direct application of coordinates characterizes the most recent, synthetic approach to geometric morphometry. The roots can be traced back to Thompson's (1917) famous book in which shape changes of biological objects (e.g., skulls, leaves) are illustrated using a regular grid, as in Fig. 7.29. Since then, Thompson's attitude has for a long time been appreciated only at a descriptive level in different interpretations of biological shape changes. A few years ago, however, it turned out that statistical mechanics has many sophisticated tools for evaluating the change of shape, and their application to plant and animal forms facilitates biological interpretation greatly. The introduction to this subject would require large space and a more thorough knowledge of mathematics, so that in this book only a short discussion is presented. These tools appear in ordination context usually as a byproduct of shape analysis, nevertheless they deserve at least a short subsection for completeness. For more details, the reader is referred to Bookstein (1991), Rohlf & Bookstein (1990), Marcus et al. (1993, 1996) and Klingenberg & Bookstein (1998).

  Geometrical morphometry distinguishes between two main components of shape change. The *affine* (or uniform) changes include all transformations associated with the size, and the rotation and reflection of the object as well as the homogeneous compression or elongation of the shape in one direction (Fig. 7.29b). Non-affine changes or *deformations* have no particular direction, they are inhomogeneous, affecting each landmark differently. As a result, a regular

13 The method has been generally known as Procrustes analysis in the literature of multivariate statistics and has been used most commonly for the comparison of ordinations (the landmark positions in fact represent special two- or three-dimensional ordinations). The details of Procrustes analysis are discussed in Chapter 9.

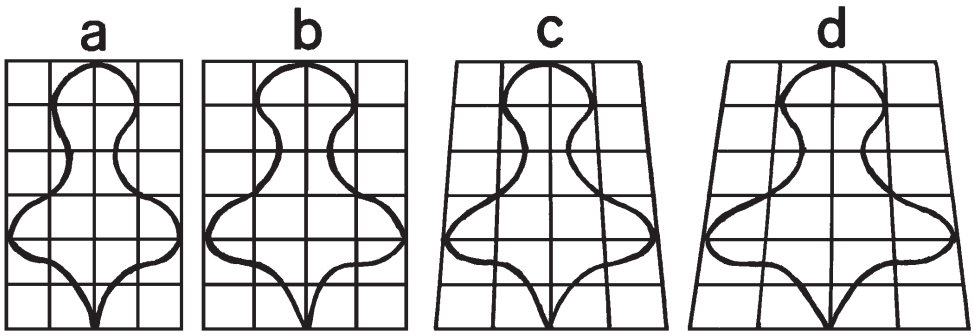

Fig. 7.29. The shape change of a hypothetical leaf **(a**) shown in the transformation grid of Thompson. **b:** homogeneous change, **c:** deformation, **d**: the superposition of the two types of change.

or systematic shape becomes irregular (Fig. 7.29c). If the set of study objects includes a reference object, such as a holotype or an average object obtained by generalized Procrustes analysis (Subsection 9.4.3) of all the other objects, then geometric morphometry may be used to separate these two components of change, often visualized by special ordinations.

As a good start, let us imagine a very thin and smooth metal plate with several points chosen on its surface, just like landmarks on biological objects. After bending this plate a little bit, we find that the points are shifted in vertical direction. The idealized energy necessary to this deformation is expressed by the binding energy matrix, $\mathbf{L}^{-1}$. This $p \times p$ matrix (where p is the number of points) is generated by inverting a matrix obtained from the interpoint distances and the coordinates of the reference object (see Rohlf 1993b: 137-138). For an affine change the energy is zero, since no bending was performed, only elongation or compression. The bending energy matrix can be formally computed for the reference object and a target object, as a special case. It is not to say that biological objects have a similar behaviour to metal plates, of course, so that this application is only formal. The $\mathbf{L}^{-1}$ matrix and the coordinates of the target object will provide the so-called. *thin plate spline,* an interpolation function which describes the mapping of the reference object to the target in terms of homogeneous and inhomogeneous components.[14]

The spectral analysis (Appendix C) of the energy matrix provides orthogonal vectors (‘*principal warps*‘); these are mathematical constructs analogous to principal components. The principal warps explain the deformation of shape at different geometric scales and can even be applied as taxonomic characters (Rohlf & Marcus 1993). The last three eigenvalues are always zero, so the associated vectors are unnecessary. The differences between the respective coordinates of the target and the reference objects, the principal warps and the eigenvalues are used to derive the ‘*partial warp scores’* of the target object. There are p-3 such scores for both the x and the y axis (Rohlf 1993b).

If our sample contains m objects, the partial warp scores are determined for each object separately. These scores are summarized in a vector of length $2(p–3)$, and then written into matrix **W** of size $m \times 2(p–3)$. This is the point where the standard multivariate techniques, such as principal components analysis, are called for. A PCA from matrix **W** (a method usually called the *relative warp analysis* in the literature) generates linear combinations of the differences between the objects and the reference. The PCA scores are then used to draw a scatter diagram of individuals. The use of PCA is not compulsory, of course, because canonical variates analysis or cluster analysis can also be performed on such input data. According

14 Of course, there is a three-dimensional case as well, with x, y and z coordinates, an opportunity not discussed in this book.

to a new proposition (Zelditch et al. 1995), cladistic analysis is also conceivable, after an appropriate transformation of coordinates. As a performance test, Naylor (1996) simulated the evolution of fishes and found that one of the trees obtained by 'morphometric cladistics' was in complete agreement with the true tree – indicating the future potential of this approach.

7.7 Literature review

The literature of ordination methods, like that of clustering, is extensive and difficult to encompass, even though we restrict ourselves to biological applications. In any case, if the title of a handbook includes the term "multivariate analysis", then it is granted that exploratory ordination procedures are discussed in great detail in that book. Reference to "multivariate statistics", on the other hand, will almost always indicate relationship to the concepts of formal statistics, such as hypothesis testing, multivariate normality and the like, a subject beyond the scope of the present book. The number of works discussing biological ordination in general is very high, therefore only a few can be mentioned here. For ecologists, the pioneering volumes edited by Whittaker (1973, 1978) still provide much useful information, especially on the 'golden age' of numerical vegetation science. For ecologists, Greig-Smith (1983) and Kershaw & Looney (1985) are also very useful. Gauch (1982) introduces the reader into the topic of ordination almost completely avoiding any mathematical formalism, with less success, I think. Advanced level introductions include, in order of difficulty, Ludwig & Reynolds (1988), Pielou (1984), Orlóci (1978), Legendre & Legendre (1987) and Digby & Kempton (1987). Notwithstanding the increased importance and popularity of certain ordination techniques, there is no book devoted exclusively to their ecological applications. There have been review articles (e.g., ter Braak & Prentice 1988) and good collections of relevant articles (ter Braak 1996). For taxonomists, Sneath & Sokal (1973) may be still considered as a good summary, whereas the short book by Dunn & Everitt (1982) remains at a more introductory level. More recent taxonomic monographs appear to forget about ordination, with the exception of Stuessy (1990). This is apparently a result of the overwhelming dominance of the tree-making approaches in contemporary systematics (but see a recent debate on the potentials of ordination in cladistic analysis [Parnell & Waldren 1996, and Faith 1997]). Of the more general – and more mathematical – texts, the exploratory function of ordinations is emphasized in Gordon (1981), whereas Cooley & Lohnes (1971), Mardia et al. (1979), Chatfield & Collins (1980) and Dillon & Goldstein (1984) provide a more formal, theory-oriented treatment of the subject matter.

Principal components analysis is introduced in great detail by Jolliffe (1986), although his examples are mostly non-biological. The relationship between PCA and other multivariate procedures, and the possibilities of their joint application, are deeply evaluated in this book. The discussion of PCA is of course part of all books on multivariate analysis, and there are hundreds of them. The illustrative examples by which the fundamentals of PCA are explained to the novice vary from book to book. Contrary to this chapter, for example, Jongman et al. (1978) discuss PCA as a special case of the least squares method. Their iterative algorithm provides a good alternative to the geometric approach. Rao (1973; see also Bookstein 1991:39) points out that the matrix product of the vectors of component scores (whose sum of squares is made equal to the eigenvalue) yields a matrix of rank 1, whose elements provide the minimum sum of squared deviations from the starting covariance matrix. A recent summary of PCA is Jackson (1991), with a thorough discussion of biplot techniques, and an even more recent treatment of biplots is found in Gower (1996).

Although the topic of factor analysis was very briefly mentioned in this book, it does not mean that the method is always neglected in the biological sciences. For instance, Cattell (1978) focused his interest to biological applications only. Unfortunately, the title of Reyment &

Jöreskog's (1993) book ("*Applied Factor Analysis in the Natural Sciences*") is somewhat misleading, since the subject is more general in scope, while factor analysis *sensu stricto* ("True factor analysis") is discussed in a subsection only. Although Wright (1954) clarified quite early the differences between FA and PCA, the terminological confusion over 'factors' seems to prevail...

The only detailed treatment of canonical correspondence analysis, with plenty of biological examples, is Gittins (1985). This book lists many references for further orientation to those still interested in this fairly old procedure. The 'bible' of correspondence analysis for the English-speaking world is Greenacre (1984, but see also van Rijkevorsel et al. 1988), although other languages – and the pioneers of the area – should be also mentioned (Benzécri et al. 1973).

The popularity of canonical correspondence analysis has increased rapidly in the past decade. This becomes most apparent if one examines the complete bibliography of its applications between 1986-1993 (Birks et al. 1996). According to Reyment (1991), the method is not necessarily restricted to recent ecological objects, a statement supported by a relevant paleoecological example. Canonical variates analysis is discussed in most books on multivariate analysis, especially in Mardia et al. (1979), although the exploratory function of this method is sometimes subordinated to the statistical aspects of discmininating between groups.

The classical text of morphometric ordination is Blackith & Reyment (1971), a book still useful for all wishing to get insight into the early approaches to the evaluation of biological form. The developments since then have been reviewed in detail by Reyment (1990, 1991) himself, who makes the point that, although ordination procedures are less emphasized in the morphometric practice, they are still useful especially in the analysis of landmark data (see also Marcus 1990, 1993). Bookstein takes the opposite view, repeatedly underestimating the importance of descriptive, ordination-oriented exploratory approaches in morphometric analysis (e.g., Bookstein 1990, 1991, 1993). He asserts that ordination implies information loss at best, and is unsuitable to the biological interpretation of shape changes. He supports very strongly the discipline of geometric morphometry in his introductory – yet not easy – text (Bookstein 1991), which is a must for all readers interested in the new developments of shape analysis. Nevertheless, ordinations along with special axes remain of central importance even in this book. Novel results in the area are summarized in Marcus et al. (1996). The literature of geometric morphometry is still limited to a few books, monographs and conference proceedings, shown vividly by the fact that they are referred to very often by the color of their cover ('blue book', orange book', 'black book', Klingenberg & Bookstein 1998 will probably be the 'light gray' softcover). It is granted that the explosive development of geometric morphometry will soon run out of the available colours...

7.7.1 Computer programs

The general user of multivariate methods is very well treated by ordination software. Commercial statistical packages almost always contain some ordination techniques, even if the documentation does not always emphasize the exploratory aspect of the methods. In addition, there are many, more specific programs designed to biological applications. In Table 7.3, I list some available packages, admitting that this selection is far from being complete. Interested people can easily find more on the internet (see Appendix B).

When evaluating the performance of an ordination package, several aspects should be considered. For example, before purchasing a program it is advisable to get information on its graphical capabilities, because numerical results do not stand on their own in ordination studies. In PCA and CoA, for example, it is extremely important to have an immediate picture of

Table 7.3. Ordination procedures in computer progam packages.

Method	Statistica	SYN-TAX	NT-SYS	CANOCO	NuCoSA	BMDP
Principal component analysis	++	++	++	++	++	++
Factor analysis	++					++
Canonical correlation analysis	++	++	++	++		++
Redundancy analysis		++		++		
Correspondence analysis		++	++	++	++	++
Canonical correspondence analysis		++		+		
Principal coordinates analysis		++	++	++	+	
Non-metric multidimensional scaling	++	++	++		+	
Discriminant analysis	++	++	++	++		++

biplots and joint plots (as in **SYN-TAX**). Program **CANOCO**, which is famous for its wide selection of (metric) ordination options (with detrending), has no own graphic routines; the job is left to **CANODRAW 3.0** (Šmilauer 1992). This program produces high quality graphics (including biplots, triplots and other diagrams) that can immediately be used in publications and theses. **CANOCO** is especially recommended for users of canonical correspondence and redundancy analyses, while conventional metric ordinations (e.g., PCoA) are more easily accessible from other packages. The advantage of the graphic output of **Statistica** is its flexibility; any parts of the diagrams may be formatted individually, Further factor to consider is user friendliness: ease with the selection of options, the format of menus and windows, on-line help, and so on. In this regard, **Statistica** is a good choice. It includes several variants of factor analysis, distinguished well from PCA, but some other important procedures are missing (Table 7.3). (Non-metric multidimensional scaling is simply designated as multidimensional scaling in the user's manual, forgetting about the metric variant.) The command language of **BMDP** is perhaps the most cumbersome. The hardware requirement of programs is also essential, those listed in Table 7.3 require a DOS/WINDOWS environment.

Some other programs not mentioned in the table also deserve the biologists' attention. Orlóci (1978), Orlóci & Kenkel (1985) and Ludwig & Reynolds (1978) provide BASIC source code for many ordination methods. A list in FORTRAN is presented in the manual of **ORDIFLEX** (Gauch 1977), a general ordination package. The **MULVA-5** program-package (Wildi & Orlóci 1996) includes many ordination methods, and is best suited to the large data tables used by ecologists and phytosociologists. The most popular program for detrended correspondence analysis is still **DECORANA** (Hill 1979b), but **CANOCO** is more up-to-date in its options and user friendliness.

Morphometric data analysis is a completely different matter. Many programs have been designed to generate input data for morphometric ordination from images or coordinates of landmarks. A Macintosh program for eigenshape analysis has been developed by MacLeod (1993). Modern methods of geometric morphometry are provided by **tpsRelw** and **tpsSplin,** developed for WINDOWS systems. Some of the procedures are available through the **NT-SYS** package as well (further information may be obtained through the Internet, Appendix B).

7.8 Imaginary dialogue

Q: *Apparently, you do not claim that all ordination procedures are covered in this boo. I suppose it would be impossible anyway. Still, there is a procedure often mentioned in the ecological literature but I do not find any reference to it here. It is the so-called polar ordination. What on earth is this, and why is it neglected?*

A: Polar ordination has only historical importance; it has been very rarely used recently. The method was invented by ecologists (Bray & Curtis 1957) in the early period of the computer age when more advanced techniques, such as PCA, were unknown to the wide audience for obvious reasons. In polar ordination, the farthest two objects are selected first, based on the pairwise distances, and then these two objects are chosen as endpoints of the first ordination axis. It is implicitly assumed that there is a strong underlying gradient in the data, which is best represented by the most distant pair of sites. The position of all other objects is determined according to their distances from the endpoint objects. Having determined the first axis, a second axis is derived by selecting the next pair of most distant objects. More details are presented in Gauch (1982). Nevertheless, I warn you not to waste too much time on this: the method is obsolete, no question. If you still want to try it, you can find its program in the **NuCoSA** package (Tóthmérész 1996).

Q: *...other procedures that would deserve at least a sentence?*

A: There are many procedures that I did not even mention, for example, Gaussian ordination, maximum likelihood ordination, hybrid nonmetric multidimensional scaling, and so on. But do not forget, this book is not only on ordinations!

Q: *I am aware of fuzzy classifications, and I wonder if there exists fuzzy ordination as well.*

A: I do not understand the motivation behind this question; perhaps you 'extrapolate' from the previous chapters, as you did several times before. You might think that in a 'fuzzy ordination' the positions of points fluctuate within certain limits, on the analogy of fuzzy cluster membership weights. As far as I know, there is no method that would provide such uncertain scores directly. However, the consensus ordinations to be discussed in Chapter 9 (see Fig. 9.18) can be understood as fuzzy ordinations. In addition to this, ordinations can be constructed on the basis of fuzzy logic. Roberts (1986) raised first the suggestion that fuzzy sets can be used as input to ordinations. These are, in some sense, direct ordinations, because these fuzzy sets must summarize known or assumed relationships between species and the environment.

We could go even further: ordinations can be obtained even from classifications! Feoli & Zuccarello (1986) have some interesting suggestions, although their method did not receive many applications.

Q: *To put it shortly: ordination or classification?*

A: Yes, there were times in plant ecology when this question was thought to be appropriate. Recall the infamous discussions about the continuity of vegetation in the sixties; proponents of clustering and advocates of ordination appeared to take irreconcilable positions in this debate. Nowadays, there has been a general view that the joint application of ordinations and classifications tells us more on data structure than any method used individually. If you insist

on some order of importance, I can assure you that we should never classify without checking the results by ordinations, whereas it is not essential the other way round: ordinations can stand on their own without any classification.

Q: *Why do not you provide a key to the selection of ordination methods, similarly to the resemblance coefficients? This would be very helpful for the novice, I think.*

A: I agree that such a vehicle is useful, although some choices were already shown in the flowchart of Figure 0.1. Here follows an extended key, which is of course just one of the many possibilities: key construction is always arbitrary.

1a The objects or variables are assigned into groups *a priori* (canonical methods)..... **2**

1b There is no such grouping **5**

2a The objects are grouped into two or more clusters according to a criterion not included in the study, whereas there is no division for the variables .....*Discriminant analysis*

2b The variables from two groups, the objects form one **3**

3a There is a symmetric relationship between the two groups of variables .....*Canonical correlation analysis*

3b The first group of variables constrains the ordination based on the second set **4**

4a There is an assumed linear relationship between the variables of the second group ..... *Redundancy analysis*

4b Variables in group 2 have an unimodal response to the background gradient .....*Canonical correspondence analysis*

5a A dissimilarity (distance) matrix for the objects is available only, or the ordination of variables is irrelevant, even though we have access to the original data **6**

5b The ordinations of objects and variables are both required **7**

6a The metric information is retained in the ordination .....*Principal coordinates analysis*

6b The metric information is lost, only the rank order of dissimilarities matters .....*Non-metric multidimensional scaling*

7a The variance shared by variables is explained only .....*Factor analysis*

7b The ordination accounts for the total variance **8**

8a The data structure is approximately linear .....*Principal components analysis*

8b The data structure is unimodal, the scores are frequencies .....*Correspondence analysis*

This key does not imply that a single method is appropriate only in a given study. The same data set should always be analyzed by alternative methods, otherwise many properties of the data remain unrevealed.

Q: *It seems to me that the arch (or horseshoe) effect manifests itself only in ecological ordinations characterized by long background gradients with high beta diversity. Is there any danger of getting curved arrangements in taxonomic or morphological ordinations?*

A: This 'effect' is not exclusive to ecological ordinations and the only 'danger' is the inappropriate use of methods. For example, Reyment (1991: 51) shows the principal coordinates ordination of *Leptograpsus* crabs in which the individuals are arranged along an almost perfect

parabola, although the input data were linear! The authors explanation ("very high correlations among almost equal variables") is obviously insufficient. I have repeated the analysis with other coefficients, and it turned out soon that from Euclidean distances the correct arrangement along a single dimension resulted. Also, if PCoA was performed from *squared* Gower coefficients the data proved to be one-dimensional. Recall that PCoA has to be launched from squared distances or dissimilarities – so that the Gower-formula is in fact distance rather than dissimilarity (just like the Manhattan metric, which is implied in Gower's formula). The conclusion is that Reyment's result was a true artefact, arising from an incorrect use of the method.

I just came across a recent suggestion (De'ath 1999) which may circumvent the problem of detrending in ecological ordinations. The method of principal curves originally proposed by Hastie & Stuetzle (1989) uses an iterative algorithm to fit an irregular curve on the points of the ordination scattergram. This curve may then be used as an abstract representation of the underlying one-dimensional gradient.

Q: *The obligate question: can we construct series in ordination spaces?*

A: Yes, I expected this question, and – not very surprisingly – my answer is positive. In addition to series generated in the ordination space by data transformation or by regular changes applied to the sampling strategy, there are possibilities of modifying the ordination algorithm successively. That is, the primary series is in the ordination space, rather than in the data or the topogaphic space. In the discussion of correspondence analysis, you could already see that the value of α is freely modified within certain limits. Thus, we can generate a series of ordinations in the function of α, and this series will be more informative than a single ordination pertaining to an arbitrarily chosen value of α. Similar parameter has to be defined by the user in the catenation method proposed by Noy-Meir (1974). For creating the biplot, we can also imagine a series: recall the Euclidean and the Mahalanobis biplots which, as Jackson (1991) asserts, are only extreme cases of a 'biplot gradient' whose members can be obtained simply by successively changing an exponent.

Q: *The section on morphometric ordination 'hangs out' from this chapter, I think. You talk a lot more on new data types (contour, landmarks, etc.) than on ordination itself. The ordination procedures that could be applied to new morphometrics are in fact the same as those discussed previously.*

Q: I admit that you are more or less right! However, I felt that – provided that someone reads this book from the beginning page by page – the reader becomes ready to digest this subject only at this point. When the procedures of classification and ordination are known, then may come the more advanced subject. Using the new terminology, the description of shape in Chapter 2 would have been too early, I think.

Q: *What is the solution if I have more than the usual species \ sites data? I can easily imagine a project running for several years, providing data in form of a species \ sites \ points of time matrix. Evidently, you forgot about this possibility, while in biology the appearance of such tables may be quite common.*

A: You are right, I did not even mention this possibility yet. There are many possibilities for analyzing such *three-way* (rather than three-dimensional!) data arrays. First of all, the array

can be sliced into conventional matrices according to either time, space or species, and then these matrices can be analyzed in the usual manner. For example, if separate analyses are performed for data pertaining to different points of time, then the resulting ordinations may be compared in a subsequent meta-analysis (see Chapter 9) to reveal the temporal trends in your data. A "quick-and-dirty" procedure is to expand each two-way 'slice' of the data into a vector, put them into a matrix and then do an ordination from this new data file (an example was the relative warp analysis). More elegant is of course the use of the non-metric method of INDSCAL (Carroll & Chang 1970) developed for this very purpose. Alternatives include the three-way extensions of factor analysis (PARAFAC; Harshman 1970, Tucker 1972) and correspondence analysis (Carlier & Kroonenberg 1996). Needless to say that PCA also has a three-way form (Kroonenberg 1983). In these papers the terminology is not biological, making our job a little more difficult than ever. In this book, I have no space to present more details, however.