2

The data matrix and data transformation

(A small step forward..., but many things are yet to be done)

Sampling, as we have seen, implies that the sampling units are described in terms of variables. The data obtained by sampling can be laid out in a rectangular table with, say, the rows as variables and columns as sampling units. The previous chapter has illustrated this already, where the dichotomization of nominal variables was introduced. For a biologist, such a table will most often have the following form:

	Sampling unit 1	Sampling unit 2	Sampling unit 3
Length	12	14	10
Width	7	9	8
Height	10	9	12

In this simple example, three variables characterize three sampling units. The heart of this table, obtained after removing row and column labels, corresponds to the *data matrix,* a rectangular array of numbers. In the sequel, this matrix will be denoted by **X** (following a mathematical convention that matrix symbols are in boldface), that is:

$$\mathbf{X}_{n,m} = \begin{bmatrix} 12 & 14 & 10 \\ 7 & 9 & 8 \\ 10 & 9 & 12 \end{bmatrix} \tag{2.1}$$

As shown, the full matrix is given in square brackets, although large regular parentheses may also be used. (But always avoid the symbol |, which has been reserved for determinants in matrix algebra, see Appendix C). The value in the i-th row and j-th column of this matrix is denoted by x_{ij}. Throughout this book, the number of rows will be n, and the number of columns will be m. The subscript n,m refers to the size of the matrix. Appendix A contains several actual

and artificial data matrices (numbered from A1 to A8) to be used later for the illustration of multivariate methods.

The reader wishing to consult further literature on this subject should always clarify for each text whether the variables are given in the rows or in the columns of the matrix. Caution prevents misinterpretation of the formulae and derivations. Some books discussing multivariate analysis with much emphasis on mathematics (e.g., Chatfield & Collins 1980, Dillon & Goldstein 1984, Mardia et al. 1979, Reyment & Jöreskog 1993, Mirkin 1996, Greenacre 1984) treat the variables as columns, whereas others (Anderson 1958, Kendall 1975) as rows. The latter practice is generally accepted in the biologically oriented literature: species and taxonomic characters most often appear as rows (e.g., Pielou 1984, Orlóci 1979, Pimentel 1979, Sneath & Sokal 1973, to mention only a few).

2.1 The principle of attribute duality and the geometric meaning of data matrices

The basic units of the analysis (i.e., what we classify, for example) will be called the *objects* in the sequel. The plant or animal individuals of a taxonomic survey will generally appear as objects and their features as *variables*. Similarly, the quadrats laid down in the community will be usually the objects of the subsequent analysis, whereas the species detected in the quadrats will serve as variables. What we have said thus far agrees well with the previous definitions: *the sampling units are the objects, whereas the properties of sampling units are the variables*. In this case, sampling units can be conceived as points in a multidimensional space determined by variables as axes, that is, the matrix **X** contains the coordinates of m points in the n-dimensional (hyper)space (see Fig. 2.1a for $n = 3$).

The investigator may just as well be interested in disclosing the relationships among variables. For example, ecological studies often attempt to recognize species groups (guilds) in the community. Now, one is faced with the reverse situation: the variables (i.e., species) are the objects of the analysis, and the units (quadrats) will become the 'variables'. The sampling units

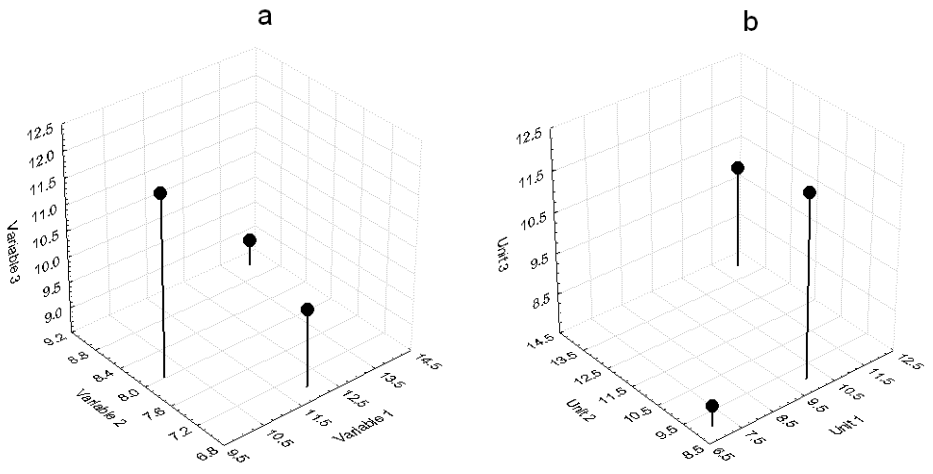


Figure 2.1. Alternative geometric representations of the data matrix (2.1). **a**: the axes are the rows and the points are the columns; **b**: the axes are the columns and the points are the rows.

act as simple replicates to facilitate measurement of the pairwise resemblance of species. The very same data matrix is now interpreted such that we have *n* points in an *m*-dimensional space (Figure 2.1b).

Keep in mind throughout that the structure of data (i.e., the relative positions of points, etc.) can be viewed in two ways for a given data matrix. This is the essence of a principle known as *attribute duality* (Williams & Dale 1965). They proposed to use the term 'attribute space' and 'individual space', referring to the alternative geometric representations of the matrix[1]. Ecologists (e.g., Gittins 1965) often refer to '*sample space*' if axes are sampling units, and to '*species space*' when the axes are species. Analogous spaces can most certainly be defined in other areas.

The terms '*R-mode*' and '*Q-mode*' analysis are widely used to distinguish between these two situations, in the first case interrelationships between variables, and in the second case interrelationships among objects are evaluated. I think, however, that such terminology often leads to unnecessary repetitions. For the majority of methods, it is immaterial whether the objects are points or axes; the variables and objects are interchangeable. For example, the same method of cluster analysis can be applied to both the rows or the columns of the matrix. In this book, therefore, these terms are not used, but cautionary notes will be given whenever interchangeability of objects and variables is questionable, illogical or inadmissible.

> An example is the product-moment correlation coefficient (Equation 3.70), which is meaningful primarily between characters, i.e., variables understood in the statistical sense, because calculation of the mean and the variance is involved. The correlation between two phytosociological quadrats or between two plant individuals, however, is more difficult to interpret, because the mean and variance are unclear for objects. Formally, the formula can be applied to such comparisons: a unit correlation results for two quadrats if the first has, say, twice as much cover of all species than the second. Also, the 'correlation' between two plants will be 1 if all size measurements for the first one are a constant times lower than for the second. That is, correlation appears to reflect proportions in size, but something seems to be incorrect when the variance is computed for a plant individual over all of its characters. An additional important difference is that although the correlation of two variables can be tested for significance (provided that the sample is random), the 'correlation' of two objects is not testable because the variables that describe them do not represent a 'random sample' (see Pielou 1984:8).
>
> It is certainly superfluous to give different names to a similarity coefficient depending on what constitute the axes of the space, as done by many texts. Consider, for example, Formula 3.25, which is called the Dice coefficient when we measure the association between species, and Sorensen index if the similarity of two quadrats is calculated. Goodall (1973a,b) lists many other instances for such double 'nomenclature'.

2.2 Displaying multivariate data structures by simple means

On the plane, only a two-dimensional configuration may be depicted, using the well-known Cartesian coordinate system. Yet, we attempted to show three dimensions in Figure 2.1, but such an illustration is never successful enough. The distances between points and their relative

1 Mirkin (1996) calls attention to a third possibility, the 'matrix space', which is *nxm* dimensional and the entire matrix is a point. This representation may be useful if we wish to find matrices that are in some sense the closest to the target matrix. Matrix comparisons (Chapter 9) are also associated with such spaces.

positions cannot be visualized perfectly, and if many more points were included the diagram would become puzzling. Four or more dimensions cannot be shown simultaneously by any means at all. The entire book is devoted to exactly this problem: how can we switch from a multidimensional space into a visible and sensible low- (preferably two-) dimensional arrangement? Before discussing mathematically more complicated techniques, it is useful to be acquainted with simple means of portraying multidimensional data structures in two dimensions. These methods use certain 'tricks' to circumvent the problem of dimension reduction, and they in fact do not apply to many (say, more than 50) variables.

2.2.1 Pictograms

The idea is to replace the objects by small pictures whose properties are determined by the original variables. The visualization is the most suggestive if the variables cannot be seen directly (e.g., species performance in ecological quadrats, measurements of environmental variables, etc.), but representation of plant or animal individuals by pictograms is less attractive, since in such cases just one real 'picture' would be replaced by another. These methods are rarely sufficient by themselves, although they can be used as supplementary tools to enhance, for example, interpretation of ordination scattergrams (by identifying the objects with pictograms). The variables are standardized by range (Formula 2.3) beforehand to ensure their commensurability.

The simplest pictorial representation is achieved by *star diagrams* and *Chernoff-faces*. In star diagrams each radial line corresponds to a variable, it is marked according to the standardized value, and the rays are connected to yield a star-like shape (Figure 2.2a). More interesting are the Chernoff-faces (Chernoff 1973); these rely on the ability of human perception

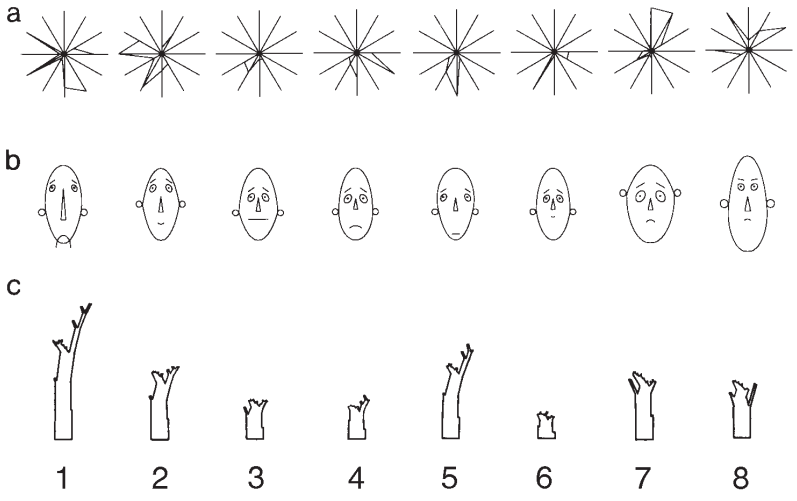


Figure 2.2. Different types of pictograms: star diagrams (**a**), Chernoff faces (**b**) and Kleiner - Hartigan trees (**c**) for the columns of Table A1. The trees in **c** were prepared using unstandardized data, the starting classification of 12 variables was complete linkage from Euclidean distances (Formula 3.47).

to distinguish among faces. Each property of the caricature-like pictures reflects an original variable, for instance, the length of the mouth is proportional to the first variable, the size of the ear with the second, and so on (Fig. 2.2b). Strict rules govern the preparation of drawings, but the interplay between facial characteristics may often have undesirable effects (for instance, the shape of a very small mouth is not striking enough, etc.).

The disadvantage of pictograms is that the correspondence between the original variables and the pictogram properties is completely arbitrary; so alternative arrangements can provide very different overall views. This is resolved by applying the Kleiner - Hartigan (1981) *trees.* The final branch lengths are proportional to the variables, and the length of intermediate branches, and of the trunk, is determined by the associated final branches (Fig. 2.2c). The correspondence between final branches and variables is determined through a hierarchical cluster analysis of variables (Chapter 5), otherwise it would be as arbitrary as above. Thus, a multivariate analysis must precede the preparation of such diagams, and a great deal of effort is needed to draw these trees.

2.2.2 Matrix of bivariate scattergrams

Another elementary technique is to project a multidimensional structure into all the possible planes each defined by two variables. For n variables we need $n(n–1)/2$ planar coordinate systems, so that, for example, a four dimensional data matrix is illustrated by six views. These scattergrams are very useful for inspecting the relationship between pairs of variables. If interchange of axes is allowed, we obtain twice as many diagrams, which can be arranged in matrix format (Fig. 2.3). The number of scattergrams is not n^2 because views in which the two axes represent the same variable are useless. Instead of these, the diagonal of the 'matrix' contains the frequency histograms (Hartigan 1975) or frequency polygons (Tukey & Tukey 1981a) of variables. The frequency histograms are worth examining at least by eye, especially if normality is a precondition of the analysis. Standardization of variables by their range is usually necessary.

2.2.3 Rotating diagrams

Rotating diagrams demonstrate very attractively a three-dimensional configuration on the plane of the computer screen (Tukey et al. 1976). The entire coordinate system and the points revolve around a fixed horizontal axis, thus giving the illusion of three dimensions. After a few rotations one may get an impression about the three-dimensional shape of the point cloud. The angle to the fixed axis can be modified gradually so that the viewer may find planes in the three-dimensional space that best reflect certain properties of the data (e.g., groups of points, linear trends, and so on, Fig. 2.4). The method is obviously limited to three variables at a time, so the number of variable triples to be examined can be very high if n is large. The method seems to perform better in illustrating three-dimensional ordinations (Chapter 7).

2.3 Data transformation and standardization

We have seen in the previous chapter that variables are often expessed in different units of measurement (lack of commensurability), although differences in magnitude may also be sig-

nificant even if the units are identical (implicit weighting). For this reason, multivariate data are very often evaluated in a format different from what we have recorded by sampling. Without data modification, the variables may have highly unequal influence upon the results, which

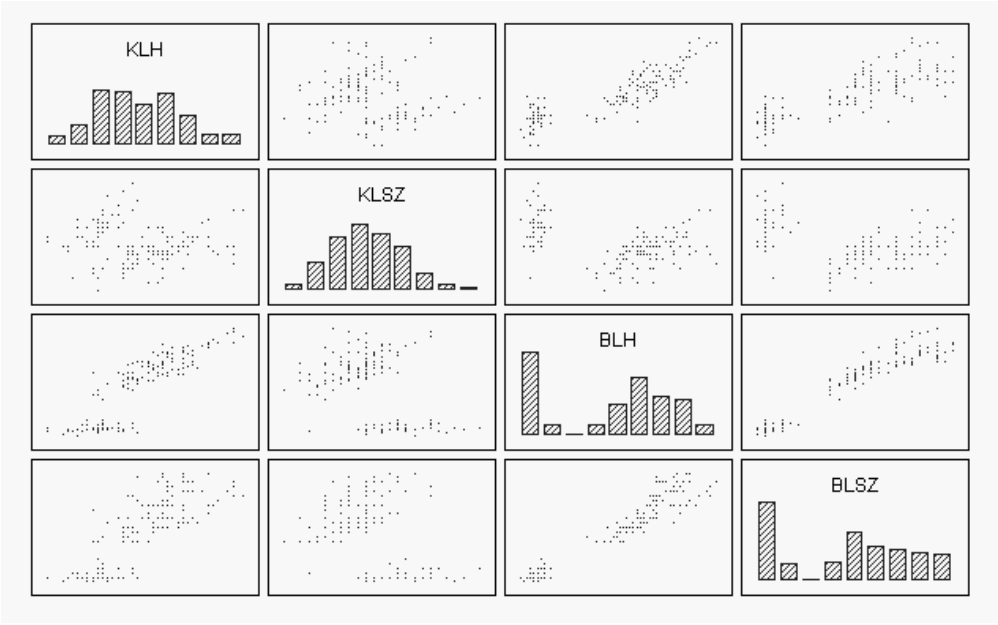


Figure 2.3. Matrix of bivariate scattergrams for the *Iris* data of Anderson (Table A2). Variables on the horizontal and vertical axes are named in the main diagonal. The individuals are arranged into two apparent groups and differences between distributional properties are also seen. Sepal width best approximates the normal distribution, but differences among species are the smallest for this variable. The more or less bimodal shape of the other histograms indicates the separation of one taxon from the other two.

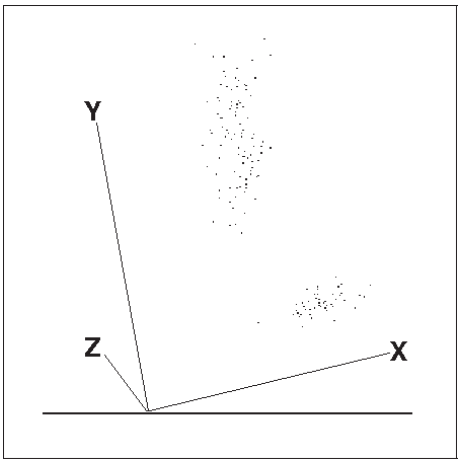


Figure 2.4. Rotating plot of the 150 *Iris* individuals as measured by Anderson (Table A2) for three characters. Rotation was arrested manually so as to maximize separation among the three taxa. X = sepal length, Y = petal length, Z = petal width. The horizontal line is the rotating axis.

is undesirable, unless this happens to be the purpose of the investigation. In ecological studies, elimination of magnitude differences among objects may also be very important in order to equalize their importance. Another good reason to transform the data is to bring the distribution of variables into a format acceptable to multivariate analysis methods. Typically, variables are transformed to reach approximate normality.

Note that in the above paragraph, variables and objects (= observations, sampling units) were understood as used in conventional statistics. This is important because, as shown below, certain transformations are feasible only for variables, so the principle of attribute duality has a limited validity for such manipulations. The methods used to modify the original data are therefore discussed separately for variables and objects.

Data can be modified in two basically different ways: *standardization* and *transformation*. One might argue that standardization is at the same time a transformation of some sort, but for our purposes it is reasonable to make a distinction which agrees well with the view of several authors (e.g., Sokal & Rohlf 1981a, Rohlf 1993a). Standardization modifies the data using statistics calculated from the data themselves; the procedure is therefore data-dependent. Such statistics are the variance, the range, the mean, the total or simply the maximum value. Standardization is often used to compensate for differences in weighting or measurement units. Data transformation, in the strict sense, uses a mathematical function whose parameters, if any, do not depend on the data. The distribution of variables is most often changed by transformation functions.

After data modofcation, the new value, obtained by changing the original score, x_{ij}, will be denoted by x'_{ij}. The procedures affecting the relative importance of variables will be demonstrated using an idealized spruce tree placed into the coordinate system of Figure 2.5a. The shape of the tree is described by two variables, the horizontal and vertical coordinates of points identified at characteristic locations along the circumference of the drawing (= 'landmark', Bookstein et al. 1985). The shape of organisms has been described such a way in a special field of numerical taxonomy, the morphometry. Differences in weighting will only be shown by changes of the tree shape, whereas methods more appropriate to changing the distribution of variables will also be exemplified by showing the original and modified frequency histograms (Fig. 2.7).

The raw data describing the tree, i.e., the coordinates of landmarks are:

2.65 3.35 0.00 2.70 3.30 6.00 1.00 2.75 3.25 5.00 1.75 2.80 3.20 4.25 2.25 2.85 3.15 3.75 3.00

0.00 0.00 2.00 2.25 2.25 2.00 3.80 4.00 4.00 3.80 5.25 5.40 5.40 5.25 6.75 7.00 7.00 6.75 8.00

Some of the resemblance coefficients (e.g., correlation, chord distance) to be discussed in the next chapter imply built-in data standardization. Thus, whenever such a function is chosen, no standardization of data is needed beforehand.

2.3.1 Standardization of variables

Centring. This is the simplest standardization method in which the mean (average) of the given variable is subtracted from each value:

$$x'_{ij} = x_{ij} - \overline{x}_i . \qquad (2.2)$$

In fact, the shape of the tree is unaffected by centring, and only the axes are shifted so as to move the origin of the coordinate system into the centroid of the tree (Fig. 2.5b). Centring is rarely used by itself, but is present in other standardization functions. Centring is part, for example, of calculating the covariance and correlation of variables (in principal components analysis and canonical correlation analysis, Chapter 7).

Linear standardization. The values of variable *i* are multiplied by a constant, a statistic that is derived from all the observations for the variable. In this example it means that the symmetry relations of the tree remain unchanged, while the shape is stretched or compressed in one direction, but is not distorted. This change is inversely proportional to the statistic being applied (e.g., range, standard error, etc.).

The first two procedures discussed below are not influenced if a constant is added to all values of a variable (i.e., the tree is translated by, say, 3 units to the left) before the standardization. They do not depend on the position of the 0 point and therefore apply to both interval and ratio scales. This is not so with the other methods; adding a constant will interact with the standardization, and they cannot be used for interval variables.

– *Standardization by range.* The variable is rescaled to the interval of [0,1]:

$$x'_{ij} = [\, x_{ij} - \min_j \{ x_{ij} \} \,] \,/\, [\, \max_j \{ x_{ij} \} - \min_j \{ x_{ij} \} \,] . \qquad (2.3)$$

The maximum and the minimum and their difference (i.e., the range) are determined for each variable. As a result, the variables are equilibrated, so the method is useful for eliminating implicit weighting as well as for ensuring commensurability.

The crown of the tree is expanded horizontally a little bit (Fig. 2.5c) because the two variables were different in range (6 for x_1 and and 8 for x_2). This operation is implicit in Formulae 3.103 and 3.104 developed for mixed data.

– *Standardization by standard deviation*. This is often referred to simply as standardization. As a result, the standard deviation becomes 1 and, due to the centring, the mean becomes 0 for each variable:

$$x'_{ij} = \{ x_{ij} - \overline{x}_i \} \,/\, s_i \qquad (2.4)$$

where

$$s_i = \left[\frac{\sum_{j=1}^{m} (x_{ij} - \overline{x}_i)^2}{m-1} \right]^{1/2} \qquad (2.5)$$

is the empirical standard deviation of variable *i* (calculated from the sample). The numerator is the sum of squared deviations from the mean, the denominator is the number of degrees of freedom. This approach is recommended if the variables are measured in very different units (pH, concentration, temperature, etc. in the same sample). The new unit of measurement will be unit deviation, so that all variables become commensurable. The product moment correlation coefficient (Formula 3.70) includes this operation.

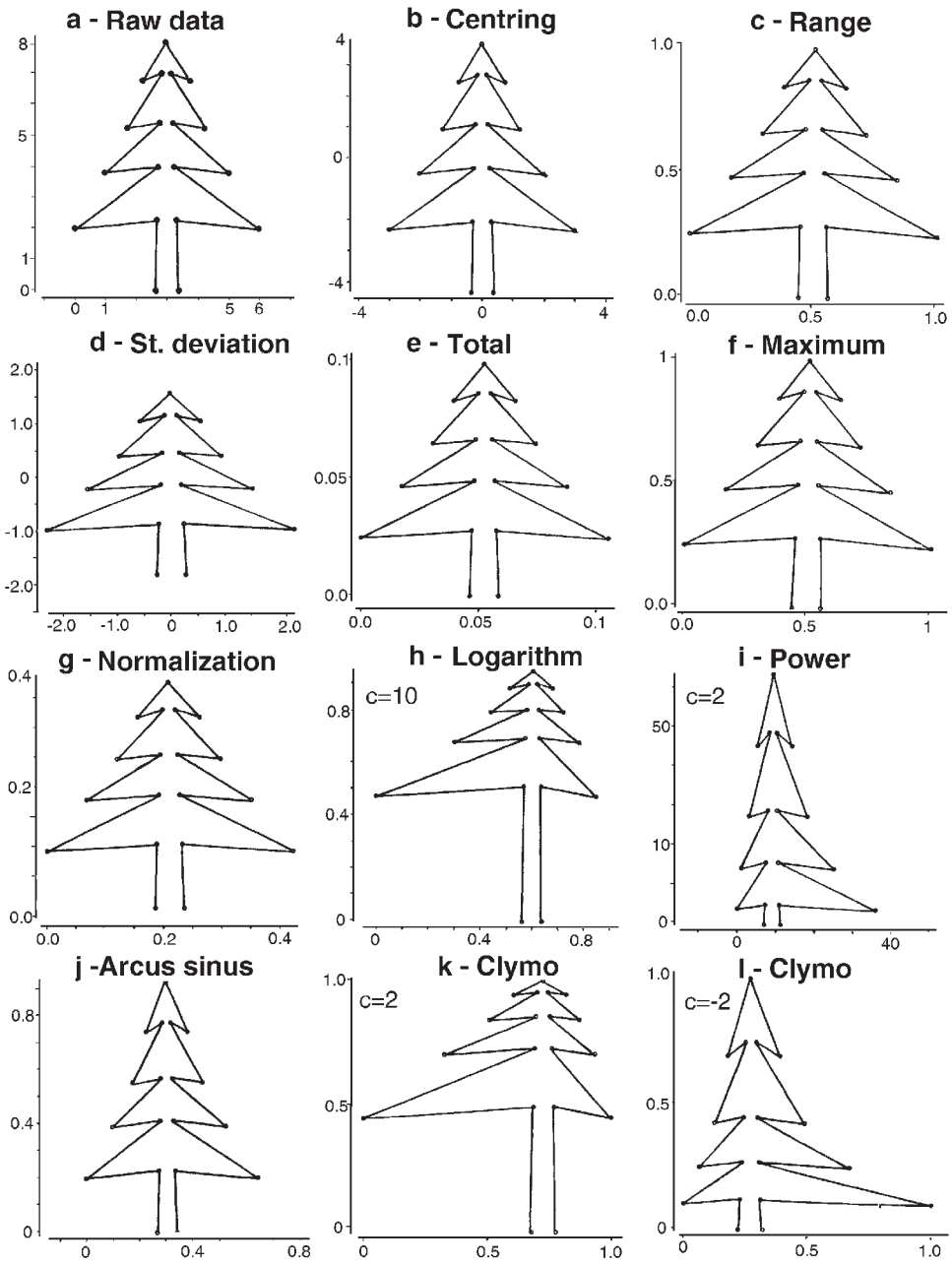


Figure 2.5. The influence of data standardization and transformation (Podani 1994). The change of shape demonstrates changes of the relative weighting of x_1 and x_2. The landmarks are indicated only on tree **a.**

Since the difference between the two tree variables is larger in standard deviation than in range, the tree is even more flattened (Fig. 2.5d).

– *Standardization with the total.* Each value is divided by the total of the respective variable:

$$x'_{ij} = x_{ij} / \sum_{j=1}^{m} x_{ij} \qquad (2.6)$$

In this way, variables having large values are diminished, and those with small values are increased in importance. Its use is logical if the total for a variable is meaningful, as in ecological samples for which the total is the sum of all individuals of species07IUC1,5,0,545,100%24i*i* in the sample. Absolute abundance differences are thus eliminated from the data.

Although this standardization is meaningless for the tree, the result is shown for completeness (Fig. 2.5e). The scores for variable x_2 became smaller than those of x_1, and the shape is very similar to the tree in Fig. 2.5c.

– *Standardization by maximum.* All values are divided by the maximum of the variable found in the sample:

$$x'_{ij} = x_{ij} / \max_j \{ x_{ij} \} . \qquad (2.7)$$

If the minimum in the sample is 0, then equations 2.3 and 2.7 provide identical results, as obvious from the comparison of Figures 2.5c and 2.5f.

– *Standardization to unit vector length (normalization*[2]*).* In the variable space, the objects are at the endpoints of vectors directed from the origin. The different contribution of variables to the length of these vectors is equalized by the formula:

$$x'_{ij} = x_{ij} / \left[\sum_{j=1}^{m} x_{ij}^2 \right]^{1/2} \qquad (2.8)$$

In other words, each value is divided by the square root of the sum of all squared values ('norm') of the variable. As a result, the sum of squares will become 1 for each variable and, in the space with objects as axes all vectors – pointing towards variable positions – will have unit length. As Figure 2.5g suggests, the equalization implied by this standardization is similar in effect to the others.

Further, rarely used standardization methods include 1 – division of all values by the range (Formula 2.3, without subtracting the minimum in the nominator), 2 – division by the square root of the sum of squared deviations of the variable, 3 – division by the square root of the total for the variable (i.e., Formula 2.6, with the denominator under the root sign), and 4 – division by the standard deviation (Formula 2.4 without subtracting the mean).

2 This should not be confused with the transformation of the data in order to approximate the normal distribution of the variable.

2.3.2 Transformation

Transformation methods do not use any data dependent statistic. Exponents or other parameters appearing in the formulae are selected completely arbitrarily by the investigator. For comparison with standardization, some methods are illustrated with the tree example.

Linear transformation. This is only a theoretical possibility for most methods of multivariate analysis. Linear transformations applied to all values (e.g., multiplication by a constant) do not change the essence of results (e.g., classifications, relative ordination positions, etc.), only the absolute numerical values change, without affecting interpretability. If multiplication is limited to a single variable, then external weighting is performed.

Nonlinear transformation. These methods, contrary to all mentioned before, destroy the original data structure, as illustrated by the distorted shape of the tree of Fig. 2.5h, for example. The 'distortion' may of course be useful for some reason, as we shall see later.

– *Logarithmic transformation.* Each value is replaced by its logarithm:

$$x'_{ij} = \log_c x_{ij} \,, \tag{2.9}$$

with c as the base of logarithm (most commonly e, for the natural logarithm, or 10). This transformation diminishes large absolute differences and is most suitable to abundance data if we wish to express magnitude discrepancy only. For the decimal logarithm, for example, the difference between abundance values 1 and 10 will become the same as the difference between 10 and 100 (Fig. 2.6a). Any ratio scale variables with a distribution strongly skewed to the right (Fig. 2.7a) can also be subjected to this transformation to make the distribution more symmetric, thus approximating normality conditions (Fig. 2.7b). This operation is also used to render linear the relationships between variables.

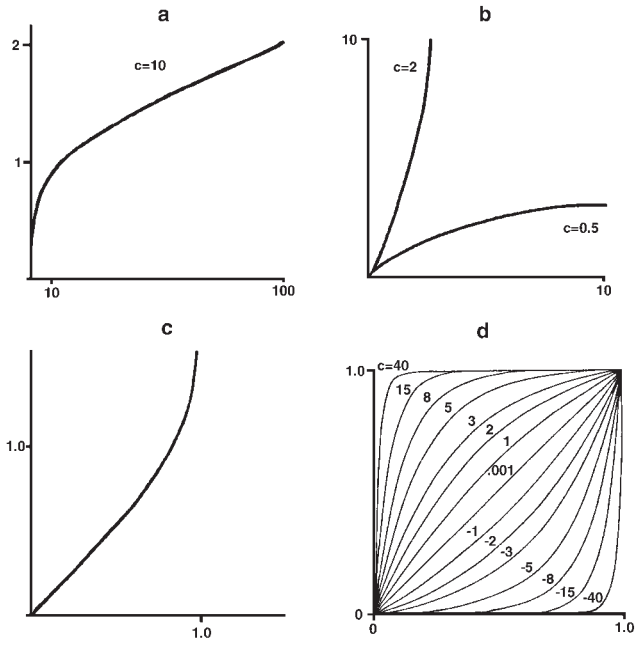


Figure 2.6. Data transformation. **a**: logarithmic transformation, **b**: power transformation, **c:** arc sin transformation, **d**: Clymo transformation. Horizontal axis: original values, vertical axis: new values.

Logarithmic transformation is an integral part of multivariate allometry, i.e., in the analysis of biological form. There are some views, however, suggesting that this transformation is not always advantageous (Reyment 1971, 1991) because the interpretation of results is more difficult.

The logarithmic function accepts positive values only, for 0 and negative scores it is undefined. Since 0 values are very common in abundance data, the above formula is replaced by:

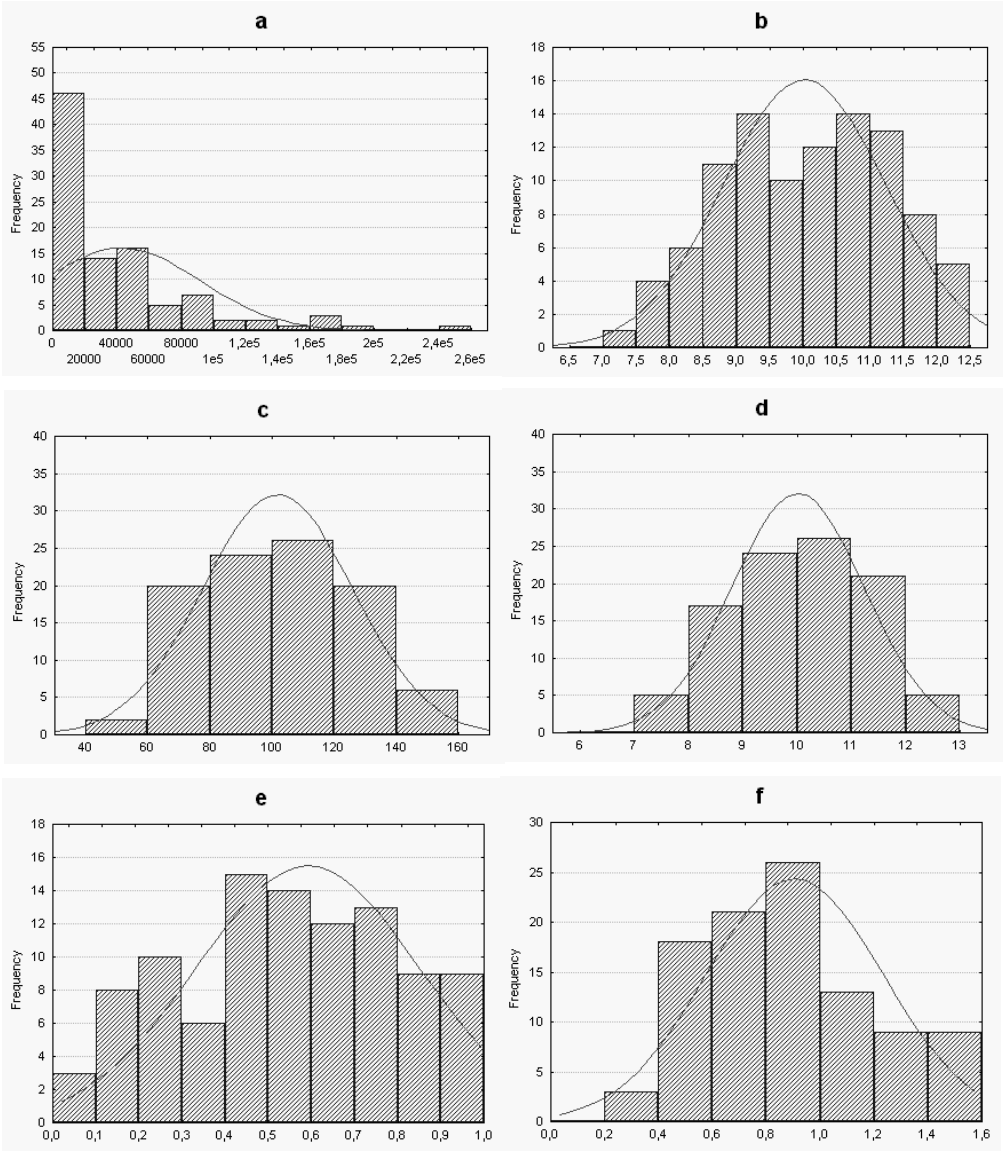


Fig. 2.7. The effect of transformation on the distribution of variables. **a-b**: logarithmic transformation from a right skewed distribution, **c-d**: square root transformation, **e-f**: arc sin - square root transformation. The continuous lines correspond to the probability density function of normal distribution fitted to the data.

$$x'_{ij} = \log_c (x_{ij}+1) . \quad (2.10)$$

Figure 2.5h illustrates well the effect of this transformation: parts of the tree coded by small values (left branches and the trunk) are overweighted, whereas the other parts are diminished.

– *Power transformation.* The original values are transformed using the power function:

$$x'_{ij} = x_{ij}^{c} \quad (2.11)$$

The result heavily depends on the selection of the c parameter (Fig. 2.6b). When $c > 1$, the large scores will become even more emphasized, but this is rarely a requirement (Fig. 2.5i). More common are transformations with $c < 1$, especially $c = 0.5$ (square root transformation), which underweight the large scores. For abundance data following the Poisson distribution (i.e., random), square root transformation will lead to close approximation to normal (Fig. 2.7c-d). In traditional statistics, this function is also used to stabilize the variance. Note that for $c = -1$ we obtain the reciprocal value, a very drastic modification of the original score.

The above methods can be considered as special cases of a general function as proposed by Box & Cox (1964):

$$x'_{ij} = (x_{ij} - 1)^{\lambda}, \text{ if } \lambda \neq 0 \ ; \quad (2.12a)$$

$$x'_{ij} = \ln x_{ij}, \text{ if } \lambda = 0. \quad (2.12b)$$

If $\lambda = 1$, then translation is performed, which has no serious consequences. For $\lambda = 0.5$, we obtain the square root transformation, whereas $\lambda = 0$ yields the logarithmic transformation. This function may be used to determine the best fit to normal distribution by changing the value of the λ parameter. The best fit is found by maximizing the following log likelihood function (Sokal & Rohlf 1981a):

$$L_i = \frac{\nu}{2} \ln s_T^2 + (\lambda - 1) \frac{\nu}{m} \sum_j \ln x_{ij} \quad (2.13)$$

where s^2_T is the variance of transformed data, ν is the degrees of freedom, and m is sample size. The λ value for which 2.13 is the highest is the optimal for the Box-Cox transformation. The procedure is a little cumbersome, and therefore its use is usually confined to univariate statistics.

Since Function 2.11 cannot be interpreted for $x_{ij} = 0$, it is modified usually for $c = 0.5$ as:

$$. x'_{ij} = \sqrt{x_{ij} + 0.5} \quad (2.14)$$

– *Arcus sinus transformation.* This function converts variables with a range of [0,1]:

$$x'_{ij} = \arcsin x_{ij} \quad (2.15)$$

and is more commonly used in combination with the square root transformation (next page). For completeness, Figures 2.5j and 2.6c show its effect on the data.

– *Clymo transformation.* It is assumed that the data reflect proportions and range from 0 to 1. (If the data are of other type, standardization by the total is performed beforehand, using Equation 2.6). The function takes then the form:

$$x'_{ij} = (1 - e^{-cx_{ij}}) / (1 - e^{-c}) \quad (2.16)$$

(van der Maarel 1979). By changing the parameter *c,* a transformation series is generated (for relative cover scores, for example) with the result shown in Figure 2.6d. For large positive values of *c,* the presence/absence data type is approximated. Values of *c* close to 0 have no practical effect, and for $c = 0$ the function is undefined. Large negative values of this parameter will overemphasize large data values and underemphasize small scores. This is illustrated by the tree example, if the scores are modified previously to fall into the range of [0,1] (Fig. 2.5k-l).

Further transformations are the *exponential function* ($x'_{ij} = e^{x_{ij}}$ and the arcus cosine function ($x'_{ij} = \arccos x_{ij}$); these rarely appear in multivariate analysis.

Binarization. Interval- and ratio-scale variables are often converted into presence/absence form to see, for example, what happens if we switch from the 'quantitative' case to the 'qualitative'. Many methods have been designed to evaluate this data type only. Now

$$x'_{ij} = 1, \text{ if } x_{ij} > p; \tag{2.17a}$$

$$x'_{ij} = 0, \text{ if } x_{ij} \leq p \tag{2.17b}$$

where p is the binarization threshold, usually 0 (all positive scores mean 'presence').

Combined transformations. Up to now, only elementary transformation functions were discussed. Often, only the combination of these yields the desired result.

– *Shape transformation*. When the data describe the outline of a shape[3] (multivariate allometry), then prior to principal components and canonical correlation analysis, Darroch & Mosimann (1985) propose a simple combination of two functions. First, the data are transformed by the logarithmic function, then centred using the new mean; so first use Equation 2.9 and then 2.2. Note, however, that centring is implied in the above analyses, so that logarithmic transformation is sufficient.

– *Arcus sinus - square root transformation for proportions.* This applies to relative frequency data only, when the scores express proportions in the range of [0,1]. The square root of all values is calculated first, then function 2.15 is used. In multivariate analysis, this is recommended for approximating the normal distribution, although its effect is less pronounced than that of the logarithmic transformation (Fig. 2.7e-f).

2.3.3 Standardization by objects

Although standardization of variables is a general practice in uni- and multivariate statistics, standardization by objects is largely restricted to ecological surveys (but see Sneath & Sokal 1973:156, for potential usage in numerical taxonomy). The aim of this operation is, for example, to equalize differences in total plant cover recorded in sampling plots. That is, a quadrat with low abundaces of many species will weight as much as another quadrat with the same number of species but higher abundances (correction for 'size').

The effect of standardization will be demonstrated by three objects, ecological quadrats, in which four species are detected. The cover of species is proportional to plant height in the illustrations in Figure 2.8. The raw data matrix is given by:

3 Methods discussed in Section 7.6 require no such transformations.

1.0 0.5 5.0

5.0 2.5 3.0

3.0 1.5 1.5

1.0 0.5 0.75

The geometric interpretation of standardization is facilitated by Figure 2.9. In this, each axis corresponds to a variable, the points represent objects. The data can be read from the figures and are not given here.

– *Centring.* The mean for the object is subtracted from all scores:

$$x'_{ij} = x_{ij} - \bar{x}_j \qquad (2.18)$$

Since negative scores also result, the effect of this operation is not shown in Figure 2.8. For two variables the effect is striking: all points move to a diagonal (Fig. 2.9a). For three dimensions, the points are projected to a plane, for more dimensions to a diagonal hyperplane.

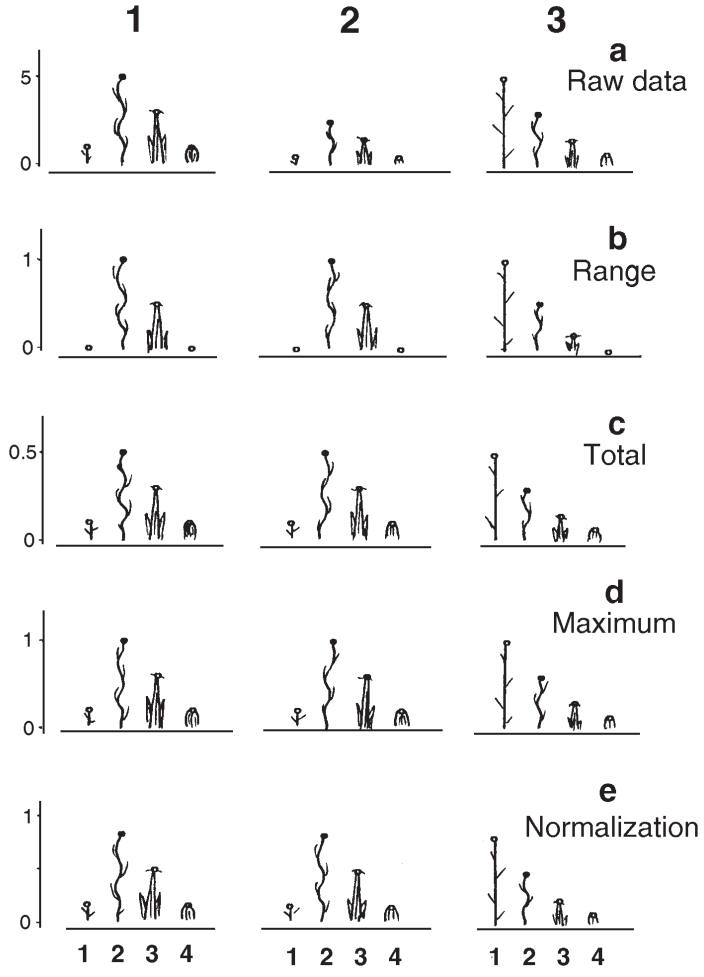


Figure 2.8. Standardization by objects. The height of plants is proportional to their cover (from Podani 1994).

Centring in fact removes one dimension, and the magnitude effect perpendicular to the line or plane is eliminated.

– *Standardization by range.* The minimum is subtracted and the result is divided by the range of the object:

$$x'_{ij} = [\, x_{ij} - \min_i \{ x_{ij} \} \,] \;/\; [\, \max_i \{ x_{ij} \} - \min_i \{ x_{ij} \} \,] \,. \qquad (2.19)$$

The new scores will fall into the range of 0 and 1 (Fig. 2.8b). Species with minimum abundance (or cover, species 1 and 4) are thus removed, which may not be our goal. For two dimensions, the new values can only be 0 or 1, so all points move into two new positions (Fig. 2.9b). For more dimensions this is not so, the points are projected to the edge of a hypercube with a side length of unity.

– *Standardization by the total.* All scores are divided by the total of the respective object:

$$x'_{ij} = x_{ij} / \sum_i x_{ij} \qquad (2.20)$$

In this way, the sum of the new values will be 1 for each object, so the standardized data will reflect proportions (Fig. 2.8c). In two dimensions, the points are projected radially to a hypotenuse of length $\sqrt{2}$ (Fig. 2.9c), in three dimensions to an equilateral triangle, in more dimensions to a 'hypertriangle'.

– *Standardization by maximum.* Each value is divided by the maximum value for the given object:

$$x'_{ij} = x_{ij} / \max_i \{ x_{ij} \} \,. \qquad (2.21)$$

This will differ from standardization by range if all variables are larger than 0 in the objects, as in the example (Fig. 2.8d). This is rarely the case; the minimum is very often 0 for abundance or cover data with many species. For two variables, the objects are projected to the periphery of a unit square (Fig. 2.9d), in more dimensions to the surface of the unit hypercube.

– *Standardization to unit vector length.* All values are divided by the square root of the sum of squares within the object:

$$x'_{ij} = \frac{x_{ij}}{\left(\sum_i x_{ij}^2 \right)^{1/2}} \qquad (2.22)$$

The effect is seen in Figure 2.8e, but this is less obvious. In the space with variables as axes, standardization 2.22 will put all the points to unit distance from the origin. In other words, the points will be projected radially onto the surface of a unit hypersphere (in two dimensions: onto a unit circle, Fig. 2.9e). Chord distance (Formula 3.54) implies this operation.

Double centring. This is the simultaneous standardization of both the objects and the variables according to:

$$x'_{ij} = x_{ij} - \bar{x}_i - \bar{x}_j - \bar{\bar{x}} \qquad (2.23)$$

where $\bar{\bar{x}}$ is the grand mean calculated for all the values of the data matrix. This operation is meaningful only if all variables are measured on the same scale. If percentage cover scores of

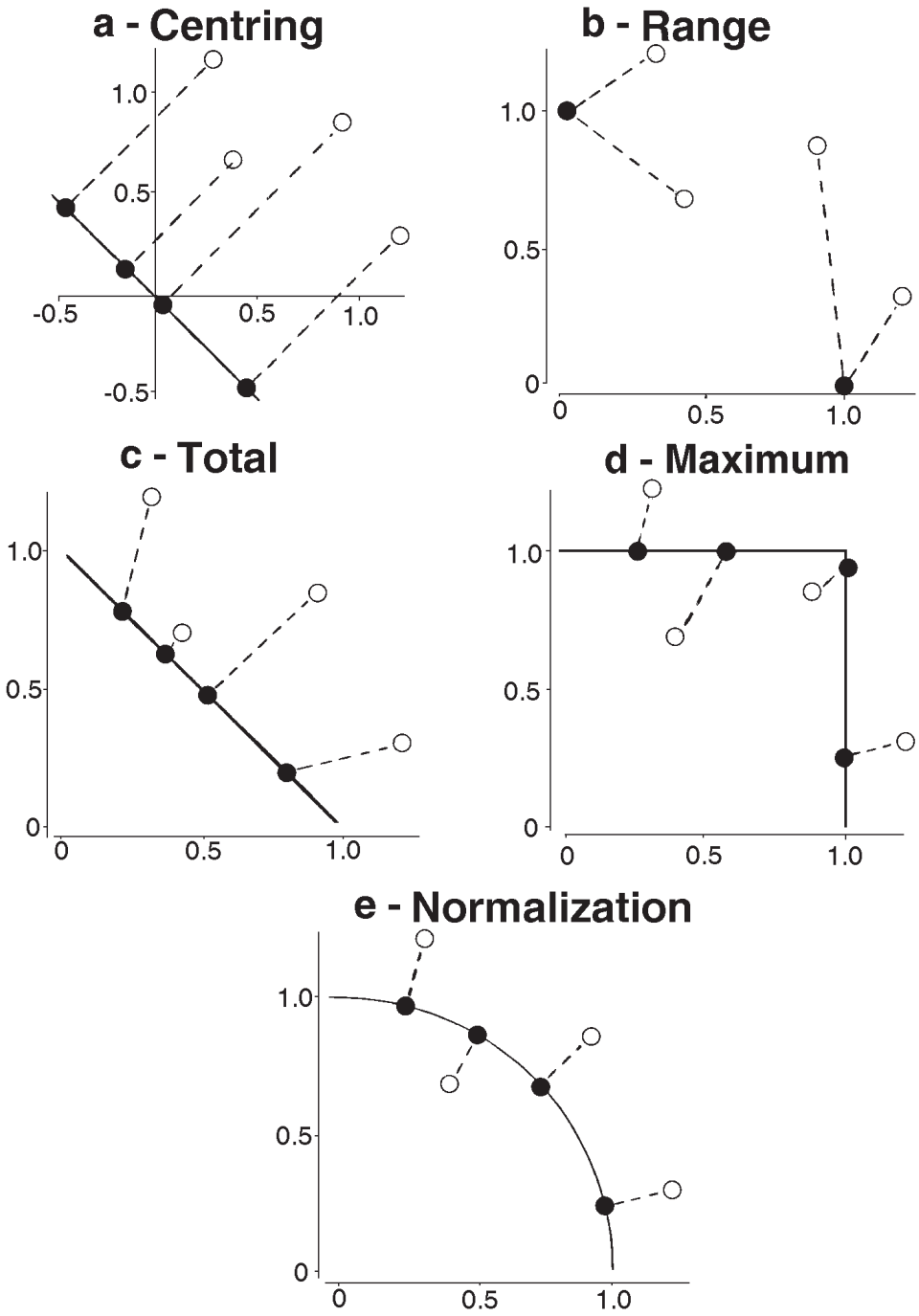


Figure 2.9. The effect of object-wise standardization for two variables. Empty symbols: original objects, full symbols: standardized objects.

species are recorded, then $\bar{\bar{x}}$ is the mean cover of all species. As a result of centring, the variables and objects are treated symmetrically (cf. attribute duality!). A rare species, if found in a species-poor quadrat, will be overweighted, whereas common species in species-rich plots will be diminished in importance. Such an implied differentation of 'individualistic' or 'unical' behaviour may be completely meaningful in ecology (Noy-Meir et al. 1975).

Double standardization by the total. Every score is divided by the respective row total and column total as well. This is implied in the χ^2-distance (Formula 3.67) , and has a central role in correspondence analysis (Section 7.3).

2.4 Literature overview

Exploration of multivariate data by various graphic means is thoroughly discussed in the volume edited by Barnett (1981), especially in Chapters 10-12 (Tukey & Tukey 1981a,b,c). Some perspective views are illustrated by examples from physics, but biological data are also used in many figures, such as the *Iris* data of Anderson (1935, 1936) and are displayed by techniques not discussed here. Barnett (1981) provides a broad overview, but the technical details are not always presented in sufficient detail. The ample bibliography may be needed to consult if one wishes to know all the algorithmic and technical details. Other important sources of information are Everitt & Nicholls (1975), Everitt (1978) and Wegmen et al. (1993).

Digby & Kempton (1987) show several examples for illustrating bi- or multivariate ecological data, some examples taken actually from Barnett. Green (1979) also has a good summary, although most figures demonstrate the diversity of graphical displays available *after*, rather than before the analysis. Reyment (1991) calls our attention to wireline diagrams, which are useful as three-dimensional perspective views, even though the examples are less convincing.

Data transformation and standardization are mentioned in most texts, at least briefly. In Gordon (1981), standardization is discussed in the context of commensurability and weighting of variables, but few details are given. Although standardization is a fundamental issue in numerical taxonomy, the topic is ignored in Dunn & Everitt (1982). For taxonomists, Sneath & Sokal (1973: 153-156) is still the best reference. Bryant (1986) discusses the utility of logarithmic transformation in systematics. Mayr & Ashlock (1991) strongly criticize and reject standardization on the grounds that less variable characters will receive too much weight in the analysis, whereas distinctive properties may decrase in importance. Stuessy (1990) takes a similar view, arguing that we should not consider all characters equally; there are many reasons behind the variability of characters, and only some are biological, others arise from measurement errors, for example. This is certainly a good point, although identification of the source of variability is not at all straightforward. The two major numerical approaches to systematics, cladistics and numerical taxonomy, have contrasting paradigmatic views on weighting, as obvious from all relevant books.

The effect of data manipulation in ecology and vegetation science has been examined by Austin & Greig-Smith (1968), Noy-Meir (1973) and Noy-Meir et al. (1975). Even though these are relatively 'old' publications, they are readable and very useful. A wide range of books on 'statistical' ecology, e.g., Digby & Kempton (1987), Jongman et al. (1987), Pielou (1984), Ludvig & Reynolds (1988) have a very limited discussion of the topic. Orlóci (1978), as Gordon (1981), discusses standardization of variables with attention to commensurability problems, whereas standardization of objects is mentioned emphasizing its relevance in certain dissimilarity and distance coefficients (see Chapter 3.).

Table 2.1. Graphical display of data structures, data standardization and transformation options in program packages (see Appendix B). + indicates methods directly applicable, * refers to methods that are available through defining a function for each variable. For preparing Kleiner-Hartigan trees, no program was found, Figure 2.2c was drawn by hand.

	Statistica	**NT-SYS**	**SYN-TAX**	**BMDP**	**NuCoSA**
Matrix of bivariate scatters	+	+			
Rotating plot			+		
Chernoff-faces	+				
Star-diagrams	+				
Histograms	+			+	+
3-dimensional perspective views	+	+			
Centring	*	+	+	*	+
Range	*	+	+	*	
Standard deviation	+	+	+	*	+
Total	*	+	+	*	+
Maximum		+	+	*	+
Normalization (unit length)		+	+	*	
Log x	*	+		*	+
Log (x+1)	*	+	+	*	+
Power (general formula)	*		+	*	+
Square root	*	+	+	*	+
Square root (x+0.5)	*	+			
Square	*	+	+	*	+
Arc sin	*	+	+	*	
Clymo			+		+
Binarization		+	+	*	+
Double centring		+	+	*	+

2.4.1 Computer program packages

Table 2.1 lists the methods introduced in this chapter and some program packages in which these methods are available. The list of programs cannot be exhaustive, of course, but the selection should be representative enough. Only multivariate data analysis packages are mentioned (see Appendix B, for information on availability), programs designed for univariate statistics are not included[4].

The implementation of data transformation methods varies with the programs. For large matrices, the use of **Statistica** and **BMDP** appears less convenient, because the variables are to be treated one by one before the analysis starts. **NT-SYS** also applies to large matrices, with an option for standardizing the variables differently, yet more easily. **SYN-TAX** and **NuCoSA** apply the same method to all variables, so that their use is simple and fast. **SYN-TAX** can be used either for run-time and permanent standardization and transformation.

4 I did my best to check the various options in these programs, but cannot bear responsibility for potential omissions and errors.

2.5 Imaginary dialogue

Q: *When I reached the end of this chapter I became a little confused over your terminology: I do not always see what you mean by sampling unit, variable, object, and whether they are interchangeable or not. Maybe it is my fault, but I would like to have a final clarification of the matter.*

A: Oh, I am sorry but I do not want to leave any 'chaos' behind me. To sum it up again: during sampling, the units are understood in the technical sense, they are *selected* from a discrete universe, or *delineated* in a continuous universe. The sampling units will be characterized by variables – understood in the statistical sense. You should not confound them yet! In data analysis, however, we usually talk about objects or observations, rather than sampling units, whereas the variables are still called the variables. From the moment that the recorded values are collected in tabular format (data matrix!) you may want to consider attribute duality. Variables and objects (i.e., rows and columns) can be interchanged for most methods, except some resemblance measures.

Q: *When I quickly inspect my data prior to analysis, I can easily find some variables that have to be transformed to achieve normality. In the same matrix, however, other variables may appear normally distributed without any modification. Is it meaningful to transform certain variables and leave the others unchanged?*

A: There is no theoretical obstacle to such operations, but you should see clearly what you are doing. As I said before, modification of data may have two objectives: reweighting or changing the distribution. Logarithmic transformation, for example, equalizes the variables and modifies the distribution simultaneously, although you may want to reach only one of these goals. The balance between weighting and transformation to normality should then be found. In multivariate studies, reweighting variables is of central importance, because this operation has general validity. Transforming the distribution is not always necessary, especially in clustering.

Application of different data transformation methods to variables of the same matrix has, in some sense, unpredictable consequences. The correlations, for example, can change significantly and, needless to say, the results of data analysis will also be affected. It is perhaps trivial that standardization of objects should always be uniformly performed for the whole data set.

Q: *If I understood correctly from Chapter 1, the objective of 'space series analysis' is to examine the effect of small systematic changes applied to the parameters of sampling in the real space. In Chapter 3 new types of space were introduced, e.g., the species space. It seems quite logical to define series in these spaces as well.*

A: You are right. Series in the real space are just the beginning. When the data matrix is completed and when you evaluate the data, you work with different *conceptual* spaces, and series can be defined in any of them. Let me remind you of the successive changes to the parameter *c* of the Clymo-function illustrated by Figure 2.6d, which is an illustration of the space series. The key parameters in the logarithm and the power transformation functions, or in the Box - Cox equation can also be changed successively.

Q: *But why? What conclusions can you draw from such experiments?*

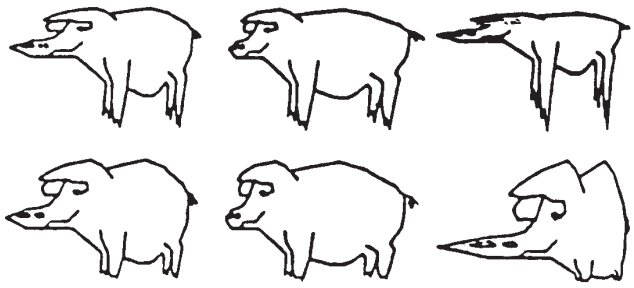
A: The series in the real space may depict the effect of our arbitrary choices of sampling parameters upon the data obtained. Analogous series applied in the data space will show the effect of arbitrary choices related to the data. For example, the logarithm on the base of 10 will more strongly reduce large abaundance values than on the base of 2. The transformation series implied by the Clymo function can be used to modify the data type in small steps. Just to repeat one point again: the Box - Cox transformation series is useful to find the best fit to the normal distribution. Such approaches require more effort, nevertheless the importance of these series has been recognized, implicitly or explicitly, by more and more people nowadays...

Q: *So there is no excuse if someone is lazy!*

A: Yes, 'unfortunately' we must live with the fact that many things in data analysis depend on our own choices: the analysis is only partly automated! Development of the sampling design, the choice of data type, and selection of transformation and standardization procedures are just the first steps that involve subjective decisions. There are many other alternatives to be mentioned in the subsequent chapters, just turn back to Figure 0.1 again to see the variety of choices you can expect! We should take the effect of our own decisions more seriously, and space series analysis may be very helpful in this regard. The closing chapter of the book will provide some more examples for you.

Q: *The modification of spruce tree shapes is very demonstrative ...*

A: I am glad if you like it, but I must admit that the idea is not completely mine. The effect of transformations was illustrated in a suggestive way on p. 15 of the journal *Münch. med. Wschr.* volume 124. number 13. Let me 'cite' some of those pigs:



These drawings are perhaps too good, too funny. In addition, you can see perhaps that the method of transformation is not the same horizontally and vertically. The original animal is in the center of bottom row. The one on top right has been transformed by the logarithm horizontally, and by square root vertically. The bottom right pig is 'squared' vertically, and log-transformed horizontally. I do not recommend such a combination of transformations, because they can have a confounding effect, as these poor animals demonstrate perfectly. In a data matrix, as I mentioned above, some of the variables may be unchanged, whereas others are transformed or standardized by the *same* method. If you feel that this restriction is incorrect, then go ahead and try to interpret results obtained for diversely transformed data...

Q: *A counter-example of clarity: I do not see why all the points moved into only two positions in Fig. 2.9b? If I had hundred points would the effect be essentially similar?*

A: Yes, of course. In two dimensions, standardization by the range of objects necessarily projects the points into the position 0,1 or 1,0. One of the values, the smaller, becomes 0 and the other will be 1 for each object. But there may be a problem, if the range is 0 (all values are equal), because division by zero does not work. We said that variables with zero range are useless, but objects that have the same value for all variables cannot be discarded! It is a little improbable for many variables, but care is always needed if you wish to try this procedure.