3

Distance, similarity, correlation...

(From data table to a new matrix)

After completing the most decisive phase of the study – sampling and subsequent data transformation – attention needs to be focused on methods that are capable of disclosing structural information hidden in the multidimensional space. The first step in this complicated methodological sequence is usually the computation of distances or the quantification of other kinds of pairwise interrelationships (dissimilarity, correlation, etc.) of the points. (Note that, as shown in Figure 0.1, this step is not always necessary, for example, in non-hierarchical clustering, Chapter 4.)

3.1 Basic terms

3.1.1 Metrics – the Euclidean distance

The first term to be clarified is the concept of *distance*. In everyday speech we have the familiar definition: the distance between two points is the length of the straight line connecting them. This is the so-called *Euclidean distance*, which later in this chapter will be extended by Formula 3.47 to many dimensions. Even though we cannot conceive of distances in the multidimensional case, this concept is the best starting point to introduce other types of interpoint relationships. For *m* points, the distances calculated between all possible pairs are written into a new matrix, the *distance matrix* of size *m*×*m*. For the three columns (sampling units) of the data matrix 2.1, we get the following result:

	unit 1	unit 2	unit 3
unit 1	0	3.0	3.0
unit 2	3.0	0	5.1
unit 3	3.0	5.1	0

or, in ‘official’ format:

$$\mathbf{D}_{3,3} = \begin{bmatrix} 0 & 3.0 & 3.0 \\ 3.0 & 0 & 5.1 \\ 3.0 & 5.1 & 0 \end{bmatrix}. \qquad (3.1)$$

Euclidean distance is just one special – although the most important – case of a family of functions, the *metric* measures, or simply: distances (in mathematics, distances and metrics are synonyms, Mirkin 1996). Many metrics can be considered for data analysis purposes. Any function d_{jk}, which satisfies the following conditions (metric axioms) for all points, is a metric:

1) If two points coincide, that is $j = k$, then it follows that $d_{jk} = 0$ (d_{jk} is zero if and only if $j = k$).

2) If the two points differ, that is $j \neq k$, then $d_{jk} > 0$.

3) According to the *symmetry axiom* $d_{jk} = d_{kj}$ (that is, the direction of measurement is immaterial).

You can easily verify that the above three axioms are satisfied for matrix 3.1. In the diagonal we have zeros, corresponding to the self-distances, whereas the off-diagonal values are positive. The entire matrix is symmetric to the main diagonal (from top left to the bottom right). Therefore, it is sufficient to provide only the three values below the diagonal ("lower semimatrix") as given below in matrix 3.2.

4) Another important metric criterion is the axiom of *triangle inequality*. It postulates that d is a metric if, for any triple i, j, k of points, the following relationship holds true: $d_{ij} + d_{ik} \geq d_{jk}$. In other word,: the distance between two points cannot be larger than the sum of their distances from a third point.

The validity of this axiom is demonstrated for visually oriented people by a two-dimensional example in Figure 3.1a. To violate the axiom for points j and k, a third point, say i, should be located such that the sum of its distances from j and k be smaller than d_{jk}. For Euclidean distance this is apparently impossible, the sum $d_{ij}+d_{ik}$ is the smallest if point i lies on the line connecting j and k. No matter where point i is moved, this sum can only increase, so the triangle inequality condition is satisfied.

One may ask the question whether it is possible to draw a configuration such that the triangle inequality holds, but we do not get Euclidean distances for the points. Figure 3.1b will illustrate the answer for four points. Let the lower semimatrix of distances be given by

$$\mathbf{D}_{4,4} = \begin{bmatrix} 0 & & & \\ 3.0 & 0 & & \\ 3.0 & 3.0 & 0 & \\ 1.6 & 1.6 & 1.6 & 0 \end{bmatrix}. \qquad (3.2)$$

Figure 3.1b shows that points 1, 2 and 3 form an isosceles triangle (equal sided triangle, each side of 3 units), and point 4 is equidistant from the others. By extrapolating what we have seen previously for three points, we realize that point 4 is the closest to the first three if it is coplanar with them (falls on the plane determined by points 1, 2 and 3), exactly in the centroid of the triangle. In this case, each distance is $\sqrt{3} = 1.73$, so the above matrix is not Euclidean. Nonetheless, the triangle inequality condition holds, since $1.6 + 1.6 > 3.0$.

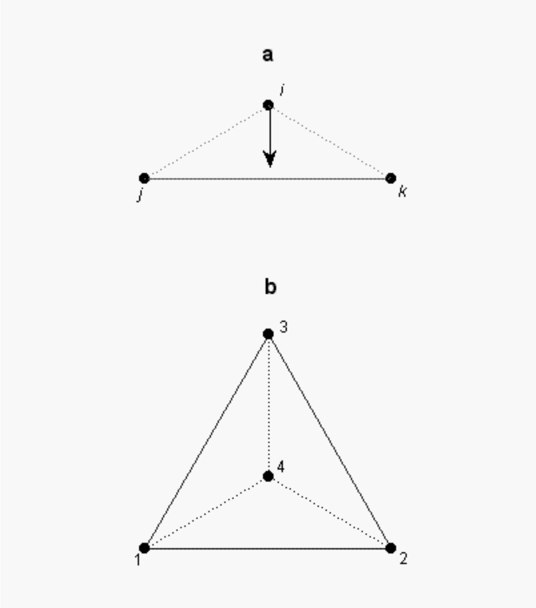

Figure 3.1. **a**: Three points cannot be drawn on the plane to violate the triangle inequality. **b**: The four points illustrate that matrix 3.2 is not Euclidean. In the Euclidean space the distance of point 4 from the others cannot be smaller than √3.

The strongest statement, therefore, on a *d* function is that it is Euclidean, because it implies the metric conditions, but the reverse – as shown above – is not always true. Matrix 3.2 was prepared completely arbitrarily, for illustrative purposes, but there do exist non-Euclidean metrics (see Table 3.4) which may provide such a result from real data. For our purposes, it is sufficient to say that any metric *d* is Euclidean, if the points can be represented in some space such that all the interpoint distances are Euclidean.

A precise matrix algebraic definition of the Euclidean property is found, for example, in Gower & Legendre (1986). A dissimilarity or distance function *d* is Euclidean if, for the matrix given by $\Delta_{m,m} = [\,-\,d_{jk}^2\,]$and for any $\mathbf{x}_m$ vector (such that $\mathbf{x}'\mathbf{1} = 0$), we have the following inequality:

$$Q(\Delta) = \mathbf{x}'\Delta\mathbf{x} \geq 0 \qquad (3.3)$$

(quadratic form, see Appendix C). For matrix 3.2, this inequality does not hold; for example with $\mathbf{x}' = [\,1\ \ 1\ \ 1\ \ -3\,]$, we obtain that $Q(\Delta) = -3.96$. If all 1.6-s were replaced by $\sqrt{3}$, then *d* would just satisfy the conditions of being Euclidean because $Q(\Delta) = 0$, and for even larger values $Q(\Delta)>0$.

One is warned that our presentation of various forms of ‘data space’ was somewhat incorrect in the previous chapter. We said that either the variables or the objects can be considered as axes in a ‘space’, and provided several scattergrams (Figures 2.1, 2.3-5 and 2.9) as illustrations. Euclidean distances were implicitly assumed to express interpoint distances, but we did not mention this. However, it is not enough to speak of points and axes only, because the definition of interpoint distances is also part of the definition of space. Consequently, a space is Euclidean if the interpoint distances are Euclidean. We talk about metric spaces if the metric axioms hold for the interpoint distances, whereas in other cases the space is non-metric.

Now, it remains to be explained why does the beginning of this chapter attribute so much importance to the Euclidean space? Does it have any advantage over the other types of space? Some arguments supporting the view that the Euclidean space is preferable include:

- It is difficult, if not impossible, to visualize – and imagine – the arrangement of points in non-Euclidean spaces, especially if the dimensionality is larger than 3. The data matrix is also conceived at the outset as a set of coordinates of points in a Euclidean space. Our results, e.g., the ordination scattergrams (Chapter 7), are also displayed in the Euclidean space, usually on the plane. Owing to these practical and mental limitations, it is advisable to require that the Euclidean, or at least the metric, conditions hold.
- The majority of multivariate methods assume that the points are in a Euclidean, or at least in a metric space. Examples are the centroid and the incremental sum of squares clustering methods. Most ordination techniques also impose metric constraints with the exception of non-metric multidimensional scaling, as its name suggests. For this reason, one must know in advance whether the selected distance or a dissimilarity function is compatible with a given data analysis technique and if so, in what form.

Despite the obvious advantages of the Euclidean property, it may very well happen that we must give up these restrictions. The biologist has a wide selection of ‘meaningful’ dissimilarity functions that reflect aspects of the data structure not detectable by other means. When the measurement scale of variables is ordinal, the subsequent data analysis stages may also remain ‘ordinal’ (i.e., non-metric). One must be careful in choosing the data analysis methods, however, and this is facilitated by tables in each chapter (e.g., Table 3.2).

In addition to the metric properties, many other criteria may govern the selection of coefficients. For example, it is extremely useful to examine the change of the resulting values in the function of certain systematic changes applied to the data. Lamont & Grant (1979), Wolda (1981) and Hajdú (1981) presented the most remarkable examples of such comparative evaluations. There is not enough space in this book to provide a full account of these criteria, but the most important aspects will be included in the forthcoming sections.

3.1.2 Dissimilarity

After defining metrics and Euclidean distances, it is time to define the term dissimilarity as well. Any function d is a *dissimilarity* if it satisfies at least the first three metric axioms (the triangle inequality need not be satisfied). Thus, the concept of dissimilarity is more general than that of metrics and the Euclidan distance; the latter are included as special cases in the big family of dissimilarity functions. The square of the Euclidean distance is a dissimilarity, for example. Its non-metric properties become evident if we take the square of all values in matrix 3.1.

Unlike Euclidean distance, most dissimilarity functions cannot exceed certain maximum values. They are constructed so that their upper bound is usually 1 (maximum dissimilarity) whereas the lower bound remains zero (that is, $0 \le d_{jk} \le 1$). Section 3.5 gives numerous examples. Many dissimilarity functions can be shown to satisfy the metric axioms, thus giving a *distance,* after the transformation $\sqrt{d_{jk}}$, which is good to know in cases where a procedural step in data analysis can only start from distance matrices. The reverse may also be true, because principal coordinates analysis – Subsection 7.4.1 – uses a matrix of squared distance values to begin with.

3.1.3 Similarity

Most biologists think in terms of similarities, rather than distances or dissimilarities, when comparing two of the study objects. To be honest, a multidimensional point swarm rarely appears in our mind: it is intuitively more straightforward to speak of similarities. Fortunately, measurement of similarity poses no problems; there are hundreds of similarity coefficients proposed in the literature. Two completely similar, that is identical, objects give the maximum similarity (usually $s_{jj} = 1$), whereas the least similar pairs reach the minimum value ($s_{jk} = 0$). That is, similarity is the complement of the dissimilarity measured in the range of [0,1], so one can be easily derived from the other:

$$s_{jk} = 1 - d_{jk} \,. \qquad (3.4)$$

Obviously, a similarity cannot be metric. Many coefficients that express similarity in the range [0,1], will become a metric, or even Euclidean, if transformed according to:

$$d_{jk} = \sqrt{(1 - s_{jk})} \qquad (3.5)$$

(see Gower & Legendre 1986).

From an **S** similarity matrix, Formula 3.5 will always yield Euclidean distance, if $0 \leq s_{jk} \leq 1$, and the matrix **S** is positive semidefinit (Appendix C).

3.1.4 Correlation and association

Disimilarities, distances and – owing to their complementary relationships – similarities have direct geometric interpretability: they express the relative positions of points in the multidimensional space. The next group of functions, on the other hand, will reveal relationships between the axes of the same space, based on the coordinates of points. Examples are coefficients of association and correlation. If the points represent a random sample from a statistical population, then the strength of association or correlation can be tested for significance by methods well-known from univariate statistics. As mentioned in Section 2.1, these measures can be calculated formally even though the axes are objects or observations, but we should remember that the principle of attribute duality has limited validity in such cases. This is especially true if we wish to apply statistical tests (not discussed in this book).

Most coefficients of association and correlation measure the strength of the relationship in the interval of [–1,1], an exception being covariance. The two extreme values reveal maximum strength differing only in direction. If necessary, then the association and correlation values are readily transformed to Euclidean distances using Equation 3.5.

3.2 Coefficients for binary data

In biology, presence/absence (binary) data are very common; and in many investigations all variables are of this type. It is not surprising therefore that the relevant literature abounds in coefficients developed specifically for binary variables, and some are almost as widely known as the familiar Euclidean distance. Their mathematical properties are quite diverse, and only the common data type justifies that they are discussed together. The formulae are presented as they appear most commonly, even though they apply only if transformed into distance form. We note in advance that the transformation $1–s_{jk}$ cannot be applied to modify either of them to satisfy the Euclidean properties, whereas the conversion $\sqrt{1-s_{jk}}$ will yield Euclidean distances in most cases (Table 3.2). In multivariate analysis, therefore, the latter operation is more admissible.

Many functions to be discussed in this section are special, simplified cases of equations presented in section 3.5. Thus, some apparently unnecessary repetitions may occur. The parallelism is not always obvious, however, so presentation of alternative forms can be helpful when comparing results based on ratio scale variables and their presence/absence versions. Each function will be abbreviated by capital letters so as to reflect its best known name (e.g., *SM, Y*1).

The formulae for presence/absence coefficients are written using the abbreviations in the so-called 2□2 contingency table:

		object 2		
		1	0	
object 1	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	n

in which

a is the number of variables present in both objects being compared (joint presence or positive match);

b is the number of variables present in object 1 and absent from object 2;

c is the number of variables present in object 2 and absent from object 1 (so that $b+c$ is the number of mismatches); and

d stands for the number of variables absent from both objects, but present in at least one object in the sample (joint absence, negative match or double zero).

The lower indices of the *a, b, c* and *d* values (such as in a_{12}) are omitted for simplicity. Clearly, $a+b+c+d = n$, that is, the number of variables describing the sample is the grand total. The marginal totals reflect the number of variables present in, and absent from, each object.

The most critical problem when selecting a presence/absence coefficient is whether the value of *d* should be considered or not, and if so, to what extent. As mentioned above, *d* is the number of variables missing from both objects, so that one might ask the proper question: why should the number of double zeros influence the comparisons at all? Although this problem was touched upon in the subsection on binary variables (1.4.2), this is the right place to present a fuller discussion. When 1 in the data represents presence and 0 means absence, that is, 1 is definitely “more” than 0 in some sense (e.g., species presence as compared to absence, or the presence of morphological characters, and so on), then one may ignore this value. It can be said that the pairwise similarity of objects is meaningful only if calculated based on characters present in at least one of the objects, and those characterizing only some others in the sample are irrelevant. Furthermore, if we use dichotomized nominal variables (see Subsection 1.4.1), then the value of *d* should most certainly be forgotten because its incorporation in the function could give an enormous weight to that variable (recall the simple example). This would contrast with the general view that *a priori* all variables should be taken with equal weight.

Under what circumstances shall we still decide that the value of *d* be considered as a factor that increases similarity? A classical example illustrating the complexity of this question is the similarity of bacterial groups on the basis of their response to standard tests (Sneath & Sokal 1973). One may feel that the number of substrates that both groups fail to decompose are just as important as the number of substrates to which both groups react positively. Consequently, the values of *a* and *d* are equally informative for us. A counter-argument is that there may be a single metabolic block responsible for many mismatches, and so it would be illogical to give too much weight to that single background character. Ecological studies also provide good examples for this controversy. In comparing the species lists of quadrats, *d* may be considered, because species absence conveys meaningful information: the given niche was taken by another competitive species or the site is not acceptable biologically to a species. The problem is that whereas species presence is most certainly an indication that a given species can survive in the area, species absence does not always mean the opposite: a species may be absent from a site just by chance (Green 1971). As Goodall (1973a) notes, the reverse may just as well be true: joint presence of common, ubiquist species may also be the result of random interactions. These two examples are sufficient to illustrate that no general recipe can be given as to the incorporation of *d* in a similarity measure. Nevertheless, if there are many rare species in the sample, whose joint presence is no more than a random coincidence, consideration of *d* would lead to an undesirably high similarity. In a sample with relatively more "balanced" presences and absences, however, *d* can be as meaningful as *a*. For those still hesitating, the intermediate coefficients (Formulae 3.19 and 3.20) can provide a good compromise.

If binary variables are in fact two-state nominal characters (i.e., the presence/absence form is just a "disguise"), then coding by 0 and 1 is arbitrary: 1 does not mean "more" than 0. In such cases, *d* is completely comparable with *a* in importance, and coefficients that treat *a* and *d* symmetrically are in order. The detailed discussion will begin with these binary indices in Subsection 3.2.1.

Our choice among presence/absence coefficients will be facilitated by graphical illustrations of their behaviour. The basis of the visual evaluation is the data matrix of Table 3.1 in which 9 objects are characterized in terms of 18 binary variables. In the direction 1□9, the objects gradually lose one character and gain another. Variables 17-18 always take zero values deliberately, so that *d* remains positive even for the pair 1/9, which has no mutual matches and therefore represents minimum similarity. In this way, we can see whether inter-object similarity reaches the zero level when *a* becomes 0. The comparison of object 1 with itself and with the other eight provides nine similarity values, whose illustration by a profile diagram shows the effect of systematic data changes upon the similarities.

3.2.1 Similarity coefficients treating a and d symmetrically

The simplest coefficients express inter-object similarity in form of ratios, which can be understood as indices of percentage agreement. Many formulae also have probabilistic interpretation.

Table 3.1. Matrix of artificial presence/absence data to be used to illustrate the performance of similarity coefficients. The objects gradually change along a background "gradient" from the starting object 1.

	Objects								
Variables	1	2	3	4	5	6	7	8	9
1	1	0	0	0	0	0	0	0	0
2	1	1	0	0	0	0	0	0	0
3	1	1	1	0	0	0	0	0	0
4	1	1	1	1	0	0	0	0	0
5	1	1	1	1	1	0	0	0	0
6	1	1	1	1	1	1	0	0	0
7	1	1	1	1	1	1	1	0	0
8	1	1	1	1	1	1	1	1	0
9	0	1	1	1	1	1	1	1	1
10	0	0	1	1	1	1	1	1	1
11	0	0	0	1	1	1	1	1	1
12	0	0	0	0	1	1	1	1	1
13	0	0	0	0	0	1	1	1	1
14	0	0	0	0	0	0	1	1	1
15	0	0	0	0	0	0	0	1	1
16	0	0	0	0	0	0	0	0	1
17	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0

The "*simple matching coefficient*" (Sokal & Michener 1958) is perhaps the most widely-known similarity index:

$$SM = \frac{a + d}{a + b + c + d} = \frac{a + d}{n} \,, \tag{3.6}$$

which is the number of agreements (matches) divided by the total number of variables. For complete agreement, $SM = 1$, whereas for maximum dissimilarity, $SM = 0$. SM is the probability that the two objects match for a randomly chosen variable. It is closely related to the *Euclidean distance*, given by

$$ED = \sqrt{b + c} \,, \tag{3.7}$$

with the notations of the contingency table. The relationship between the two is:

$$ED^2 = n\,(1 - SM)\,. \tag{3.8}$$

The range of Euclidean distance is $[0, \sqrt{n}]$. It follows from Equation 3.8 that it is usually immaterial whether 3.6 or 3.7 is used. Since the range of 3.6 is independent of n, it is favoured, because similarity values obtained for different n are comparable.

It follows immediately from Formula 3.8 that $\sqrt{1 - SM}$ is Euclidean (Table 3.2). Another advantage of SM is its linear response to the gradual change of objects, which is little-affected after conversion into distance (Fig. 3.2ab).

A variant of the simple matching coefficient (3.6) is the *Rogers - Tanimoto* (1960) index:

Table 3.2. Metric and Euclidean properties of presence/absence coefficients. Abbreviations: N = non-metric, M = metric, E = Euclidean. All similarity functions were transformed according to Equation 3.5 before examined. Exceptions are the Mountford and the Ochiai indices; they are represented here by the exponential function 3.32 and chord distance, respectively.

Function	Property	Function	Property
symmetric for a and d		*unsymmetric for a and d*	
Simple matching coefficient, *SM* (3.6)	E	Baroni-Urbani - Buser I, *BB*1 (3.20)	E
Euclidean distance, *ED* (3.7)	E	Baroni-Urbani - Buser II, *BB*2 (3.19)	E
Rogers - Tanimoto, *RT* (3.9)	E	Russell - Rao, *RR* (3.23)	N
Sokal - Sneath I, *SS*1 (3.11)	M	Faith I, *FA*1 (3.21)	N
Anderberg I, *A*1 (3.12)	E	Faith II, *FA*2 (3.22)	N
Anderberg II, *A*2 (3.13)	N	*functions ignoring d*	
correlation, *PHI* (3.14)	E	Jaccard, *JAC* (3.24)	E
Yule I, *Y*1 (3.16)	N	Sorensen, *SOR* (3.25)	N
Yule II, *Y*2 (3.17)	N	Chord distance, *CH* (3.28)	E
Hamann, *HAM* (3.18)	E	Kulczynski, *KUL* (3.29)	N
		Sokal - Sneath II, SS2 (3.30)	N
		Mountford, *MFD* (3.32)	M?

$$RT = \frac{a + d}{a + 2b + 2c + d}, \tag{3.9}$$

which gives double weight to mismatching variables, so it is always smaller than *SM*, except for the trivial case of $b+c = 0$. The denominator is interpreted by Anderberg (1973) as the number of all realized character states for *n* variables, and the numerator is the number of character states in which the objects actually agree.

Gower & Legendre (1986) have shown that for the following family of similarity functions

$$s = \frac{a + d}{a + d + \theta\,(b + c)}, \tag{3.10}$$

the transformation $\sqrt{1-s}$ will always be Euclidean if θ □ 1. When the matches are weighted doubly, as in the following index attributed to Sokal & Sneath (1963):

$$SS1 = \frac{2a + 2d}{2a + b + c + 2d}, \tag{3.11}$$

then the square root of its complement is a non-Euclidean metric.

At first glance, the following two similarity functions (Anderberg 1973) will have a probabilistic interpretation. In the first, given by

$$A1 = \left(\frac{a}{a+b} \frac{a}{a+c} \frac{d}{b+d} \frac{d}{c+d} \right)^{1/2}, \tag{3.12}$$

each term may be interpreted as a conditional probability. For example, *a/(a+b)* is the probability that a randomly selected variable will take the value of 1 for object 2, provided that its value is also 1 for object 1. Equation 3.12 is therefore the square of the geometric mean of four conditional probabilities.

The meaning of this function is further clarified by the following considerations. As shown later, expression 3.26, which contains the first two terms of formula 3.12, is the cosine of the angle of vectors pointing towards objects 1 and 2 in the multidimensional space. If the vectors coincide (0^o), then its value is 1 (maximum similarity), whereas for the largest possible angle (90^o) it becomes 0 (minimum similarity). Formula 3.26 is of course unsymmetric for *a* and *d*, therefore reverse coding (all zeros replaced by 1, and all 1-s replaced by 0) will usually yield different results. It is easy to see, however, that Formula 3.12 is the squared geometric mean of two cosine values calculated by Equation 3.26 with the two different coding systems. *A*1 is therefore most useful in situations when decision between the two coding systems is arbitrary or difficult to make.

*A*1 falls into the range of [0,1] (Figure 3.2). For maximum similarity, we have $b = c = 0$, so that all terms are 1, giving 1 as the final result. For minimum similarity, $a = d = 0$, and the expression yields zero.

Sokal & Sneath (1963:130) and Anderberg (1973) proposed a related formula, which is the arithmetic mean of the four conditional probabilities:

$$A2 = \frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right). \qquad (3.13)$$

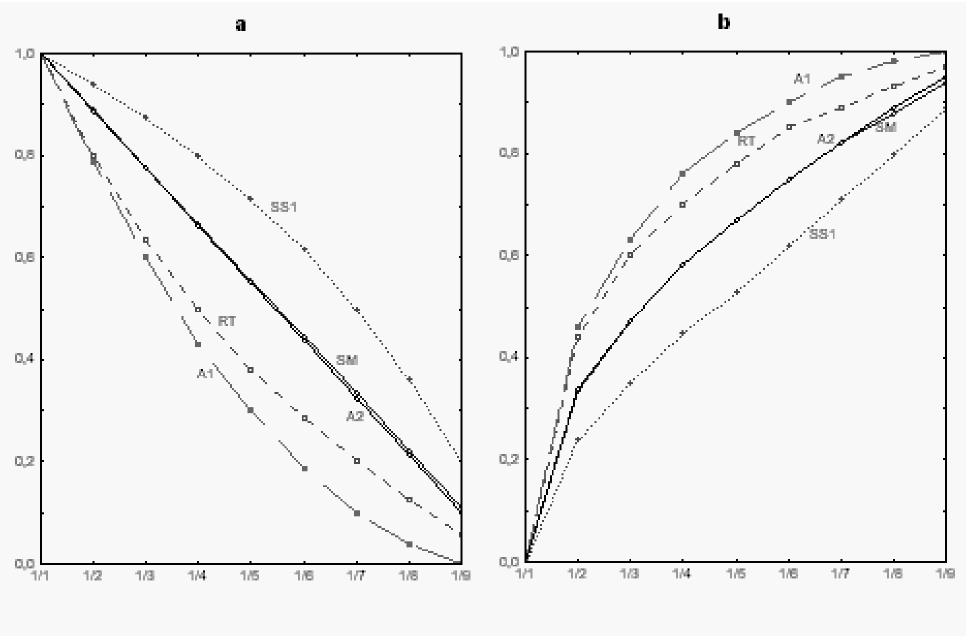

Figure 3.2. The response of similarity functions symmetric for *a* and *d* to the gradual change of data in Table 3.2 (object 1 compared with all the others). **a**: original formulae measuring similarity within the range [0,1], **b**: the formulae converted into dissimilarity according to Equation 3.5.

The result is the same as the mean of the two similarities calculated by the Kulczynski-index (3.29) based on the two coding systems. This function, like Formula 3.12, is applicable to ambiguous situations. The complements of these formulae, however, are not Euclidean.

In the binary case, the *product moment correlation coefficient* (3.70) can be expressed using the notations of the 2×2 contingency table:

$$PHI = \frac{ad - bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}} . \qquad (3.14)$$

The properties of this function are the same as those of 3.70, and will be detailed later. Now it is sufficient to point out that its range is [-1,1] (Figure 3.3a) and that the removal of *bc* from the numerator produces Equation 3.12. The *PHI* coefficient and the χ^2 statistic, measuring independence (or association) of nominal characters, are closely related:

$$PHI^2 = \chi^2 / n . \qquad (3.15)$$

Another formula for measuring association of variables is Yule's *predictability index:*

$$Y1 = \frac{\sqrt{ad} - \sqrt{bc}}{\sqrt{ad} + \sqrt{bc}} . \qquad (3.16)$$

This expresses the predictability of one variable provided that the realized value of the other is known. *Y*1 = 1 or *Y*1 = –1 indicate complete predictability: if *bc* = 0, then the two variables agree in all objects, whereas if *ad* = 0, then they always take different values (positive and negative association, respectively). *Y*1 is undefined if any marginal total of the 2×2 contingency

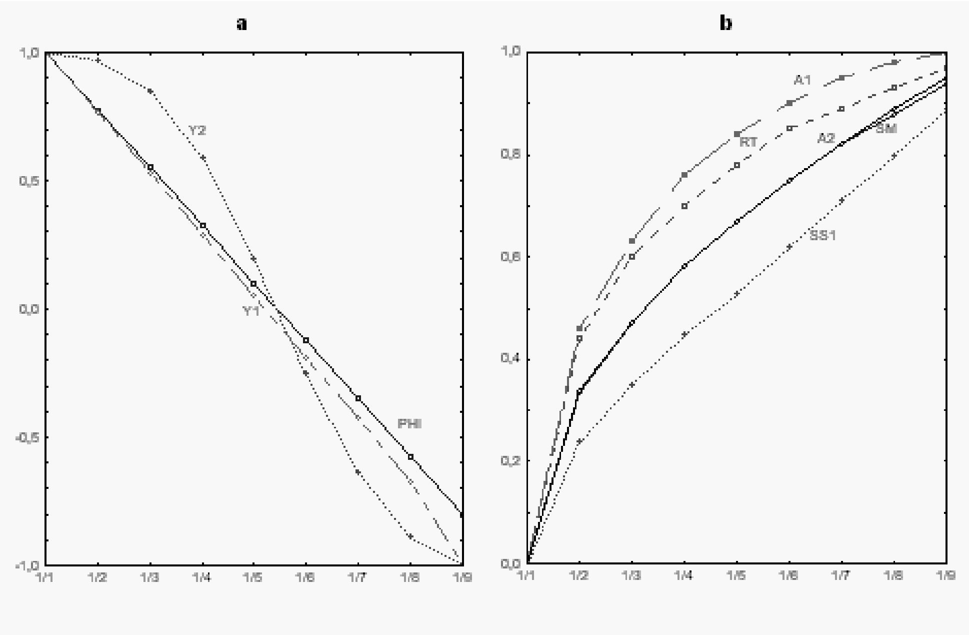

Figure 3.3. The response of similarity functions symmetric for *a* and *d* to the gradual change of data in Table 3.2 (object 1 compared with all the others). **a**: original formulae measuring similarity within the range [-1,1], **b**: the formulae converted into dissimilarity according to Equation 3.5.

table is zero (i.e., either variable is in fact a constant). The same is true of the *PHI* coefficient. Yule's second coefficient can be easily derived from the first:

$$Y2 = \frac{ad - bc}{ad + bc}\,. \tag{3.17}$$

Neither of Yule's functions can be transformed into the Euclidean distance, and the behaviour of *Y*2 seems less acceptable than the other, because it responds non-linearly to the gradual change of data (Fig. 3.3a).

There is another index measuring similarity in the range [–1,1], the *Hamann* index:

$$HAM = \frac{a + d - b - c}{a + d + b + c}\,. \tag{3.18}$$

This function adds nothing to the simple matching coefficient (3.6), only its range is extended. The two functions can be expressed from one another according to the relationship *SM* = (*HAM* + 1) / 2.

The comparison of Figures 3.2-3 reveals that the similarity functions *SM*, *PHI* and (approximately) *A*2 modifies linearly along the ordered series 1/1 → 1/9 of Table 3.1 When converted into distance form according to Formula 3.5, this property is lost, especially in the first step of the series. *SM* and *PHI* are Euclidean, so they appear the most useful for our purposes. *A*1, *RT* and *SS*1 have a nonlinear response even in similarity form, and the nonlinearity is further enhanced for the first two when given in distance form. *SS*1 remains fairly linear, but this measure in non-Euclidean. The change of *Y*2 is very different: it has an inflexion point in the middle range. Functions *Y1, Y2* and *A1* reach zero similarity if *a* or *d* is 0, and this may be undesirable.

Information theory functions, such as pooled entropy, also treat presence/absence symmetrically. They are discussed in Section 3.7, among functions applicable to more than two objects as well (heterogeneity measures).

3.2.2 Similarity measures unsymmetric for a and d

The two functions given below represent transitions between those introduced above and the measures that completely ignore the value of *d*. Baroni-Urbani & Buser (1976) took the view that the number of joint absences should not be completely neglected and, at the same time, *a* and *d* should not be weighted equally either. Hence, they modified the simple matching coefficient such that *d* is replaced by the geometric mean of *a* and *d:*

$$BB2 = \frac{\sqrt{ad} + a}{\sqrt{ad} + a + b + c}\,, \tag{3.19}$$

whereas the Hamann-index becomes

$$BB1 = \frac{\sqrt{ad} + a - b - c}{\sqrt{ad} + a + b + c}\,. \tag{3.20}$$

The two measures differ only in their range, just as the two functions from which they were derived: *BB*2 = (*BB*1+1)/2. The use of *BB*2, which has a range of [0,1], is more comfortable in multivariate analysis. Although the authors provided a detailed simulation study to demonstrate advantages of their formula, applications are scarce. In a biogeography-oriented comparative study, Kenkel & Booth (1987) found *BB*1 to be the most interpretable.

Faith (1983) has shown that the similarity calculated by *BB*2 can slightly increase when *d* increases on account of the value of *a* (for example, if $a = 10$, $d = 1$ and $b+c = 5$, then $BB2 = 0.247$, whereas for $a = 9, d = 2$ and $b = c = 5$ we obtain $BB2 = 0.259$). That is, even though we wish to give less weight to double zeros, the replacement of a double zero by a double presence produces an unexpected result. To overcome this problem, Faith proposed the following similarity coefficient:

$$FA1 = \frac{a - b - c}{a + b + c + d}\,, \tag{3.21}$$

in which *a* increases, *b* and *c* decrease the similarity, whereas *d* appears only in the denominator. Consequently, if *d* increases on account of *a*, the similarity will always decrease. Formula 3.21 measures similarity in the range of [–1,1], so the use of a formula converted according to the relationship $FA2=(FA1+1)/2$ may be easier:

$$FA2 = \frac{a + d \big/ 2}{a + b + c + d}\,. \tag{3.22}$$

The appearance of *d* in the numerator is misleading at a first glance. The formula considers *b* and *c* negatively, because they are excluded from the numerator, whereas *a* is double-weighted compared to *d*.

The *Russell - Rao index* also includes *d* in the denominator:

$$RR = a/(a+b+c+d) \tag{3.23}$$

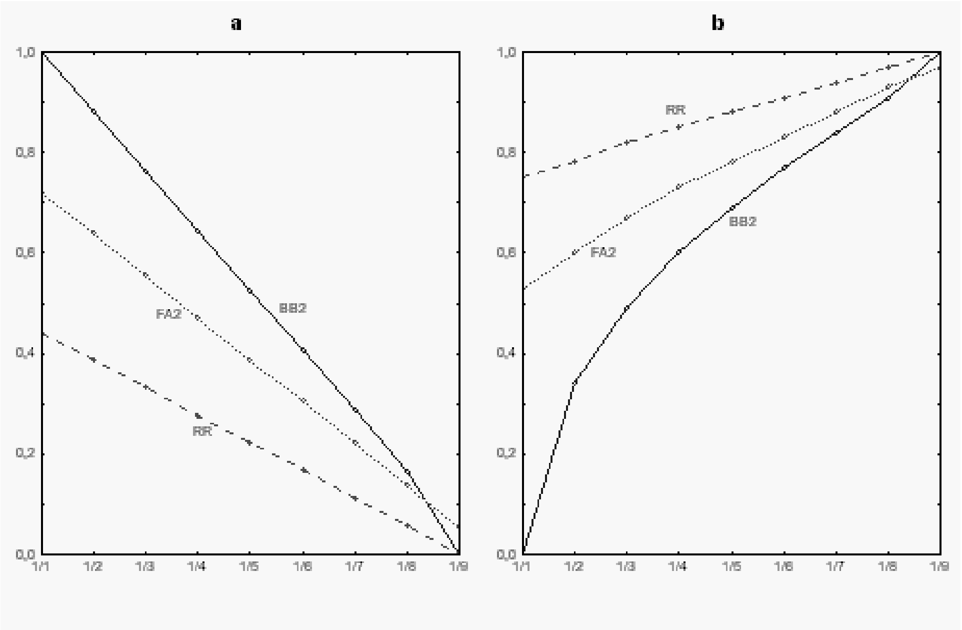

Figure 3.4. The change of similarity functions unsymmetric for *a* and *d* in an ordered series (Table 3.2). **a**: original functions, **b:** similarity converted to “distance” according to Equation 3.5.

so the increase of d will diminish the similarity value. This formula represents a relative frequency: this is the estimated probability that a randomly chosen variable is present in both objects. However, a relatively large value of d may have an overwhelming effect upon RR.

An undesirable property of $FA2$ and RR is that, even though their theoretical range is $[0,1]$, the self-similarity of objects is usually less than one (Figure 3.4a). As a consequence, their complements are not metric (contrary to Table 2 in Gower & Legendre 1986), although metric axioms 2-4 are obeyed.

A graphic evaluation of these functions is presented in Figure 3.4 ($BB1$ and $FA1$ need not be shown, because of their relationship with $BB2$ and $FA1$, respectively). $BB2$ is almost linear, the other two are perfectly linear (Fig. 3.4a) for the sample data. The full range of [0,1] is utilized only by $BB2$, whereas $FA2$ reaches neither the upper nor the lower bound (for a= 0 $FA2$ is larger than zero!). If converted to "distance", then RR remains approximately linear, but is forced into a narrow interval.

3.2.3 Coefficients ignoring the value of d

In the forthcoming formulae, d never appears, so the number of double zeros will have no effect at all upon the results. These are popular mostly in ecology. The best known of these is the *Jaccard index*:

$$JAC = a / (a+b+c)\ , \tag{3.24}$$

which is the estimated probability of the event that objects agree in a randomly chosen variable, from the set of those variables appearing in at least one of the objects being compared. It is therefore a conditional probability, with potential values ranging from 0 to 1. Conversion by Equation 3.5 produces a Euclidean measure (Table 3.2).

The *Sorensen- (Dice-) index* differs from the previous one in that the value of a is double-weighted in both the numerator and the denominator:

$$SOR = 2a / (2a+b+c)\ . \tag{3.25}$$

Double weighting emphasizes the shared part of joint presences, whereas the sum $b+c$ is responsible for the differences (compare with Formula 3.59). Owing to weighting, however, SOR cannot be brought into a Euclidean form.

The *Ochiai coefficient* (for some authors *Otsuka* is the original proponent) is given by:

$$OCH = \frac{a}{\sqrt{(a+b)\,(a+c)}} \tag{3.26}$$

for which geometric interpretation is the most straightforward: OCH is the cosine of the two vectors pointing towards the two objects in the multidimensional space (recall that this, and the other cosine obtained by reverse coding are used for calculating $A1$, Formula 3.12). For complete agreement, we get a value of 1; maximum dissimilarity is $OCH = 0$. Formula 3.26 corresponds to Equation 3.55, simplified to presence/absence data. Fager & McGowan (1963) proposed the use of a correction factor:

$$FAG = \frac{a}{\sqrt{(a+b)\,(a+c)}} - \frac{1}{2\sqrt{\max\{(a+b),\,(a+c)\}}}\ , \tag{3.27}$$

which does not influence the result seriously, and only the self-similarity gets below 0, so metric axiom 1 cannot fulfil for its complement.

The *chord distance* is closely related to the Ochiai coefficient:

$$CH = \left[2 \left(1 - \frac{a}{\sqrt{(a+b)\,(a+c)}} \right) \right]^{1/2} . \qquad (3.28)$$

This implies that the points are projected onto the surface of the unit hypersphere (standarization by Equation 2.22) and then their Euclidean distance is measured (compare with Formula 3.54).

The arithmetic mean of two conditional probabilities with respect to *a* is the so-called *Kulczynski index:*

$$KUL = \frac{1}{2} \left[\frac{a}{(a+b)} + \frac{a}{(a+c)} \right]. \qquad (3.29)$$

The second coefficient of Sokal - Sneath (1963) is given by:

$$SS2 = a \,/\, (a+2b+2c) \,. \qquad (3.30)$$

Neither of the above two formulae can be recommended, because they are not Euclidean. For the other properties, see Figure 3.5 and the related text on next page.

On the basis of the logarithmic distribution of species abundances, Mountford (1962) proposed a special similarity measure. The α parameter of this distribution is used frequently in ecology as a diversity measure (see Pielou 1975:43-45). To compare two sites on the basis of

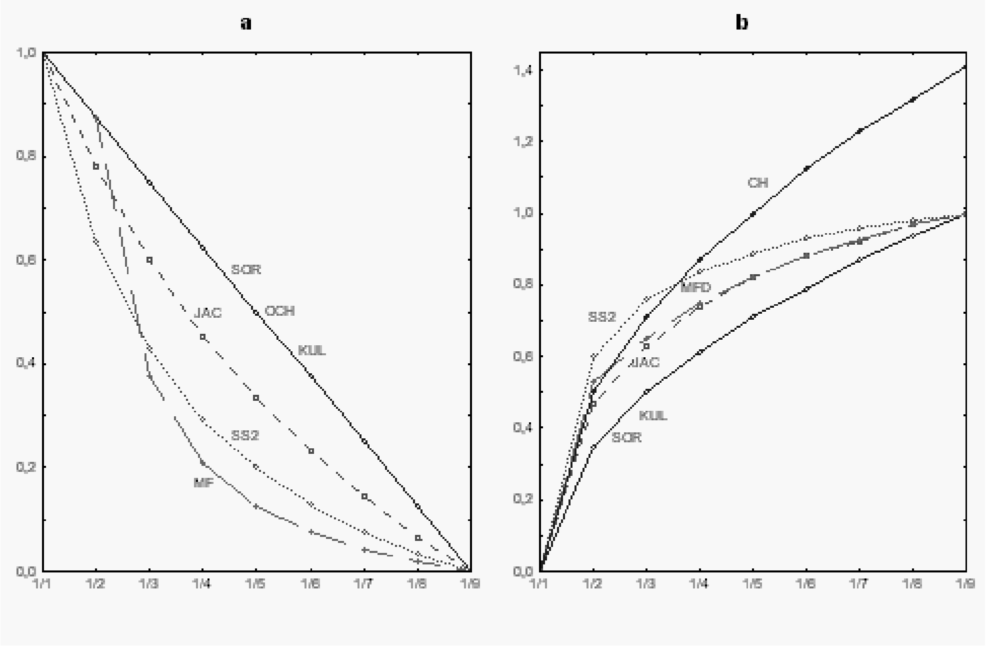

Figure 3.5. Response of similarity functions disregarding the value of *d* to data changes in the ordered series of Table 3.2. **a:** original functions, **b:** formulae after transformation using Equation 3.5, except *MFD* and *CH*.

their species composition he suggested its reciprocal value, $MF = 1/\alpha$, which is relatively independent of sample size and, in turn, of the presence of rare species. One may estimate *MF* based on the 2×2 contingency table according to:

$$MF \square \frac{2a}{ab + ac + 2bc} \,. \qquad (3.31)$$

A disadvantage ot this is that for complete agreement the denominator becomes zero. For maximum disagreement *MF* is 0. Orlóci (1978) suggested conversion of *MF* into a relativized distance measure which, as confirmed by simulation experiments, seems to be metric:

$$MFD = e^{-MF} \,. \qquad (3.32)$$

If $b = c = 0$, then *MF* must be replaced by a sufficiently large number to set $MFD = 0$. When species composition gradually changes, the similarity given by Equation 3.31 first decreases rapidly, and then levels off, which is clearly undesirable in multivariate analysis (Kenkel & Booth 1987, Wolda 1981; see also Figure 3.5b).

The graphical comparison of similarity functions discussed in Subsection 3.2.3 is Figure 3.5. *SOR, OCH* and *KUL* are identical – and linear – for the sample series. The other three functions, *JAC, SS*2, and *MF* deviate from linearity in that order. Out of the distance versions, *CH* appears to reflect best the gradual background changes and *CH* is also Euclidean. The diagrams suggest that the Mountford index is the least acceptable, since its changes are not proportional to the changes in the data series.

Coefficients not satisfying the symmetry axiom are little if ever used in multivariate studies. For the sake of completeness, we mention Kulczynski's second index ($s = a/(b+c)$), the Simpson index ($s = a/(a+b)$) and the Braun-Blanquet index ($s = a/(a+c)$). These equations apply in cases when the direction of comparison is important, as in procedures developed for asymmetric matrices (Gower 1977).

3.3 Coefficients for nominal data

When all the variables are of the nominal type with more than two states (="*multistate nominal*"), then the comparison of objects is the easiest with the generalized forms of presence/absence similarity functions that consider *a* and *d* symmetrically. If *u* is the number of variables for which the two objects agree, then the simple matching coefficient becomes

$$SM = u/n \,. \qquad (3.33)$$

The generalized form of the Rogers - Tanimoto index is:

$$RT = u \,/\, (2n\text{-}u) \,, \qquad (3.34)$$

whereas Sokal - Sneath's first coefficient (3.11) is rewritten as:

$$SS1 = 2u \,/\, (n+u) \,. \qquad (3.35)$$

Gower's general similarity coefficient (3.103) compares the objects for nominal variables according to Equation 3.33.

In order to expand the scope of the *PHI* coefficient, a new contingency table needs to be prepared. Since this function is usually calculated to measure association between variables (in the statistical sense), the table is shown with symbols referring to variables:

		variable 2			
		1	j	q	
	1	f_{11}			$f_{1.}$
variable 1	i		f_{ij}		$f_{i.}$
	p			f_{pq}	$f_{p.}$
		$f_{.1}$	$f_{.j}$	$f_{.q}$	$f_{..}$

In this table, f_{ij} is the frequency of the event that state i of variable 1 co-occurs with state j of variable 2 in the sample. $f_{i.}$ and $f_{.j}$ are marginal frequencies, whereas $f_{..} = m$, the sample size. p and q are the numbers of states for variable 1 and 2, respectively. The association of the two variables can be measured by the chi-square statistic:

$$\chi^2 = \sum_{i=1}^{p} \sum_{j=1}^{q} \frac{\left(f_{ij} - \dfrac{f_{i.}\, f_{.j}}{f_{..}} \right)^2}{\dfrac{f_{i.}\, f_{.j}}{f_{..}}} \,. \tag{3.36}$$

χ^2 increases if $f_{..}$ increases so that some standarization is in order. A solution implied by Equation 3.15 (applied to the 2×2 contingency table) comes to our mind first, i.e., division by $f_{..}$. The maximum of $\chi^2/f_{..}$ is $\min\{(p-1),(q-1)\}$ (see Anderberg 1973:76), which needs to be considered as well to obtain the final formula

$$CR = \left(\frac{\chi^2 \,/\, f_{..}}{\min\,[\,(p-1)\,,(q-1)\,]} \right)^{0,5} , \tag{3.37}$$

known as *Cramér-index* (Cramér 1946) in the literature. Its value ranges from 0 to 1 for any values of p and q. However, standard use of CR as a similarity function in multivariate analysis is questioned by many authors; notwithstanding the fixed range, it is unlikely that $CR = 0.5$ indicates the same association strength for a 5×5 and for a 3×6 table, for example (analogous problems will be discussed in Chapter 9). CR can only be recommended if the number of states is the same for all variables, and this is rarely satisfied. A general solution is offered by the coefficient of Goodman & Kruskal (1954), which measures the predictability of variable 2 provided that the score for variable 1 is known. To follow the derivation of this formula, let us assume that first we wish to guess the value of variable 2 for a given object without knowing the value of variable 1. Clearly, the best guess is the most common state of variable 2, because it minimizes the probability of wrong guesses. That is, we look for $\max_j\,[f_{.j}]$ in the table. However, if we know already that variable 1 realizes state i for that object, then the ith row of the table becomes of primary concern, and $\max_j\,[f_{ij}]$ is to be found to minimize the chance of bad guesses. The decrease in the probability of erroneous guesses is proportional to the difference between these maxima. That is, given that variable 1 is known, the relative predictability of variable 2 will be obtained by the formula:

$$LAS = \frac{\sum_{i=1}^{p} \max_j\,[\,f_{ij}] - \max_j\,[\,f_{.j}]}{f_{..} - \max_j\,[\,f_{.j}]} \,. \tag{3.38}$$

LAS takes the value of 0 if variable 1 is completely uninformative with respect to variable 2 (independence), whereas *LAS* = 1 when knowledge of the realized state of variable 1 excludes all but one state for variable 2 (in the latter case, each row and column of the contingency table has only one nonzero value). *LAS* is an unsymmetric measure: knowledge of variable 1 does not increase predictability of variable 2 by the same amount as vice versa. The symmetry axiom satisfies, however, if the two predictability values are averaged:

$$\Lambda = \frac{\sum_{i=1}^{p} \max_j [f_{ij}] + \sum_{j=1}^{q} \max_i [f_{ij}] - \max_i [f_{i.}] - \max_j [f_{.j}]}{2f_{..} - \max_i [f_{i.}] - \max_j [f_{.j}]} \tag{3.39}$$

(Goodman - Kruskal lambda). In the presence/absence case, the formula simplifies to $Y1$ (Equation 3.16). Since $Y1$ is not metric, then the same must be true of Λ. In this regard, therefore, Λ appears less practical than *CR,* satisfying the conditions for being Euclidean.

3.3.1 Comparison of sequences

Biology abounds in sequential information: it is needless to emphasize that DNA, RNA and protein sequences are of central importance in biochemistry and in contemporary molecular taxonomy. As a matter of fact, the information in such sequences is of the nominal type even if it is evident that the order of the basic units has enormous consequences as to the biological functionality of these molecules. Comparisons usually involve checking the positional agreements in analogous locations of alternative sequences. The starting point of the evaluation is the sequence of units itself, rather than a data matrix of some sort. Nevertheless, the discussion of distance functions would be incomplete without mentioning at least some simple methods for comparing biological sequences.

The most critical step of the comparison is to find the best (“optimal”) alignment of the two sequences. A classical procedure is the optimalization algorithm proposed by Needleman & Wunsch (1970), which relies on the following considerations:

- the number of *positional agreements*, M, is the maximum;
- the number of *positional disagreements*, U (when in the given position the two sequences have different units, = *Hamming distance*) is minimized;
- the two sequences may not have the same length, so let the difference be denoted by G. In order to find the best fit, one or both of the chains must be broken if G>0, so that certain units will be unpaired. In the calculation of similarity, these breaks (*“indel”*) are weighted by “penalty points”, denoted by w. The variants of the method differ in the specification of this – admittedly arbitrary – weighting function. If we simply disregard the breaks, then $w = 0$. For some authors, indels are the same as positional disagreements, so $w = 1$. Swofford & Olsen (1990) suggest that large breaks be excluded from the comparison because they will considerably distort the result if alingned parts have a good fit. For short breaks, $w = 0.5$ is a good compromise.
- the *effective chain length* is given by

$$L = M + U + wG\,, \tag{3.40}$$

from which we derive the similarity analogous to the simple matching coefficient (3.6):

$$S = M / L \tag{3.41}$$

Finding the best alignment, the maximalization of *S,* needs to be assisted by a computer, although shorter sequences may be compared by hand calculations. Further details of the method are presented in Kruskal (1983), Weir (1990) or Waterman et al. (1991). The *S* coefficient has universal validity, being applicable to both DNA, RNA and protein sequences. The value of *S* can directly be submitted to further analysis, but its complement, 1–*S* is recommended instead.

For the nucleic acid sequences CTGTATC and CTATAATCCC, the algorithm finds several equally optimal solutions, each with $M = 6$, $U = 1$ and $G = 3$. A possible optimal alignment is:

```
CTGTA T C
CTATAATCCC
```

and, with $w = 1$, the similarity of the sequences is $S = 6/(6+1+1\times3) = 0.6$.

The temporal change of nucleic acid sequences is characterized by decreasing *S* values, if the assumption that any nucleotide may be substituted by any other nucleotide with the same probability is valid. If μ is the rate of mutation and *t* is the time elapsed, then the quantity

$$2\mu t \;\square\; K = \frac{3}{4}\ln\left(\frac{3}{4S-1}\right) \tag{3.42}$$

can be used to estimate the *evolutionary distance* (Jukes & Cantor 1969). In this, the increase of *K* over time is approximately linear, but not without upper bound; when *S* reaches 0.25, we have the expected similarity of two randomly generated sequences and *K* becomes irrelevant. A disadvantage of this function is that the possibility of several mutations on the same location is disregarded. The fact that the changes A↔G and T↔C (*transitions,* see sections 6.3-4) are much more common than the others is also neglected (the Kimura distance does consider this fundamental observation, see Waterman et al. 1991). For proteins, 3 is replaced by 19 and 4 is replaced by 20 in the above equation, if we allow the simplification that all amino acids are equally frequent. Since the ratio 19/20 is close to 1, Function 3.42 simplifies to $K = -\ln S$.

The above discussion just touched upon the complex topic of sequence comparisons for molecular data, for a recent account, see Gusfield (1997). Related formuale that allow within-population variability are presented by Weir (1990). Other types of sequences also exist in biology, such quadrats in a line transect in which presence/absence of species is recorded, linear "plots" along which the order of species touched is recorded, or chains of events observed on a temporal scale. In such cases, however, ordinal information is involved, rather than positional data (se see the next section).

3.4 The case of ordinal variables

For the comparison of variables of the ordinal type, we have several well-known and widely tested rank statistics that may also be considered in multivariate analysis. Regardless of the coding of our data, the original values need to be converted into ranks in order to be able to use these measures. The smallest value of a variable gets rank 1, the next one gets 2, and so on.

Thus, the original score x_{ij} is replaced by the rank r_{ij} showing the position of x_{ij} in the rank order of all values for variable *i*. The agreement of two variables, now two rank orders, is expressed most easily by the *Spearman rank correlation:*

$$RHO_{hi} \; = 1 - \frac{6\sum_{j=1}^{m} (r_{hj} - r_{ij})^2}{m(m^2 - 1)} \, , \tag{3.43}$$

which yields 1 for completely identical orderings, and –1 for opposite rankings. *RHO* will be around zero if there is no relationship between the rank orders. The usefulness of rank correlation is strongly limited by the presence of equal values in the original data, which produce *tied ranks*. A few ties can be treated by correction formulae, but too many will diminish the sensitivity of *RHO*, and therefore *TAU* is recommended (see below). Rank correlation is best suited to problems in which the observations directly provide orderings (e.g., arrivals of species at a trap, etc.). The derivation of the formula is found in Legendre & Legendre (1983:206-207).

Spearman rank correlation overemphasizes large differences in ranks, and small changes are slightly manifested in the resulting coefficient. It may be an advantage if small differences between the original data are largely due to unreliable sampling or measurement. However, if we wish to consider all differences equally, because small differences are as meaningful as the large ones, then the *Kendall* τ is preferred:

$$TAU_{hi} \; = \frac{4\sum_{j=1}^{m} C_j \; - \; m(m-1)}{m(m-1)} \, . \tag{3.44}$$

C_j is determined as follows: the values for the first variable are listed in ascending order, and the corresponding ranks for the second variable are written besides these values. For each rank score of variable 2 we count the number of larger ranks that appear afterwards in the list. The total of these values for complete agreement of the two orders is $m(m–1)/2$, so this sum in the denominator of (3.44) is multiplied by 4 to have *TAU* = 1 in this case. For opposite orderings, the total of C_j values becomes zero, and the function will yield –1. A more complex formula is used if there are tied ranks in the sequences (see the alternative method of calculation on next page).

The calculation of *TAU* is illustrated by a simple example. Let the data table for the two variables and six objects be given by:

variable 1 12 16 18 14 17 20
variable 2 15 18 19 13 12 17

The following lists are prepared first:

Values of var. 1. ordered	Corresponding values of variable 2	Ranks for variable 2	Number of rank scores larger than the rank in the previous column
12	15	3	3 (5, 6, 4)
14	13	2	3 (5, 6, 4)
16	18	5	1 (6)

17	12	1	2 (6, 4)
18	19	6	0
20	17	4	0
			Total: 9

then *TAU* is calculated using formula 3.44 to yield *TAU*= (4×9 – 6×5) / (6×5) = 0.2.

The pairwise comparison of *objects* based on ordinal variables is a more difficult and unfortunately neglected topic. In many studies, ordinal data are evaluated by procedures applicable to interval and ratio scale variables only. There is no need to emphasize the incorrectness inherent in such unwise generalizations of the data type: *for ordinal variables differences between states and their ratios are not interpreted*. One must admit, however, that the sequential information of the possible states of ordinal variables is difficult to incorporate into a dissimilarity coefficient. We do not advise against the application of coefficients developed for nominal variables, but we lose information in that way: this is the consequence of converting the ordinal scale “down” to the nominal. If functions to be discussed in Section 3.5 are used, then we shift to the interval scale implicitly, and attribute meaning to undefined differences. As a solution, the above-mentioned rank statistics may be applied to objects formally, specially the complement of *TAU* (Diday & Simon 1976). Equation 3.44 applies to objects as well, and only *m* is to be replaced by *n*. The use of this coefficient is further illustrated by a different computational procedure, based on the raw scores. Let *j* and *k* be the two objects to be compared, and let us define a Δ^j_{hi} auxiliary variable as follows:

$$\Delta^j_{hi} = \begin{cases} 1 \text{ if } & x_{hj} > x_{ij} \\ -1 \text{ if } & x_{hj} < x_{ij}\,, \\ 0 \text{ if } & x_{hj} = x_{ij} \end{cases}$$

If T_j denotes the number of variable pairs for which $\Delta^j_{hi} = 0$ for object *j,* and T_k is an analogous quantity for object *k*, then we can write that:

$$DTAU_{jk} = 1 - \frac{2}{\sqrt{[n(n-1)-T_j]\,[n(n-1)-T_k]}} \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} \Delta^j_{hi}\,\Delta^k_{hi} = 1 - TAU_{jk}\,. \qquad (3.45)$$

The function is undefined for the presumably rare situation in which either or both variables are constant for all objects, because $T = n(n–1)$ and the denominator becomes zero.

Formula (3.45) treats tied values with ease, which is important because ordinal data can be very “poor” in values. Consider, for example, a vegetation data table in which quadrats are characterized by AD values of the Braun-Blanquet style (recall Table 1.1). A species can take six different values plus zero, but mostly the first few states of the ordinal scale are taken (in a quadrat we cannot have many very abundant species!). The species composition in a quadrat is represented by a species list ordered according to the AD scores. Owing to the limited number of states it is a necessity that this order has many tied values; the AD values of species in a quadrat represent the so-called *partially ranked data type*. Critchlow (1985) lists several coefficients for such data. Their adaptation to biological problems was first attempted by Dale (1989): he proposed using a special case of Levenshtein distance (Ulam distance) in multivariate analysis of vegetation data. This measure can be interpreted as the number of replacements to be made in the (partial) species ordering of one quadrat to obtain the (partial) ordering of the other quadrat.

The auxiliary variables serve as a basis to define an alternative formula, developed by Goodman & Kruskal (1954) to measure association of ordinal variables. With the same denotations as above, for the comparison of objects the Goodman-Kruskal γ will take the following form:

$$\gamma_{jk} = \frac{\sum_{h=1}^{n-1} \sum_{i=h+1}^{n} \Delta_{hi}^{j} \ \Delta_{hi}^{k}}{\sum_{h=1}^{n-1} \sum_{i=h+1}^{n} |\Delta_{hi}^{j}||\Delta_{hi}^{k}|} . \tag{3.46}$$

This is a simple ratio. The denominator is the number of variable pairs that are ordered (i.e., not tied) in both *j* and *k*. Based on the relative proportion of products 1×1and 1×–1 the numerator decides whether this ordering is in the same or in the opposite direction for the two objects, and to what extent. For full positive agreement we have $\gamma_{jk} = 1$, for completely reversed orders we obtain $\gamma_{jk} = -1$. A zero value means that the positively ordered pairs are fully compensated by the negatively ordered pairs. The complement of 3.46 can be used as a dissimilarity measure.

This coefficient is also burdened by the presence of many ties, and therefore the result often depends on a small, potentially negligible, subset of variable pairs. It is especially critical in vegetation data with AD scores of which many are zeros. As a possible solution, Podani (1997a) proposed a modification of Goodman-Kruskal's formula, which behaves as a pre-sence/absence-based dissimilarity index for tied pairs of variables. This "hybrid" *coefficient of discordance* thus gives priority to ordinal data, but when the ordinal information is insufficient, presence/ absence relations become decisive in determining inter-object dissimilarity.

3.5 Coefficients for interval and ratio scale variables

As far as most formulae are concerned, there is no difference between the treatment of interval and ratio scale (often called "quantitative") variables, so these two data types are discussed together. Exceptions are coefficients that are sensitive to the translation of objects in the variable space (i.e., to the addition of a constant to all values), such as chord distance, angular separation, cross products and covariance. Therefore, these coefficients should never be applied to interval scale variables, i.e., those with arbitrary zero point! The behaviour of most functions will be illustrated using the fabricated data matrix of Table 3.3, in the same way as the presence/absence coefficients. These data describe the gradual change of 9 objects along an imaginary "gradient" such that each variable is characterized by a simplified unimodal response curve (Figure 7.9a). This is sufficient to get some preliminary insight into the problem of evaluating coefficients. A more detailed survey with different ordered data series, but for fewer coefficients, is presented in Hajdu (1981). Following this example, the reader can prepare her or his own artificial ordered data series with any systematic change, examining and comparing the available coefficents to facilite choice among them.

The most appropriate starting point of this discussion is the *Euclidean distance*:

Table 3.3. Matrix of artificial data for evaluating coefficients developed for interval and ratio scale variables. The objects gradually diverge from the starting object along a “gradient”, such that their response to the gradient is unimodal.

	Objects								
Variables	1	2	3	4	5	6	7	8	9
1	1	0	0	0	0	0	0	0	0
2	2	1	0	0	0	0	0	0	0
3	3	2	1	0	0	0	0	0	0
4	4	3	2	1	0	0	0	0	0
5	4	4	3	2	1	0	0	0	0
6	3	4	4	3	2	1	0	0	0
7	2	3	4	4	3	2	1	0	0
8	1	2	3	4	4	3	2	1	0
9	0	1	2	3	4	4	3	2	1
10	0	0	1	2	3	4	4	3	2
11	0	0	0	1	2	3	4	4	3
12	0	0	0	0	1	2	3	4	4
13	0	0	0	0	0	1	2	3	4
14	0	0	0	0	0	0	1	2	3
15	0	0	0	0	0	0	0	1	2
16	0	0	0	0	0	0	0	0	1

$$EU_{jk} = \left[\sum_{i=1}^{n} (x_{ij} - x_{ik})^2 \right]^{1/2}, \tag{3.47}$$

mentioned many times from the beginning of this chapter as a function that reflects our every-day, intuitive feeling about distances (Figure 3.6). It is the generalization of the well-known Pythagorean theorem to many dimensions. Euclidean distance is the standard reference for evaluating any other dissimilarity or distance measure, as emphasized earlier. Since the differences are squared before pooled, EU_{jk} is most sensitive to large differences. Its minimum is 0, and there is no upper bound.

For the data of Table 3.3, the distance from object 1 first increases rapidly, but then the change is less pronounced (Fig. 3.7a). If the series were continued, then we would remain on the level of the distance between objects 1 and 9, which has enormous consequences in ordinations (see Chapter 7, for more details).

The *Manhattan-metric* is the sum of absolute differences:

$$CB_{jk} = \sum_{i=1}^{n} |x_{ij} - x_{ik}|, \tag{3.48}$$

which is also called the “*city block*” metric. Both names remind us of the north American cities with perfectly regular street systems in which we cannot always get from one point to another along a straight line: one has to walk round the blocks (Fig. 3.6). As its name suggests, function 3.48 is a metric, although not Euclidean (Table 3.4).

The Euclidean distance and Manhattan metric are special cases of a general family of functions, the so-called Minkowski-metrics:

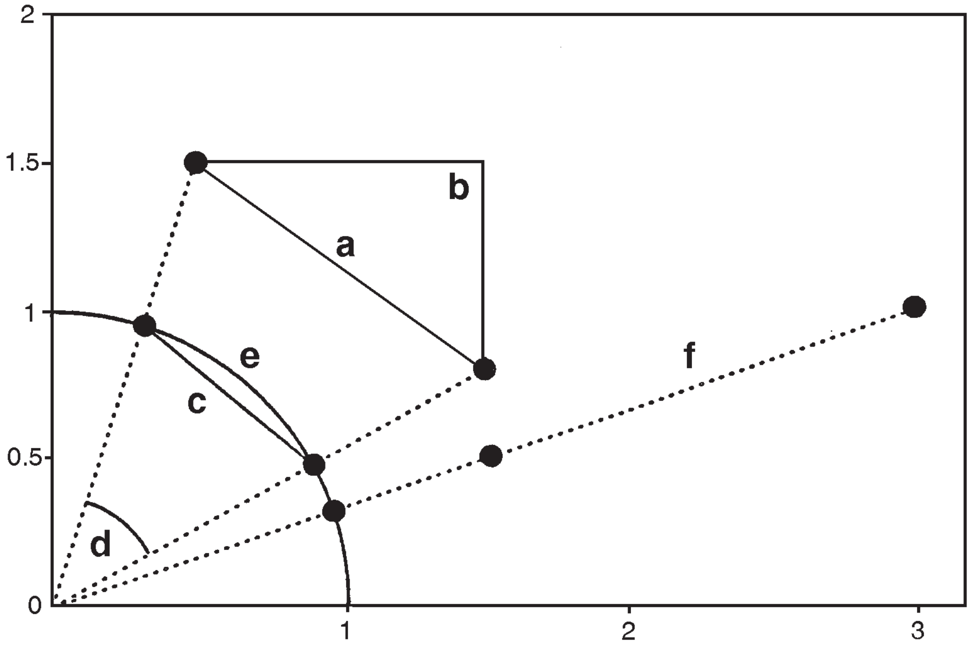

Figure 3.6. Different metrics in the two-dimensional variable space. **a**: Euclidean distance, **b**: Manhattan-metric, **c:** chord distance, **d:** angular separation, **e:** geodesic metric, **f:** chord distance becomes zero, if the proportion of variables is identical in the two objects.

$$MNK_{jk}^{(r)} = \left[\sum_{i=1}^{n} | x_{ij} - x_{ik} |^{r} \right]^{1/r} , \qquad (3.49)$$

in which $1 \geq r$. For $r = 1$, we have the Manhattan-metric, for $r = 2$ the Euclidean distance. For higher values of r, large differences between variables are extremely overweighted, which is not justified in most multivariate situations.

Division by the number of variables gives the average contribution of variables to the distance:

$$AVD_{jk} = \frac{1}{n} \left[\sum_{i=1}^{n} (x_{ij} - x_{ik})^{2} \right]^{1/2} , \qquad (3.50)$$

or to the sum of absolute differences:

$$MC_{jk} = \frac{1}{n} \sum_{i=1}^{n} | x_{ij} - x_{ik} | . \qquad (3.51)$$

The latter function is known in numerical taxonomy under the name *mean character difference* (Cain & Harrison 1958). In fact, this is the formula applied by Czekanowski in his anthropological investigations ("*durchschnittlische Differenz*"); rather than formula 3.59, which is often – and erroneously – called the Czekanowski index in the literature.

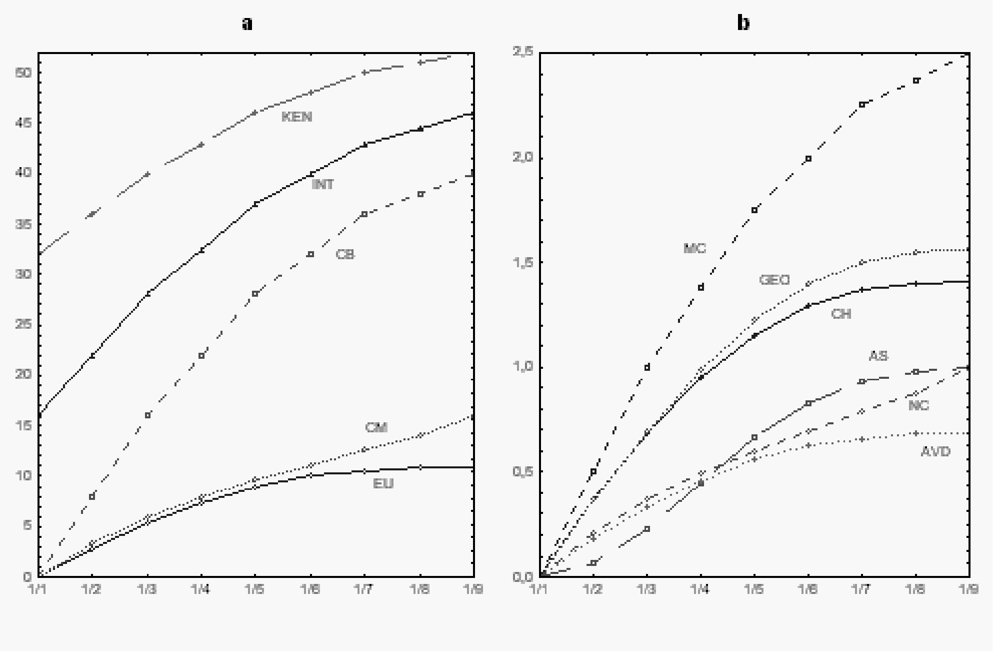

Figure 3.7. Graphical comparison of several distance functions for ratio scale variables. Relativized versions of *EU* and *CM* (abbreviated by *AVD* and *NC,* respectively) are included in **b.**

As Figure 3.7a-b demonstrates, division by *n* does not influence the shape of the curves, yet comparison with other distance formulae becomes easier.

The *Canberra metric* is obtained from the Manhattan metric by standardizing absolute differences for each variable separately with the sum of the two values (Lance & Williams 1967b):

$$CM_{jk} = \sum_{i=1}^{n} \frac{|x_{ij} - x_{ik}|}{|x_{ij}| + |x_{ik}|} \quad . \qquad (3.52)$$

As a result of this operation, the variables will be equalized in importance. For example, in phytosociological quadrats the same absolute difference for rare species will be more influential than for more common species. Use of the absolute values in the denominator was proposed by Gower & Legendre (1986) to extend the applicability of the function to negative values (for example, those standardized by standard deviation previously). Variables taking zero value for both objects have to be excluded from the comparison.

CM is not Euclidean, and its change in the example is only approximately linear (Fig. 3.7a).

The range of Canberra metric is [0,*n*], so division by *n* gives a standardized measure:

Table 3.4. Metric and Euclidean properties of dissimilarity and distance coefficients for interval and ratio scale variables. N = non metric, M = metric, E = Euclidean. * = see text, for explanation.

Function	Property	Function	Property
Euclidean distance	E	Pinkham - Pearson	M
Manhattan metric	M	Gleason	N
Canberra-metric	M	Ellenberg	N
Chord distance	N/E*	Pandeya	N
Angular separation	N	chi-square distance	E
Geodesic metric	E	1 – correlation	N
Clark	E	1 – similarity ratio	M ?
Bray - Curtis	N	Kendall difference	N
Marczewski - Steinhaus	M	Faith intermediate coefficient	N
1 – Kulczynski	N	Uppsala coefficient	N?

$$NC_{jk} = \frac{1}{n}\sum_{i=1}^{n}\frac{|\,x_{ij}-x_{ik}\,|}{|\,x_{ij}\,|+|\,x_{ik}\,|}\ , \tag{3.53}$$

with a result falling into the interval [0,1]. Clifford & Stephenson (1975) proposed consideration of only the number of those variables that have nonzero values in at least one of the objects being compared – a most reasonable suggestion.

If the vectors pointing to the objects are normalized previously to unit length (standardization by norm, Equation 2.22) and we compute Euclidean distance afterwards, then we obtain the *chord distance* (Orlóci 1978). Normalization has been included in the formula below, so standardization is not needed before the calculations:

$$CH_{jk} = \left(2\left[1-\frac{\sum_{i=1}^{n} x_{ij}\,x_{ik}}{\left(\sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{ik}^2\right)^{1/2}}\right]\right)^{1/2} . \tag{3.54}$$

This corresponds to the length of the chord between the two points projected onto the surface of the unit hypersphere (Fig. 3.6c). Consequently, if two objects agree in the proportion (ratio) of variables, chord distance becomes zero (Fig. 3.6f). That is, chord distance for the original objects is not metric, because axiom 1 is not satisfied. For the projected points, the distance is of course Euclidean.

The above formula incorporates the angular separation:

$$AS_{jk} = \ 1 - \frac{\sum_{i=1}^{n} x_{ij}\, x_{ik}}{\left(\sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{ik}^2 \right)^{1/2}} , \tag{3.55}$$

which is the complement of the cosine (Fig. 3.6d) between the two vectors. That is, *AS* becomes 0 if the angle is 0^o ($\cos 0^o = 1$), whereas the function yields 1 for the right angle ($\cos 90^o = 0$). Thus, angular separation can be zero if the two points are not identical.

The *geodesic metric* is related to the previous two coefficients, corresponding to the arch length between the two points on the surface of the unit hypersphere (Fig. 3.6e):

$$GEO_{jk} = \text{arc cos} \frac{\sum_{i=1}^{n} x_{ij}\, x_{ik}}{\left(\sum_{i=1}^{n} x_{ij}^2 \sum_{i=1}^{n} x_{ik}^2 \right)^{1/2}} \quad . \tag{3.56}$$

GEO ranges between 0 and $\pi / 2$. The name reflects that on the surface of the Earth, this is the walking distance between two points rather than the Euclidean distance. Chord distance and geodesic distance are closely related, which is obvious from their formulae (Figure 3.7b). I think chord distance is easier to interpret than the geodesic metric, and further analyses using the geodesic metric provide no additional information anyway.

Euclidean distance and Canberra metric each represent a family of functions, the first based on the *separation* or *differences* over variables (3.47-51), the other sensitive to the *ratios* or *proportions* (3.52-56). The second group includes many more functions, but these are always variants of the equations already known. The closest relative to the Canberra metric is the *coefficient of divergence* proposed by Clark (1952):

$$CL_{jk} = \left(\frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_{ij} - x_{ik}}{x_{ij} + x_{ik}} \right)^2 \right)^{1/2} \tag{3.57}$$

Each term in the summation is squared, so the relationship between this function and Canberra metric is almost the same as between the Euclidean distance and the Manhattan metric (that is, large differences are more pronounced when squared). The result is zero for complete agreement and, due to division by *n*, 1 for maximum difference.

The following formula differs substantially from the Canberra metric, because there is a separate summation for the numerator and the denominator:

$$BC_{jk} = \frac{\sum_{i=1}^{n} | x_{ij} - x_{ik} |}{\sum_{i=1}^{n} (x_{ij} + x_{ik})} . \tag{3.58}$$

This is a simple index, reflecting the proportion of the total in which the two objects differ. The formula is best known as the Bray - Curtis (1957) index, although Pielou (1984) preferred the name *"percentage difference"*. Its complement, often attributed erroneously to Czekanowski, is a well-known similarity coefficient given by:

$$1 - BC_{jk} = \frac{2 \sum_{i=1}^{n} \min [\, x_{ij}, x_{ik} \,]}{\sum_{i=1}^{n} (\, x_{ij} + x_{ik} \,)} \tag{3.59}$$

For presence/absence data, 1-*BC* reduces to the Sorensen index (3.25), so *BC* cannot be metric (Table 3.4). Nevertheless, its linear response to gradual changes in the data is an obvious advantage (Fig. 3.8a).

The *Marczewski - Steinhaus coefficient* (Holgate 1971, Lewandowsky 1972) relates the sum of differences to the sum of the maximum values over the variables:

$$MS_{jk} = \frac{\sum_{i=1}^{n} |\, x_{ij} - x_{ik} \,|}{\sum_{i=1}^{n} \max[\, x_{ij}, x_{ik} \,]} . \tag{3.60}$$

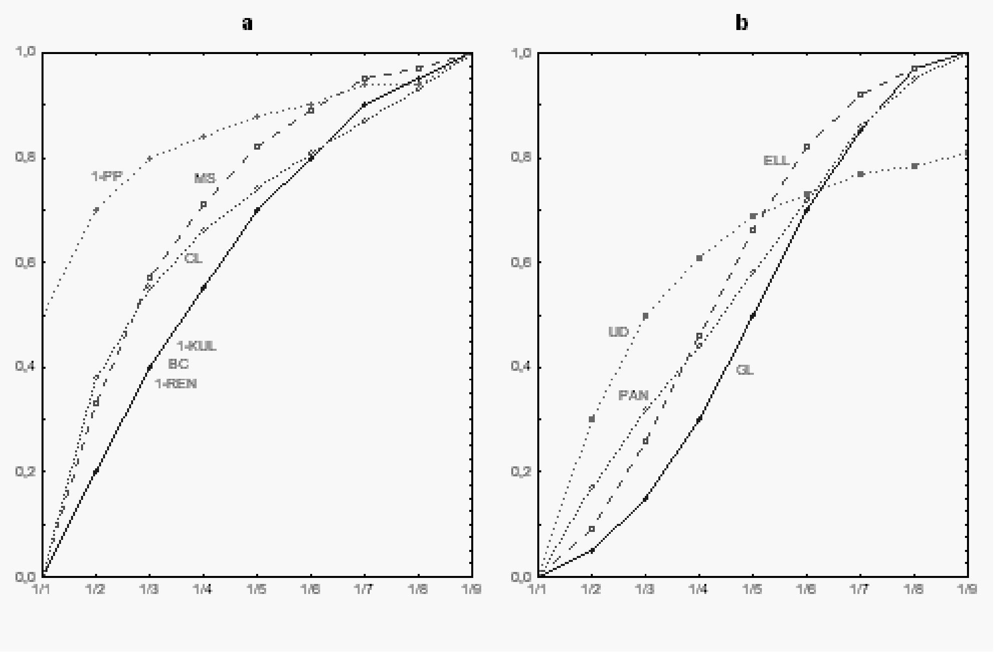

Figure 3.8. Graphical evaluation of some dissimilarity functions for interval variables (**a**) and some special formulae developed for vegetation data (**b**) based on the artificial data of Table 3.3.

This equation may be interpreted in terms of set theory. The numerator is the symmetric difference of the sets representing objects *j* and *k*, and the denominator is the union (Orlóci 1978). *MS* is metric, although not Euclidean. Its complement is known as *Ruzicka index*, written in two alternative forms:

$$RUZ_{jk} = 1 - MS_{jk} = \frac{\sum_{i=1}^{n} \min[x_{ij}, x_{ik}]}{\sum_{i=1}^{n} \max[x_{ij}, x_{ik}]} = \frac{\sum_{i=1}^{n} \min[x_{ij}, x_{ik}]}{\sum_{i=1}^{n} x_{ij} + \sum_{i=1}^{n} x_{ik} - \sum_{i=1}^{n} \min[x_{ij}, x_{ik}]} \tag{3.61}$$

This equation becomes the Jaccard index (3.24) for presence/absence data.

For interval variables, the *Kulczynski index* (3. 29) takes the following form:

$$\frac{1}{2}\left(\frac{1}{\sum_{i=1}^{n} x_{ij}} + \frac{1}{\sum_{i=1}^{n} x_{ik}}\right)\sum_{i=1}^{n} |x_{ij} - x_{ik}| = 1 - KUL_{jk} = 1 - \frac{1}{2}\left(\frac{\sum_{i=1}^{n} \min[x_{ij}, x_{ik}]}{\sum_{i=1}^{n} x_{ij}} + \frac{\sum_{i=1}^{n} \min[x_{ij}, x_{ik}]}{\sum_{i=1}^{n} x_{ik}}\right)$$
(3.62)

For the sample data, 1– *KUL*= *BC* (Fig. 3.8a) because all objects have the same total.

The relationship between the minimum and maximum may be expressed by taking the ratio within the summation. Then, the maximum difference becomes *n*, so division by *n* provides a dissimilarity coefficient with a range of [0,1]:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{|x_{ij} - x_{ik}|}{\max[x_{ij}, x_{ik}]}\right) = 1 - PP_{jk} = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\min[x_{ij}, x_{ik}]}{\max[x_{ij}, x_{ik}]}\right) \tag{3.63}$$

The similarity form is known as the *Pinkham - Pearson coefficient.* The analogy between 1–*PP* and *MS* is the same as the analogy between the normalized Canberra-metric (*NC*, 3.53) and the Bray - Curtis index (*BC*, 3.58). 1–*PP* is not metric, however, since the self dissimilarity is non-zero. The change of this function in the ordered comparison series is quite irregular (Fig. 3.8a). The first problem is alleviated by division using the number of variables that are positive for at least one of the objects, rather than using *n.*

For readers most interested in phytosociological comparisons, some less frequently used formuale are also discussed. These were proposed to meet special requirements in vegetation analysis. For example, absolute differences between species that are present in both sampling units may not contribute to dissimilarity, so the comparison emphasizes differences for species present only in one of the quadrats being compared. Such a similarity coefficient was proposed by Gleason (1920):

$$GL_{jk} = \frac{\sum_{i \to A} (x_{ij} + x_{ik})}{\sum_{i=1}^{n} (x_{ij} + x_{ik})}, \tag{3.64}$$

where *A* is the set of species present in both *j* and *k*. The difference between the numerator and the denominator is caused by species present either in *j* or in *k*. Its complement is a dissimilarity measure, whose behaviour for the sample data set is shown in Figure 3.8b. Ellenberg (1956) suggested considering twice the difference mentioned above (see also Goodall 1973a):

$$ELL_{jk} = \frac{\sum_{i \to A} (x_{ij} + x_{ik})}{\sum_{i=1}^{n} (x_{ij} + x_{ik}) + \sum_{i \downarrow A} x_{ij} + \sum_{i \downarrow A} x_{ik}} . \tag{3.65}$$

For the sample data, 1-*GL* and 1-*ELL* have similar curves with their discrepancy being the highest in the middle of the gradient (Fig. 3.8b). A related similarity measure is due to Pandeya (1961),

$$PAN_{jk} = \frac{\sum_{i \to A} (x_{ij} + x_{ik})}{\sum_{i=1}^{n} (x_{ij} + x_{ik}) + \sum_{i \to A} | x_{ij} - x_{ik} |} \tag{3.66}$$

(see also Goodall 1973a), which does consider differences between the common species as a factor increasing dissimilarity. 1-*PAN* is more linear than the previous two coefficients for the sample data (Fig. 3.8b). Thanks to the set theoretical restrictions applied in 3.64-66, neither of their complements satisfies all metric axioms.

Of the distance measures, the so-called χ^2-*distance* also requires mention. This is the Euclidean distance calculated for objects after standardizing each value by the column and the row total as well:

$$CHISQD_{jk} = \left[\sum_{i=1}^{n} \frac{1}{\sum_{h=1}^{m} x_{ih}} \left(\frac{x_{ij}}{\sum_{s=1}^{n} x_{sj}} - \frac{x_{ik}}{\sum_{s=1}^{n} x_{sk}} \right)^2 \right]^{1/2} \tag{3.67}$$

so its application is limited to data matrices for which variable-, object totals and the grand total are all meaningful (*aggregable* data, Mirkin 1996). The importance of χ^2-distance becomes evident in interpreting correspondence analysis (Section 7.3) because it is rarely used as a distance function by itself.

In addition to the three measures (*AS, CH, GEO*) already mentioned as examples of ratio-sensitive coefficients, some more functions with similar purpose will be introduced. All equations include ratios of the scalar products of vectors (Appendix C), an indication that the function reflects proportions of variables. The scalar product itself, calculated for two column vectors of the data table, is often called the *cross products.* For column vectors *j* and *k,* it is simply:

$$CP_{jk} = \sum_{i=1}^{n} x_{ij} \, x_{ik} \tag{3.68}$$

which is rarely used for raw data (e.g., non-centered PCA, Subsection 7.1.5). Most commonly, the data matrix is centered previously by columns, and the subsequent application of Formula 3.68 yields the cross products of centered data, which in turn, divided by *n*–1, produces the *covariance* of *j* and *k*. Its formula for raw data is given by:

$$COV_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)\,(x_{ik} - \bar{x}_k)}{n-1} \qquad (3.69)$$

which is well-known from standard statistics for expressing linear relationship between variables. Since covariance has neither lower nor upper bounds, and its use is conditioned upon commensurability, the product moment correlation coefficient is used. It is obtained by Formula 3.68 if the data are standardized beforehand by columns or (from raw data) by:

$$COR_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)\,(x_{ik} - \bar{x}_k)}{\sqrt{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \sum_{i=1}^{n}(x_{ik} - \bar{x}_k)^2}} \qquad (3.70)$$

It applies to objects with the reservations mentioned in Section 2.1. The most typical field of application for Functions 3.68-70 is the quantification of inter-variable relationships in principal components and canonical correlation analysis (with reverse indexing, see Sections 7.1-2). The dissimilarity form does not satisfy all the metric properties because zero dissimilarity results for two non-identical objects if one is obtained from the other by multiplication with a constant.

This group of coefficients includes the *similarity ratio* (Wishart 1969, van der Maarel 1979) given by

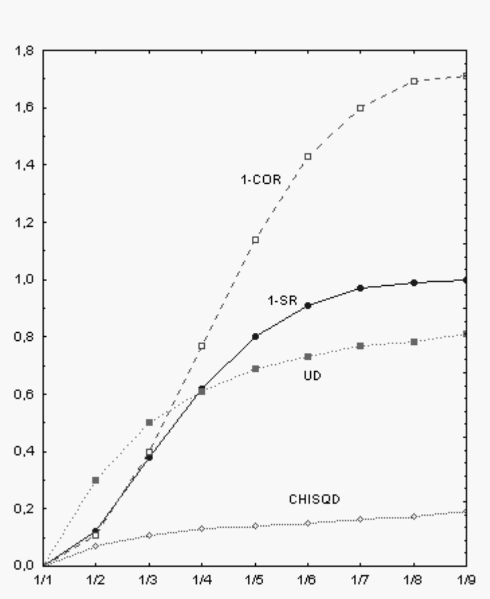

Figure 3.9. The response of some distance and dissimilarity functions to the data changes in Table 3.3.

$$SR_{jk} = \frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{\sum_{i=1}^{n} x_{ij}^2 + \sum_{i=1}^{n} x_{ik}^2 - \sum_{i=1}^{n} x_{ij}x_{ik}} . \tag{3.71}$$

Its value ranges from 0 (complete disagreement) to 1 (identity). For presence/absence data, *SR* is identical to the Jaccard index. Its close relationship with the product moment correlation coefficient is obvious from Figure 3.9.

In addition to separation- and ratio-sensitive measures, there is a third group of coefficients which rely on the minimum agreement between the objects being compared ("minimum value sensitivity", Faith 1984). The basic type of this group is the *minimum value similarity measure* of Kendall (1970):

$$KEN_{jk} = \sum_{i=1}^{n} \min\,[\, x_{ij}, x_{ik} \,]. \tag{3.72}$$

KEN corresponds to the intersection of two sets. Its minimum is 0, whereas there is no fixed upper limit, unlike in most other similarity measures. It is therefore best used for standardized data (see Formula 3.74 below). If we take the maximum value of each variable in the sample as the theoretical upper limit, i.e., $\max_h[\, x_{ih} \,]$, then 3.72 can be expressed as dissimilarity:

$$DKEN_{jk} = \sum_{i=1}^{n} \left\{ \max_h[\, x_{ih} \,] - \min\,[\, x_{ij}, x_{ik} \,] \right\} . \tag{3.73}$$

A version of *KEN* uses standardization by column (object) totals, the formula is well-known in animal ecology as the *Renkonen index:*

$$REN_{jk} = \sum_{i=1}^{n} \min \left\{ \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}, \frac{x_{ik}}{\sum_{i=1}^{n} x_{ik}} \right\} = 1 - 0{,}5 \sum_{i=1}^{n} \left| \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} - \frac{x_{ik}}{\sum_{i=1}^{n} x_{ik}} \right| . \tag{3.74}$$

Another name ("*percentage similarity of distribution*", Whittaker & Fairbanks 1958) explains the meaning of this formula: standardization of abundance data by object (site) totals implies that a relative frequency distribution is obtained for each site, so that 100□*REN* measures their percentage agreement. Owing to standardization by objects, the ratios of variables within the objects become important, so that the distinction between minimum value and ratio sensitivity vanishes. For the sample sequence, 1–*REN* is identical with *BC* (Fig. 3.8a) because the object totals are the same over all objects (hence the agreement with 1-*KUL* as well).

Intermediate forms. It is possible to derive coefficients that represent transitions between measures of different sensitivity, so the result is influenced by both. Faith (1984) and Faith et al. (1987) proposed the use of the arithmetic mean of the Manhattan-metric and *DKEN* ("*intermediate coefficient*"):

$$INT_{jk} = \frac{1}{2}\left[\sum_{i=1}^{n} |x_{ij} - x_{ik}| + \max_h[x_{ih}] - \min[x_{ij}, x_{ik}]\right]. \qquad (3.75)$$

For presence/absence data, we obtain that $INT_{jk} = b+c+d/2$, i.e., the denominator of 1-*FA*2 (cf. Formula 3.22). This function has no upper limit, but this can be resolved by division with *n* (as in Equation 3.22). Another transitional formula is the "*Uppsala coefficient*" (Noest & van der Maarel 1989):

$$UD_{jk} = \frac{1}{n - z_{jk}} \sum_{i=1}^{n} \frac{1}{2}\left[\frac{|x_{ij} - x_{ik}|}{x_{ij} + x_{ik}} + \frac{|x_{ij} - x_{ik}|}{x_{\max} - x_{\min}}\right], \qquad (3.76)$$

where z_{jk} is the number of variables absent from both *j* and *k* (so division is usually not by *n*) and $x_{\max} - x_{\min}$ is the theoretical – rather than the actual – range of the variables. The function is intermediate between the Bray-Curtis index and the Manhattan metric standardized by theoretical range (cf. Gower-index, 3.103, which uses the actual range as a rescaling factor). A feature of this index is that differences between values at the beginning of the scale are more weighted than differences between large values. For example, if $x_{\max} - x_{\min} = 9$, then the differnce between 0 and 1 will contribute to the total by 0.566, whereas the difference between 8 and 9 will add only 0.085. What is done is implicitly analogous to the underweighting of large scores by the logarithmic transformation.

The possibilities for intermediate forms cannot be exhausted, and arithmetic averaging is not the only combining operation. For example, the geometric mean of the correlation coefficient+1 (3.70) and the Gleason-index divided by 2 (see Equation 3.64) was proposed for use in vegetation analysis by Sgardelis & Stamou (1990). The authors claim that their coefficient responds the "best way" to the ordered comparison sequences they tested (they were all different from the gradient shown in Table 3.3 and used throughout in this book).

Genetic distances. Allele frequencies represent special cases of variables measured on the ratio scale. The objects are now populations, whereas the variables are assigned to as many groups as the number of loci examined. The allele frequencies are standardized by the total for each locus, and data tables are presented usually in this relativized form. There are several special formulae for expressing distance between populations based on gene frequencies; all of them consider the above-mentioned grouping of variables and are more or less interpretable genetically. If loci were not considered separately, then any distance function discussed earlier would do the job, but the result were not "genetic". Interpretability has to do with the requirement that genetic distances should reflect the time elapsed since the populations diverged during evolution, and therefore we must have meaningful models on mutations and genetic drift. Of course, distance can be computed formally without any modeling, but the results will have only geometric meaning and cannot serve the basis of biological explanation of evolutionary processes (Weir 1990).

The performance of genetic distances will be illustrated using the simple case of one locus with two alleles. The sample matrix will reflect a gradual drift from the first population until the first allele is completely replaced by the second (we will not dwell on the potential causes of such processes):

allele 1	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.0
allele 2	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0

The number of loci will be denoted by L, whereas a value of the data matrix will be x_{hij}, reflecting the relative frequency of allele i of locus h in population j. n_h is the number of alleles on locus h. Since relative frequencies are used, the points representing populations are on a hyperplane for each locus (for two alleles, they fall onto a line segment shown in Figures 2.9c and 3.11).

A slight modification of Euclidean distance is the *genetic distance of Rogers* (1972):

$$ROG_{jk} = \frac{1}{2L} \sum_{h=1}^{L} \left[\sum_{i=1}^{n_h} (x_{hij} - x_{hik})^2 \right]^{1/2}, \qquad (3.77)$$

which is strongly affected by within-population heterozigosity. The greatest disadvantage of this formula is perhaps the same as the problem with Euclidean distance in ecology: it may happen that the distance between two populations, which have no alleles in common, is smaller than between other two which agree in the presence of many alleles. The *Prevosti distance* (cf. Wright 1978), i.e., the mean character difference per locus, given by:

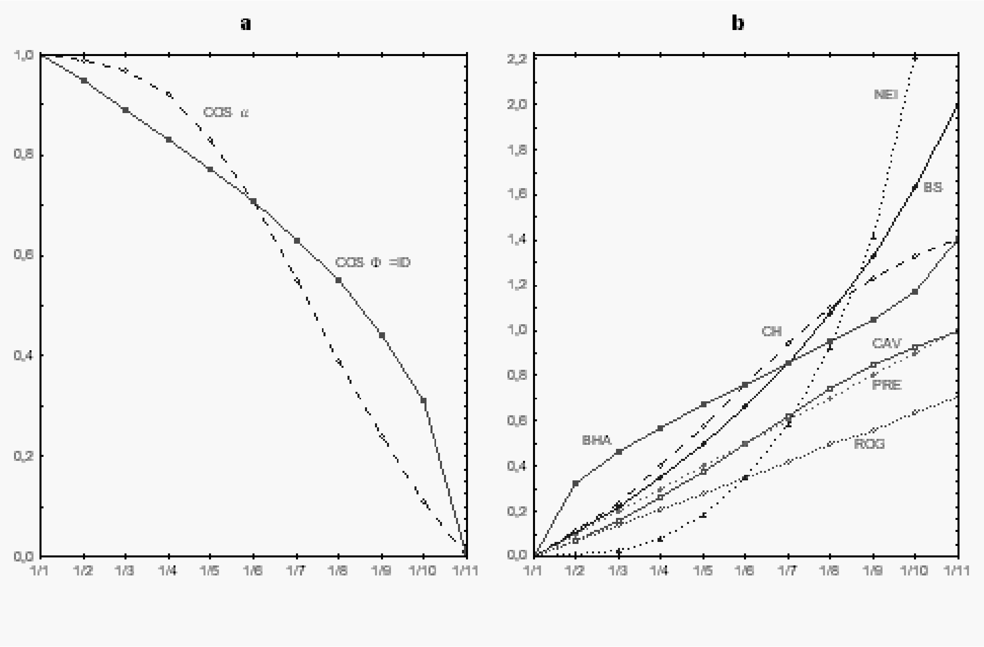

Figure 3.10. The change of genetic cosine functions (**a**) and distances (**b**) along the gradual differentiation of two populations for one locus with two alleles (for frequencies, see text). *CH* is chord distance according to Equation 3.54.

$$PRE_{jk} = \frac{1}{2L} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \; | x_{hij} - x_{hik} | \tag{3.78}$$

may be criticized on similar grounds.

The use of relative frequency differences leads to geometrically interpretable results, but the difficulty mentioned above led most biologists to prefer ratio-sensitive measures. The most common of these are Nei's *genetic identity* (Nei 1972, 1978) and its derivatives. For one locus, genetic identity is in fact measured by Formula 3.55 (without subtraction from 1), which is the cosine of the angle (α, Fig. 3.11) of two vectors pointing towards the populations, Its value is 1 for full agreement and 0 for maximum difference (Fig. 3.10a). Since the formula incorporates relative frequencies, the result has probabilistic interpretation as well. The numerator is the estimated probability that two individuals taken from the different populations will bear the same allele on the locus ($\hat{q}_{jk}$). The two sums of squares in the denominator estimate the probabilities that two individuals from the same population have identical alleles ($\hat{q}_j$ and $\hat{q}_k$, respectively). The denominator is the geometric mean of these two probabilities:

$$ID_{jk} = \frac{\sum_i x_{ij} x_{ik}}{\left(\sum_i x_{ij}^2 \sum_i x_{ik}^2 \right)^{1/2}} = \frac{\hat{q}_{jk}}{\sqrt{\hat{q}_j \, \hat{q}_k}} = \cos \alpha \, . \tag{3.79}$$

This formula may be conceived as the ratio of the genetic identity of populations *j* and *k*, and the mean of within population genetic identities. For *L* loci, the function is generalized as follows:

$$ID_{jk} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{hij} \, x_{hik}}{\left(\sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{hij}^2 \sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{hik}^2 \right)^{1/2}} \; , \tag{3.80}$$

which is a biased estimate of genetic identity, and therefore for small sample size (*m,* the same for all populations) a corrected version (Nei 1978) is recommended:

$$IDC_{jk} = \frac{(m-1) \sum_{h=1}^{L} \sum_{i=1}^{n_h} x_{hij} \, x_{hik}}{\left(\sum_{h=1}^{L} \; (2m \sum_{i=1}^{n_h} x_{hij}^2 - 1) \square \sum_{h=1}^{L} \; (2m \sum_{i=1}^{n_h} x_{hik}^2 - 1) \right)^{1/2}} \, . \tag{3.81}$$

Nei's genetic identity is best understood genetically if we can measure the *time* elapsed since the two populations separated. To do this, one may choose among several models describing population change. In the simplest case, we assume that the mutation from an allele into any other allele has the constant rate of μ and obtain the following relationship:

$$NEI_{jk} = -\ln ID \approx 2\mu t \,, \tag{3.82}$$

which is *Nei's genetic distance.* It is undefined for the case when all alleles appear only in one of the populations (Fig. 3.10b). Nei's distance implies major oversimplifications on population change, because a constant rate of mutation is assumed after the divergence on all loci and on both sister lines (Hillis 1984). Hillis suggested overcoming this problem by calculating the arithmetic mean of genetic identities obtained for the loci separately:

$$HIL_{jk} = -\ln \left[\frac{1}{L} \sum_{h=1}^{L} \frac{\sum_{i=1}^{n_h} x_{hij}\, x_{hik}}{\left(\sum_{i=1}^{n_h} x_{hij}^2 \sum_{i=1}^{n_h} x_{hik}^2 \right)^{1/2}} \right]. \tag{3.83}$$

The unbiased estimate of this quantity is obtained analogously as in Formula 3.81 (Swofford & Olsen 1990). For a single locus, *HIL* is identical to *NEI.*

Nei's distance is inapplicable to cases where genetic drift only is responsible for the separation of populations. Then, a geometric measure, the *Balakrishnan - Shangvi distance* (Weir 1990) is in order:

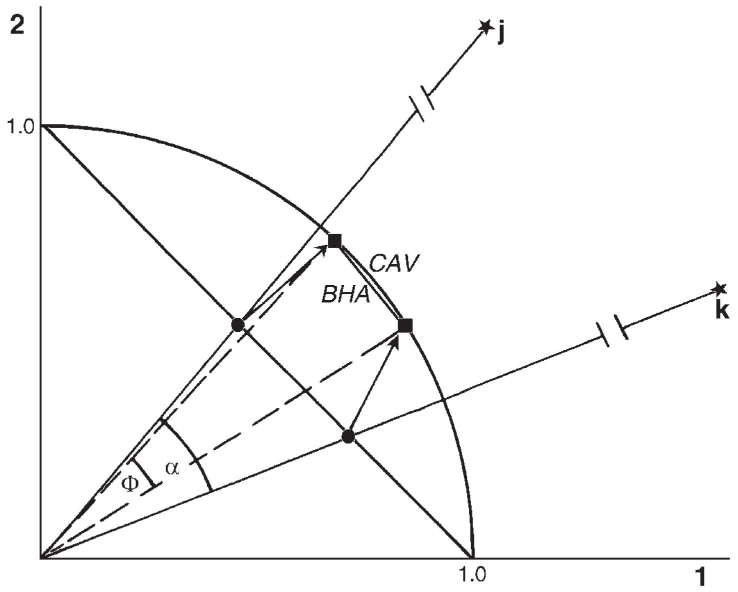

Figure 3.11. Geometric illustration of some genetic distance measures for one locus and two alleles. ✭ denotes individual frequency values (10 and 12 for *j;* and 20 and 8 for *k).* ● denotes relative frequencies and ■ indicates the relative frequencies, squared, thus projected to the unit circle.

$$BS^2_{jk} = \frac{1}{\sum_{h=1}^{L} n_h - 1} \sum_{h=1}^{L} \sum_{i=1}^{n_h} \frac{(x_{ij} - x_{ik})^2}{x_{ij} + x_{ik}} . \tag{3.84}$$

Defining and interpreting genetic distances is a complicated matter, best illustrated by the formulae proposed by Cavalli-Sforza and co-workers. For a given locus *h,* if the relative frequencies of alleles are squared, then the points representing populations will fall onto the surface of unit hypersphere (Fig. 3.11). Afterwards, the angle between vectors pointing to *j* and *k* is obtained as:

$$\cos \Phi = \sum_{i=1}^{n_h} \sqrt{x_{hij}\, x_{hik}} . \tag{3.85}$$

Starting from this, Cavalli-Sforza & Edwards (1967) derived a genetic distance measure by standardizing and averaging the arc lengths (geodesic distances, Equation 3.56, Fig. 3.11) pertaining to the loci:

$$CAV_{jk} = \left[\frac{1}{L} \sum_{h=1}^{L} \left(\frac{2}{\pi} \arccos \sum_{i=1}^{n_h} \sqrt{x_{hij} x_{hik}} \right)^2 \right]^{1/2} \tag{3.86}$$

Alternatively, the chord distance between points projected to the surface of the hypersphere can be calculated for each locus:

$$BHA_{jk} = \left[2 - 2 \sum_i \sqrt{x_{ij}\, x_{ik}} \right]^{1/2} = \left[\sum_i \left(\sqrt{x_{ij}} - \sqrt{x_{ik}} \right)^2 \right]^{1/2} \tag{3.87}$$

(*Bhattacharyya distance*, cf. Mardia et al. 1979, Fig. 3.11), and then averaged over the loci. Weir (1990) argues that these are merely geometric measures without any genetic implications. A problem is that Φ and α, the latter appearing in Nei's genetic identity are different angles (Fig. 3.11), and it seems that Nei's measure and its α are even more understandable geometrically than Cavalli-Sforza et al.'s functions. A practical argument in favour of using the angle Φ is that cos Φ is approximately linear over increasing differences between the allele frequencies (Tóthmérész 1986), and this is not so with cos α (Fig. 3.10a). Swofford & Olsen (1990) strongly support Cavalli-Sforza's measures by giving a genetic justification as well. They maintain that genetic drift is well-expressed by Function 3.86 because the distance value itself is independent of the initial gene frequencies. Mardia et al. (1979: 379) show mathematically that, for a single locus case, perhaps Swofford & Olsen are right; there is a simple arithmetic relationship between the Balakrishnan - Shangvi distance and the Bhattacharyya distance.

Measures of niche overlap. The measurement of niche breadth and overlap of species may be a good starting point for a multivariate analysis of interspecific relationships. The measures of *niche overlap* are special distance or similarity functions, and it is no surprise that coefficients well-known in other areas often have different names in "niche jargon". An example is the *Schoener* (1970) index, which corresponds to the Renkonen index as applied to species (that is, standardization by the total of each species). The Horn (1966) formula is based on the concepts of information theory. Let *n* now be the number of sampling units, given in rows of the data matrix. Column vector *j* can be conceived as the frequency distribution of species *j*, and the *niche breadth* of that species is expressed by Shannon's entropy function:

$$\hat{H}_j = -\sum_i \frac{x_{ij}}{\sum_h x_{hj}} \log \frac{x_{ij}}{\sum_h x_{hj}} \,. \tag{3.88}$$

Species *j* and *k* are in complete overlap if the above quantity does not change upon merging the two column vectors. This is the minimum value for the merged vectors, denoted by $\hat{H}_{\min}$. The difference between the species is the greatest, i.e., their overlap is zero, if they never co-occur. For such a case, let $\hat{H}_{\max}$ be the entropy of the merged vector. All actual values, $\hat{H}_{obs}$, will fall between these extreme values. If rescaled according to the maximum range:

$$HN_{jk} = \frac{\hat{H}_{\max} - \hat{H}_{obs}}{\hat{H}_{\max} - \hat{H}_{\min}} \,, \tag{3.89}$$

the function will take a value of 0 for full disagreement, and 1 for complete agreement of species niche. A formula more suitable for calculations is given by

$$HN_{jk} = \frac{\sum_{i=1}^{n} (x_{ij} + x_{ik}) \log (x_{ij} + x_{ik}) - \sum_{i=1}^{n} x_{ij} \log x_{ij} - \sum_{i=1}^{n} x_{ik} \log x_{ik}}{(x_{.j} + x_{.k}) \log (x_{.j} + x_{.k}) - x_{.j} \log x_{.j} - x_{.k} \log x_{.k}} \,, \tag{3.90}$$

in which $x_{.j}$ and $x_{.k}$ are the totals of the *j*-th and *k*-th columns, respectively. The formula may also be used as a similarity index between sample sites.

Similarity of shapes. Penrose (1954) suggests that Euclidean distance can be partitioned into two components, one due to *size* differences only, and the other to *shape* differences only:

$$d_{jk}^2 = (n-1)\, SHAPE_{jk}^2 + n\; SIZE_{jk}^2 \tag{3.91}$$

If one wishes to disregard size when comparing two objects, the shape coefficient of Penrose may be used:

$$SHAPE_{jk} = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_{ij} - x_{ik})^2 - \frac{1}{n(n-1)} \left[\sum_{i=1}^{n} (x_{ij} - x_{ik}) \right]^2 . \tag{3.92}$$

This quantity is essentially the *variance* of deviates for each character of the objects compared (mean square – square mean). Its expectation is large if the objects differ both in magnitude and direction (i.e., sign) of deviations. The size coefficient is given by

$$SIZE_{jk} = \left[\frac{1}{n^2} \left[\sum_{i=1}^{n} (x_{ij} - x_{ik}) \right]^2 \right]^{1/2} . \tag{3.93}$$

which will provide a large value if the differences are generally of the same sign. Note the unsymmetry involved in this size and shape comparison.

Rohlf & Sokal (1965) suggest that correlation (Equation 3.70) is preferable over the *SHAPE* function of Penrose in expressing shape similarity. Some very special variants of component analysis discussed in section 7.6 provide an even more sophisticated morphometric means for describing and analyzing shape variation, so the Penrose coefficients have lost their practical and theoretical importance, and need attention from a historic perspective only.

Generalized distance. When comparing two objects by Euclidean distance, the effect of correlated variables is overemphasized in the result. This is an "undercover" version of implicit weighting, which is practically always present because biological variables are correlated (if not much, we can still have spurious correlation among them). Let us illustrate the effect of this implicit weighting by the following artificial data matrix:

variable 1	5.1	6.2	7.1	8.0
variable 2	4.0	5.0	6.2	7.3
variable 3	3.0	2.0	9.0	6.0

The first two variables have high positive correlation and it is likely that they reflect the influence of a third, background variable that was not examined directly. So, consideration of both of them will overweight the background variable compared to variable 3 in the matrix. It *may*

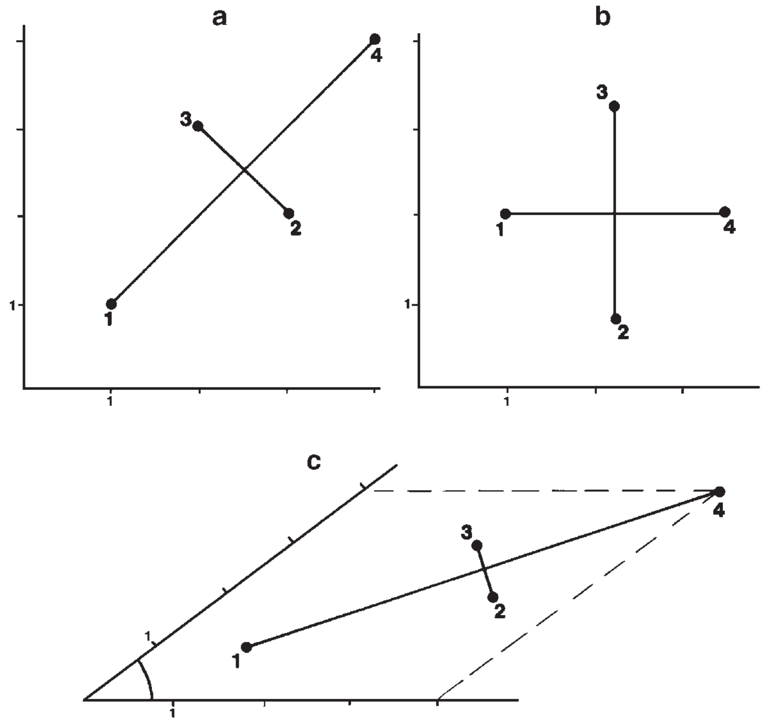

Figure 3.12. Euclidean distances of points in an *orthogonal* coordinate system (**a**), generalized distances in an *arbitrary,* orthogonal coordinate system *(***b**) and Euclidean distances in an *oblique* coordinate system (**c**). COR_{12} = 0.8, therefore the angle between axes is arc cos 0.8 = 36.8°. Dotted lines help determine the position of point 4 in the oblique system.

be undesirable in interpreting the results. Overweighting due to correlated variables can be completely eliminated from the comparisons using the *Mahalanobis* (1936) *generalized distance:*

$$GEND^2_{jk} = \sum_{h=1}^{n} \sum_{i=1}^{n} w_{hi}\,(x_{hj} - x_{hk})\,(x_{ij} - x_{ik}) \qquad (3.94)$$

or, in matrix algebraic terms:

$$GEND^2_{jk} = (\mathbf{x}_j - \mathbf{x}_k)'\,\mathbf{W}^{-1}\,(\mathbf{x}_j - \mathbf{x}_k)\,, \qquad (3.95)$$

where $\mathbf{x}_j$ and $\mathbf{x}_k$ are the column vectors corresponding to objects *j* and *k,* respectively, $\mathbf{W}^{-1}$ is the inverse of the variance-covariance matrix of the *n* variables (Appendix C), and w_{hi} is one of its elements. Thus, the Mahalanobis distance involves standardizing all variables to unit variance. Therefore, if the original variables are perfectly uncorrelated, then the result of Equation 3.95 will be identical to the squared Euclidean distance calculated from standardized data. The matrix of generalized distances conveys metric information for standardized and orthogonal axes. A principal coordinates analysis (metric multidimensional scaling, see Subsection 7.4.1) from such a matrix will derive equally "important" axes (i.e., the total variance will be shared equally by the ordination axes). These are important points, even though full understanding is conditioned upon our knowledge of some ordination theory.

The equilibrating effect of the Mahalanobis distance is demonstrated in Figures 3.12a-b. The Euclidean and Mahalanobis distances between the four points are included in the following two semi-matrices:

0					0			
2.23	0			and	1.73	0		
2.23	**1.41**	0			1.73	**2.45**	0	
4.24	2.23	2.23	0		**2.45**	1.73	1.73	0

The Mahalanobis distance is sensitive to the two main directions in the point scatter, according to the positions of points 1-4 and 2-3, and they are considered equally. Consequently, distances d_{14} and d_{23} (boldface in the above tables) will be equal (Fig. 3.12b). These main "directions" will be treated in more detail in Chapter 7.

Generalized distance has originally been suggested to measure distance between groups of objects (e.g., populations). For this case, we have the following formula:

$$GEND^2_{jk} = (\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_k)'\,\mathbf{W}^{-1}\,(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_k)\,, \qquad (3.96)$$

where $\bar{\mathbf{x}}_j$ and $\bar{\mathbf{x}}_k$ are the mean vectors of groups *j* and *k* (that is, the means of variables presented in a column vector), and $\mathbf{W}^{-1}$ is the inverse of the **W** variance/covariance matrix (computed for all groups based on the pooled data). The distance is meaningful only if the within-group covariances are homogeneous (more precisely, they estimate the same common covariance matrix), and the distribution of variables is multivariate normal. Sneath & Sokal (1973) emphasize, however, that the distance measure is less sensitive to the violation of these conditions (robustness). Note, further, that computation of generalized distance is possible only if the number of objects is not less than the number of variables. Otherwise matrix **W** is singular (Appendix C) and cannot be inverted. The same problem arises if the correlation of any two variables is –1 or 1, or if the variance of one or more variables is zero.

Distance in oblique coordinate systems. We did not say explicitly, because it appeared so obvious, that our data were visualized in a coordinate system with axes orthogonal to each other (i.e., the angle between any two is 90$^{\mathrm{o}}$). After shifting from this to an oblique coordinate system, the interpoint distances will be influenced by the relationships among variables, since the cosine of the angles between axes corresponds to the correlation of the variables they represent. The formula is given by:

$$OBL_{jk} = \left[\sum_{h=1}^{n} (x_{hj} - x_{hk})^2 + 2 \sum_{h=1}^{n-1} \sum_{i=h+1}^{n} (x_{hj} - x_{hk})\,(x_{ij} - x_{ik})\, COR_{hi} \right]^{1/2}, \qquad (3.97)$$

in which COR_{hi} is the product moment correlation of variables h and i (Formula 3.70, see Orlóci 1978: 49). One component of the function is the squared Euclidean distance, the other is a correction factor. This factor is positive if objects *j* and *k* “support” the correlation of variables (e.g., points 1 and 4 in Fig. 3.12a). As a result, the new distance will be larger than the Euclidean. If the relative position of the two objects being compared is contradictory with the correlations (such as points 2 and 3, Fig. 3.12a and c), then the correction factor takes a negative value, and the new distance will be smaller than the Euclidean distance. The application of oblique coordinate systems implies therefore that, depending on the object pair in question, the effect of correlated variables is either enhanced or diminished.

To sum up, contrary to the Mahalanobis distance, measure 3.97 emphasizes the main direction or trend behind the correlation of the two variables, whereas the effect perpendicular to it is neglected. The distance matrix of points in Figure 3.12c is as follows:

$$\begin{array}{cccc} 0 & & & \\ 2.92 & 0 & & \\ 2.92 & \mathbf{0.64} & 0 & \\ \mathbf{5.69} & 2.92 & 2.92 & 0 \end{array}$$

Special measures. There are several coefficients developed for the interval or ratio scale variables that do not fit logically into any of the aforementioned groups. Such a measure is the *Calhoun distance* (Bartels et al. 1970), which is sensitive to the topological relationships of points. The basic idea is that the distance of two points, *j* and *k*, is determined by the number of other points that appear in between the first two in the multidimensional space. That is, like in Formulae 3.94 and 3.97, interpoint distances are influenced by the other points as well, this effect being the most apparent for the Calhoun measure. The way “betweenness” is considered in the calculations is illustrated by Figure 3.12, using the small data matrix given by:

variable 1	2	5	1	2	3	6	7	7	7
variable 2	2	5	1	6	4	3	2	7	6

For a given pair of objects, we find an interval for each variable, such that these intervals determine a hypercube in the multidimensional space. For points 1 and 2 in the above matrix, we are concerned with the *unshaded* portions in the diagram of Figure 3.13a.

To obtain the Calhoun measure, we introduce several auxiliary parameters:

n_1 = the number of points falling *inside* the hypercube determined by the variables (in two dimensions, the number of points found within the square of Fig. 3.13a, which are points 5 and

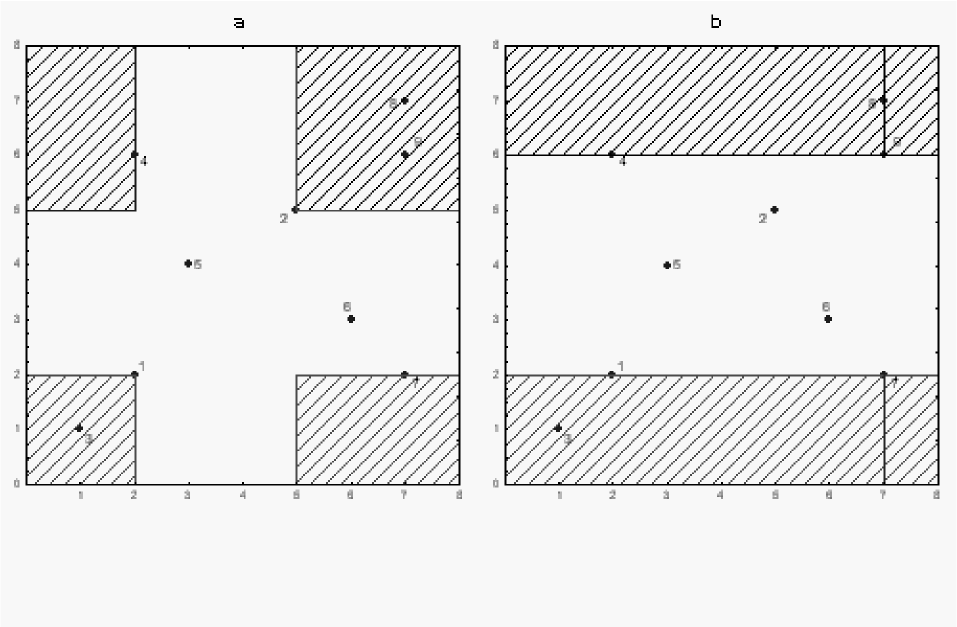

Figure 3.13. Determining Calhoun "distance" for an artificial two-variable case. **a:** relationships for points 1 and 2, **b**: and for points 7 and 9.

6);

n_2 = the number of points falling exactly onto the surface of the hypercube, i.e., the number of points that agree with object *j* or *k* in at least one variable (points 4 and 7, Fig. 3.13a);

n_3 = the number of points that agree with both objects being compared in at least one variable and fall outside the hypercube (in Figure 3.12a there is no such point; however, in calculating the Calhoun distance between 7 and 9, we find that point 8 takes such a position, Fig. 3.13.b).

Then, the measure sought is obtained as:

$$CAL_{jk} = w_1\, n_1 + w_2\, n_2 + w_3\, n_3 \qquad (3.98)$$

in which w_1, w_2 and w_3 are arbitary weights (according to the proponents of the measure, they should be equal to 6, 3 and 2, respectively). For Orlóci (1978), a more logical choice is defining $CAL_{jk} = n_1$ (that is, $w_2 = w_3 = 0$), since actually only n_1 points fall between points *j* and *k,* so that the arbitrariness in selecting the weights is thus eliminated. The Calhoun "distance" is not metric, so it should not be called "distance" (*CAL* may be zero for two points even though they are different). There is an advantage, however, that scale differences do not influence the result.

Goodall (1964, 1966) proposed the *probabilistic similarity index,* which defines the similarity of two objects as the function of the other similarities. The pairwise similarity is thus affected by the whole sample, recalling the analogous behaviour of Equations 3.94, 3.97 and 3.98. Yet, as shown by the computational example below, the basic idea is greatly different from that of all other similarity functions.

1. Let $d_{i,jk} = |\, x_{ij} - x_{ik}\,|$ be the Manhattan distance of objects *j* and *k* for variable *i*. In the sample of size *m,* there are $m(m–1)/2$ such values for each variable. Let us find the rank order of

these $d_{i,jk}$ values for each variable.

2. For variable *i,* define the dissimilarity of objects *j* and *k* as the proportion of values smaller than or equal to $d_{i,jk}$ in the sample. That is, let

$$p_{i,jk} = \frac{\#\,(d \le d_{i,jk})}{m(m-1) \big/ 2}\,. \tag{3.99}$$

The larger this value the greater the discrepancy between the two objects in relation to the entire sample. $p_{i,jk}$ is the probability of the event that, for a pair of objects, the dissimilarity value for variable *i* is not larger than $d_{i,jk}$ if x_{ij} and x_{ik} were chosen at random from the total of *m* values. This probability is therefore inversely proportional to similarity.

3. Having determined $p_{i,jk}$ for all variables, obtain the following product:

$$q_{jk} = \prod_{i=1}^{n} p_{i,jk}\,. \tag{3.100}$$

4. This step involves a second ranking procedure, the $m(m–1)/2$ different *q* values are arranged in ascending order. The similarity of objects *j* and *k* is defined as the proportion of *q* values larger than q_{jk}:

$$GD_{jk} = \frac{\#\,(q > q_{jk})}{m(m-1) \big/ 2}\,. \tag{3.101}$$

This is the probability that objects *j* and *k* are at least as similar to each other as if their original data values were chosen completely randomly from the sample.

Goodall's index is without question a very ingenious expression of within-sample relative similarity of objects. At the same time, such a relativization may prove to be a disadvantage, because the similarities are valid for the given sample only; addition of a new object or a variable can completely upset the original similarity structure. Although metric information in the data is lost thanks to the ranking procedure, index 3.101 may be very useful in biological classifications. Whether the variables are commensurable or not is irrelevant, because the scores are ranked separately for each variable. Suitable modification of Formula 3.99 may extend its utility to ordinal and nominal variables.

If the variables are considered stochastically independent from one another, then the similarity is not calculated formally only. Fisher (1963) has shown that the quantity

$$X^2 = -\ln \sum_{i=1}^{n} \ln p_{i,jk} \tag{3.102}$$

follows the χ^2 distribution with $2n$ degrees of freedom. The larger the value of 3.102, the higher the similarity between the two objects.

An extension of the above formula is the *affinity index* (Goodall 1968), measuring the "affection" of an object to a group, considering also its similarities to objects outside this group. This index may serve as a basis of deciding whether the objects should be assigned to the group. The *deviant index* (Goodall 1966) does the opposite; it expresses the deviation of each object from the group to which it is classified.

3.6 Coefficients for mixed data

Similarity and distance functions discussed thus far do not apply to data sets containing variables measured on different scales. This problem could be solved by the appropriate conversion of variables into the same scale, which either involves loss if information or requires incorporation of some external information. If one wishes to retain the variables in their original form (which is usually the case), then a reasonable solution is offered by formuale developed specifically for mixed data types. The best-known of these is the Gower (1971b) index, which has the additional advantage of allowing some missing scores in the data. The formula is given by:

$$GOW_{jk} = \frac{\sum_{i=1}^{n} w_{ijk}\ s_{ijk}}{\sum_{i=1}^{n} w_{ijk}}, \qquad (3.103)$$

where $w_{ijk} = 0$ if objects j and k cannot be compared for variable i because either x_{ij} or x_{ik} is unknown. In addition, for

a) binary variables we define

$w_{ijk} = 1$ and $s_{ijk} = 0$ if $x_{ij} \neq x_{ik}$
$w_{ijk} = s_{ijk} = 1$ if $x_{ij} = x_{ik} = 1$ or if $x_{ij} = x_{ik} = 0$ and double zeros (mutual absences) are included;
$w_{ijk} = s_{ijk} = 0$ if $x_{ij} = x_{ik} = 0$ and the double zeros are excluded from the comparison;

b) for nominal variables:

$w_{ijk} = 1$ if x_{ij} and x_{ik} are known; then let
$s_{ijk} = 0$ if $x_{ij} \neq x_{ik}$
$s_{ijk} = 1$ if $x_{ij} = x_{ik}$

c) for variables measured on the interval and ratio scale:

$w_{ijk} = 1$ if x_{ij} and x_{ik} are both known, and $s_{ijk} = 1 - \{\, | x_{ij} - x_{ik} | \,/\, (\text{range of variable } i\,)\}$.

The complement of 3.103 is a *dissimilarity* measure. Note that for the presence/absence case, with double zeros included, the index reduces to the simple matching coefficient (3.6), with double zeros excluded we obtain the Jaccard index (3.24). For nominal variables, Gower's formula implies index 3.33; and for interval and ratio scale variables, the dissimilarity form is the mean character difference (3.51) calculated such that each variable is standardized by range previously.

The Gower index, in its original form, does not incorporate ordinal variables, a serious shortcoming if we consider that mixed data sets often have ordinal-type variables. As a solution, I recommend adding the following, straightforward procedure (Podani 1999) to the above definition of the index:

d) for ordinal variables:

w_{ijk} is the same as above; all x_{ij} are replaced by their ranks r_{ij} determined over all objects (such that ties are also considered, as in rank correlation), and then $s_{ijk} = 1 -$

[$|r_{ij}–r_{ik}|$ / (max $\{r_i\}$–min $\{r_i\}$)]. If ties appear, then correction terms are added to both the denominator and the numerator.

As seen, the idea is to involve differences in ranks for two items within the same rank order, which is somewhat analogous to taking differences between ranks for the same item in two orders, as implied in Spearman's *RHO* (3.43). Standardization by the range of ranks for each variable ensures commensurability with the other variable types. If the index is calculated for purely ordinal variables, then the above definition provides a new ordinal similarity measure allowing simultaneous presence of non-commensurable ordinal variables (contrary to Equations 3.45-46 that require commensurability, e.g., all variables were species AD scores, which need not be so with the new coefficient).

An alternative to Gower's index is the following distance coefficient (Podani 1980):

$$DM_{jk} = \left(\sum_{i=1}^{n} w_{ijk} \left[\frac{x_{ij} - x_{ik}}{q_{ijk}} \right]^2 \right)^{½}, \qquad (3.104)$$

where $w_{ijk} = 0$ if comparison of objects j and k for variable i is invalid for lack of data, otherwise $w_{ijk} = 1$;

a) for binary variables:
$q_{ijk} = 1$.

b) for nominal variables:
$q_{ijk} = x_{ij} - x_{ik}$ if $x_{ij} \neq x_{ik}$
$q_{ijk} = 1$ if $x_{ij} = x_{ik}$

c) for interval and ratio scale variables:
q_{ijk} = max (x_{ih}) – min (x_{ih}) ; h=1,..., m.

d) for ordinal variables, analogously with the above new definition:
q_{ijk} = max $\{r_i\}$ – min $\{r_i\}$.

3.7 Generalization of distances to more than two objects (heterogeneity measures)

Pairwise distances are merely special, but favoured, forms of comparison, and they are not always applicable. A number of clustering procedures, for example, rely on measures expressing some *internal property* of the group of two or more objects. We shall refer to these properties by the term *heterogeneity* (and its counterpart being *homogeneity*), admitting that this is perhaps not the best usage of the word. The heterogeneity of object clusters, say A, will be measured by statistics well-known from traditional biometry, as well as by information-theory functions.

A fundamental heterogeneity measure is the *sum of squares* (i.e., sum of squared deviations from the mean) within the subset of objects, given by the familiar formula:

$$SSQ_A = \sum_{i=1}^{n} \sum_{j \to A} (x_{ij} - \bar{x}_{iA})^2 , \tag{3.105}$$

where $\bar{x}_{iA}$ is the average of variable i in the subset A of objects. Formula 3.105 can be written using all the pairwise Euclidean distances within A:

$$SSQ_A = \frac{\sum_{j=1}^{m_A} \sum_{k=1}^{m_A} d_{jk}^2}{2m_A} , \tag{3.106}$$

in which m_A is the number of objects in A. Thus, for two objects the sum of squares is half of their squared Euclidean distance:

$$SSQ_{jk} = d_{jk}^2 \big/ 2 . \tag{3.107}$$

Sum of squares divided by the number of objects yields the *variance:*

$$VAR_A = SSQ_A / m_A = \frac{\sum_{i=1}^{n} \sum_{j \to A} (x_{ij} - \bar{x}_{iA})^2}{m_A} , \tag{3.108}$$

which can also be expressed as:

$$VAR_A = \frac{\sum_{j=1}^{m_A} \sum_{k=1}^{m_A} d_{jk}^2}{2m_A^2} . \tag{3.109}$$

The variance for two objects is given by

$$VAR_{jk} = d_{jk}^2 \big/ 4 . \tag{3.110}$$

The heterogeneity of a set of objects can also be expressed as the average of their distances or dissimilarities (abbreviated as DIS_{jk}):

$$AVG_A = \frac{\sum_{j=1}^{m_A - 1} \sum_{k=j}^{m_A} DIS_{jk}}{(m_A^2 - m_A) \big/ 2} , \quad j,k \to A . \tag{3.111}$$

Its advantage is that averages are meaningful for any symmetric measure, even in non-metric spaces, whereas variance and sum of squares are closely linked to the concept of the Euclidean distance.

If the m_A objects in A are described in terms of n nominal variables in which the number of states for variable i is p_i, then hererogeneity may be expressed by the weighted pooled entropy of variables:

$$H_A = n \, m_A \log m_A - \sum_{i=1}^{n} \sum_{h=1}^{p_i} f_{hi} \log f_{hi} , \tag{3.112}$$

where f_{hi} is the frequency of state h of variable i in A. Formula 3.112 measures the *disorder* of the set of objects. Disorder is the minimum if all objects are the same for all variables, and maximum is reached if, for each state h of each variable i, we have $f_{hi} = m_A / p_i$. For the case where $p = 2$ and there are two objects, the above formula can be written using the abbreviations of the 2×2 contingency table:

$$H = 2\,(b+c)\log 2 \qquad (3.113)$$

which is even simpler if the base of the logarithm is 2:

$$H = 2\,(b+c) \qquad (3.114)$$

The other information theory measure for characterizing set A is the (mutual) *information* conveyed by the variables. Low value indicates high coincidence of variables, implying high inter-object similarity. For binary data, information is obtained as:

$$I_A = (n-1)m_A \log m_A - \sum_{i=1}^{n} \left[\, f_i \log f_i - (m_A - f_i)\log(m_A - f_i)\right] + \sum_{g=1}^{\omega} f_g \log f_g \qquad (3.115)$$

where f_i is the number of occurrences of variable i in A, and f_g is the frequency of variable combination g in A. The number of possible variable combinations is $\omega = 2^n$. For two objects we have:

$$I = 2(b+c-1)\log 2 \qquad \text{if } b+c > 0; \qquad (3.116)$$

and

$$I = 0 \qquad \text{if } b+c = 0. \qquad (3.117)$$

Note that Formula 3.115 has a central role in the analysis of multispecies patterns (see Juhász-Nagy 1976).

3.8 Literature overview

In sharp contrast with the limited literature discussing the relevance of sampling and data transformation in multivariate studies, a little library is available on distance and dissimilarity functions. The choice of a formula best suited to the given problem has been covered by many papers and book chapters. The possibilities are by no means exhausted and new coefficients are suggested and "discovered" for special purposes at a rate of perhaps one new measure per week. The recent literature is fairly rich in the evaluation of the mathematical properties of dissimilarity coefficients. Therefore, only some fundamental sources that provide a thorough overview of a particular topic will be mentioned.

The most complete overview of dissimilarity functions used in ecology is found in Goodall (1973a) and Orlóci (1978). Legendre & Legendre (1983:170-215) also list several formulae. Pielou (1984) and Greig-Smith (1983:194-195) restrict themselves to much fewer coefficients whose properties are thoroughly examined though. The distinction between R-mode and Q-mode coefficients is almost always present, which is not necessarily desirable, as I emphasized in the previous chapter. The first detailed analysis of presence/absence coefficients is due to Cheetham & Hazel (1969), with emphasis on paleontological applications. Kenkel & Booth (1987) have examined the usefulness of presence/absence coefficients in biogeographic studies. Lamont & Grant (1979) and Hajdu (1981) provide comparisons using ordered series of object pairs, to evaluate the performance of many coefficients under different circumstances. Their graphical method inspired the comparative approach of this book. The method was also used by Shi (1993) in evaluating and comparing not less than 39 presence/absence coefficients. Further comparative studies include Campbell (1978), Janson &

Vegelius (1981), Hubálek (1982), Wolda (1981), Jackson et al. (1989) and – more recently – Batagelj & Bren (1995). For taxonomists, but also for people with general interest in biological applications, Sneath & Sokal (1973) is still the best reference. A great value of this monograph is the complete bibliography of the first two decades of numerical taxonomy. For microbiologists, Austin & Colwell's (1977) comparative analysis of presence/absence coefficients is recommended.

From the viewpoint of mathematics, Anderberg's (1973) pioneer work (which will be cited many times later) is a good summary. The metric and Euclidean properties of dissimilarity functions were examined by Gower & Legendre (1986). A full account of information theory measures is given in Feoli et al. (1984). Mathematical properties of some ordinal measures are discussed very recently – and deeply – by Monjardet (1997).

For the comparison of molecular sequences, this book is admittedly incomplete: there are a wider a range of methods that were not covered here (see e.g., Miyamoto & Cracraft [1991], for a fuller account of the topic). Measures of niche-overlap are summarized in more detail by Abrams (1980), Hurlbert (1982) and Ganis (1991). Similarity of shapes, as mentioned above, has been rarely defined in terms of distances recently. There were significant developments in the analysis of biological form in the past ten years. Some methodological results of this new approach, called geometric morphometry, will be revisited in Section 7.6.

Any newcomer consulting the aforementioned literature may have the immediate impression of being lost in a jungle. It is doubtful whether there is a single coefficient judged unanimously as good or bad by the specialists of the area. We might say that different objectives, different objects, and various aspects of evaluation are mixed up in an apparently messy and incomprehensible manner. Completely contrasting conclusions often appear when examining the utility of the same coefficient. For instance, chord distance was found by Kenkel & Orlóci (1986) to outperform the others in ecological ordinations, whereas Faith et al. (1987) believe that this measure is ecologically irrelevant. Strange enough: both studies may very well be correct on their own right. Clearly, there is a need of an up-to-date, comprehensive and impartial review of the topic, an important job yet to be done. Maybe a new theoretical foundation awaits development. It is comparatively less critical that ambiguities and errors often appear when Euclidean or metric properties of measures are discussed (we just hope that this book is free of such errors). The Russell - Rao index, to mention just an example, is often assigned to the group of metric measures. Its dissimilarity form cannot be metric, however, because self-similarity is zero if and only if d=0. That is, the first metric axiom is not satisfied. This list could be continued...

3.8.1 Computer programs

Large commercial packages usually include just a few, but well-known and very widely applicable coefficients. On the other hand, most specialized packages offer a much more exhaustive menu (Table 3.5). Ironically, such a wide selection of coefficients makes the life of a general user much more "difficult": choice among coefficients is not always easy.

The table cannot include all relevant program packages. There are, for example, several programs for evaluating sequences, and just a few of them are mentioned here. The University of Wisconsin Genetics Computer Group (Devereux et al. 1984) has developed the **Uni** package for aligning DNA and RNA sequences, and for computing the Jukes - Cantor distance. Nei's (1991) package computes distances between sequences and provides other tools for further analysis. But see Appendix B for more information.

Table 3.5. Availability of similarity and distance coefficients in selected program packages. Formulae not discussed in the book are not included in the table.

	BMDP 7	Statistica	NT-SYS	SYN-TAX	NuCoSA
simple matching coefficient	+	+	+	+	+
Rogers - Tanimoto			+	+	+
Anderberg I				+	
Anderberg II			+	+	
PHI	+		+	+	+
Yule II			+	+	+
Baroni-Urbani - Buser I				+	+
Baroni-Urbani - Buser II				+	+
Russell - Rao			+	+	+
Kulczynski (p/a)			+	+	+
Jaccard	+		+	+	+
Sorensen/Dice	+		+	+	+
Ochiai	+		+	+	+
Fager	+				
Spearman Rho					+
Kendall Tau				+	+
Jukes - Cantor			+		
Euclidean	+	+	+	+	+
Manhattan-metric		+		+	+
Minkowski general formula	+	+			
Average distance			+	+	
Mean character difference			+	+	
Canberra-metric			+	+	+
Canberra-metric/n				+	
chord distance				+	
angular separation	+		+	+	+
geodesic metric					
Pinkham - Pearson					+
Bray-Curtis/percentage diff.	+		+	+	
Marczewski-Steinhaus/Ruzicka				+	
Kulczynski					
chi-squared distance	+		+		+
cross- product				+	
covariance	+	+		+	
correlation	+	+	+	+	+
similarity ratio				+	+
Kendall/Renkonen			+		
Rogers			+		
Prevosti			+		
Nei			+		
Goodman-Kruskal gamma				+	
Horn				+	
Penrose size			+	+	
Penrose shape			+	+	
generalized distance				+	
distance in oblique systems					+
Gower formula				+	
Distance for mixed data				+	

Goodall et al. (1991) developed a package for calculating the probabilistic index and related measures (affinity index, deviant index, etc.). The Calhoun distance can be calculated using the BASIC program listed by Orlóci (1978). Ludwig & Reynolds (1988) provide a package in BASIC, containing most similarity and distance measures. FORTRAN lists of programs for computing information theory measures are found in Feoli et al. (1984).

Many formulae discussed in the book are excluded from the table, and I am not aware of programs which provide, for example, Gleason's and Ellenberg's coefficients. If required, it is better to write a small BASIC or Pascal routine, and the distance matrix thus obtained may be input to other programs, such as **SYN-TAX** (Podani 1997b) and **NuCoSA** (Tóthmérész 1996a).

3.9 Imaginary dialogue

Q: *I must confess that I am completely overwhelmed and exhausted by your enumeration of coefficients. When I finished reading this chapter, my head was a little confused and dazzled. Those names will not let me sleep well tonight.*

A: Yes, I admit that it was perhaps the most tiresome, though very important chapter of the book – but there was no escape! I was probably successful in illuminating the high methodological diversity of the subject. It is not by accident that many of the formulae introduced here were derived by biologists or statisticians involved in biological data analysis. And if you knew how many more can be found in the literature! As I seemed to suggest in the overview, the largest and least comprehensible literature of similarity and distance functions has to do with their application in biology.

Q: *It was a bit disturbing from the beginning that in one place you were talking about distances, in another you mentioned dissimilarities, or sometimes similarities. Now I understand their differences, I hope, but it would be nice to find a collective term to refer to all of these coefficients by a single word.*

A: I agree with you. In many cases it was uneasy to find the appropriate words, and sometimes even I got entangled in a terminological confusion. Do not forget that there is a collective term, the *"resemblance"*, first used probably by Orlóci (1972, 1978) for this purpose. Although its original meaning is similarity, the literature seems to accept the word with its extended meaning.

Q: *OK, but! Having received such an intensive training on these resemblance coefficients, I would be refreshed by some guidance as to the appropriate choice among them. The text, tables and figures do not help me enough to see under what circumstances will a given coefficient be useful to me!*

A: I am afraid that an absolutely unambiguous answer to your question does not exist. Nobody can assure you that in a given situation this and only this function can be recommended. It is you who makes the decision and, in order to be able to pass this important methodological step successfully, it is better to understand the meaning of each function, and you must see yourself their behaviour under controlled circumstances. Nevertheless, a very general guide is not too difficult to provide, following the style of Legendre & Legendre (1983) and Gower & Legendre (1986). I can present to you a concise "identification key" to most of the resemblance coefficients covered in my book, excluding those that were already declared to be applicable to specific problems only (e.g., niche overlap measures, genetic distances, etc.):

1a The variables are measured on different scales ....Gower (3.103), distance (3.104)
1b All variables are of the same type **2**
2a The variables are nominal (in the binary case, too, because coding is arbitrary) **3**
2b The measurement scale of variables is not nominal **7**
3a Simple ratios, mostly for comparing objects **4**
3b Coefficients measuring independence or predictability, suitable mostly to compare variables **5**
4a Variables causing agreements and disagreements are equally weighted.. simple matching coefficient (3.33)
4b Agreements are double-weighted ....Sokal - Sneath I (3.35)
4c Disagreements are double-weighted ....Rogers - Tanimoto (3.34)
5a Metric, measuring independence of variables ....Cramér (3.37)
5b Non-metric, measuring mutual predictability **6**
6a The data are binary ....Yule I (3.16)
6b The variables have several states ....Goodman - Kruskal lambda (3.39)
7a The variables are measured on the ordinal scale **8**
7b The variables are expressed on the interval and ratio scales (may be binary, too)....................**9**
8a Best used to compare variables, few ties allowed, large differences emphasized ....Spearman rho (3.43)
8b Variables and objects can be both compared, many ties tolerated, differences in ranks equally treated ....Kendall tau (3.44-45), Goodman - Kruskal gamma (3.46)
9a We have binary (presence/absence) data **10**
9b The scores are not binary **17**
10a The number of mutual absences is influential **11**
10b Mutual absences (*d*) are completely ignored **16**
11a Mutual absences are as important as mutual presences **12**
11b Mutual absences are less decisive than mutual presences **15**
12a Agreements and disagreements are equally weighted **13**
12b Agreements and disagreements are not equally important **14**
13a The coefficient is metric .... .simple matching coefficient (3.6), Euclidean distance (3.7), Anderberg I (3.12), PHI (3.15)
13b The coefficient is not a metric ....Yule I, II (3.16-17), Anderberg II (3.13)
14a Agreements are double-weighted ....Sokal - Sneath I (3.11)
14b Disagreements are double-weighted ....Rogers -Tanimoto (3.9)
15a The number of mutual absences (*d*) decreases similarity ....Russell - Rao (3.23)
15b Mutual absences have intermediate effect.....Baroni-Urbani - Buser I, II; Faith I, II (3.19-22)
16a The coefficient is a metric ....Jaccard (3.24), Ochiai (3.26)
16b Metric criteria not satisfied..................Sorensen (3.25), Kulczynski (3.29), Mounford (3.31)
17a Adding a constant to the data does not alter the result (only these apply to the interval scale, and they work in case of ratio scale as well, of course) **18**
17b Adding a constant to all data values influences the result (good only for the ratio scale, cannot be recommended for the interval scale) **21**
18a The formula contains implicit standardization **19**
18b No standardization implied **20**

19a Standardization by column and row totals ....chi^2 distance (3.67)

19b Standardization to unit dispersion ....correlation (3.70)

20a Differences between values matter ....Euclidean distance (3.47), Manhattan-metric (3.48)

20b Minimum agreements (intersections)
are accumulated ....Kendall (3.73), Renkonen (3.74)

21a Coefficients sensitive to ratios **22**

21b Coefficients sensitive to products of variables **23**

21c Coefficients sensitive to absolute differences among variables **24**

22a Measures chord length in unit hypersphere ....chord distance (3.54)

22b Measures arc length in unit hypersphere ....geodesic metric (3.56)

22c Measures angle between vectors ....angular separation (3.55)

23a Range is infinite ....cross product (3.68), covariance (3.69)

23b The similarity falls between 0 and 1 ....similarity ratio (3.71)

24a The agreements (or disagreements, for dissimilarities) of the objects are first summarized
by variables, then relativized by possible maximum for the given pair;
range between 0 and 1 ....**25**

24b Agreements between objects are relativized first, and
then summed up ....Canberra (3.52), Clark (3.57)

25a Differences between variables present in both objects
are ignored ....Gleason (3.64), Ellenberg (3.65)

25b Differences are always considered ....
..Bray - Curtis (3.58), Marczewski - Steinhaus (3.60), Kulczynski (3.62), Pandeya (3.66).

When you end up with a group of coefficients along the above key, then minor things become decisive. Choice in favour of a particular function is almost impossible without knowing the problem you are faced with, and it is therefore very useful to examine the behaviour of coefficients upon successive modifications of your data, as shown in this book. Of course, you may want to test other types of sequences, whichever is best suited to your problem. I strongly advise trying several coefficients with the same set of data and then comparing the results obtained. The novice can benefit much of such experiments, but... please do *not* use this comparative approach *a posteriori,* selecting the result that best reflects your intuitive ideas and expectations! That would be unfair to yourself as well as to the scientific community.

Q: *When I have made up my mind and picked up a coefficient, and my data are measured on the interval and ratio scales, I am still uncertain whether standardization of data before comparing the objects is meaningful or "illegal"! Can you advise?*

A: Yes, you just identified a weak point in my discussion! It is very true that, having chosen a similarity measure, some data standardization procedures can certainly be excluded. In other cases standardization is implied in the formula, as pointed out earlier. A compatibility matrix showing meaningful and invalid combinations of resemblance coefficients and standardization methods would hopefully be of great value for you (Table 3.6). The table calls attention to illogical or doubtful combinations, which can have unpredictable effects. I can guarantee you that the more specialized a coefficient, the less tolerant it is for data modifications. Be warned that a plus sign in the table does not imply that metric properties of the given function are al-

Table 3.6. Compatibility matrix of some distance functions and standardization methods. Abbreviations: + = acceptable combination, N = the standardization does not influence the result, E = inadmissible, not recommended for some reason (e.g., meaningless, leads to division by 0, etc.). Numbers refer to combinations that require some notes: (1) identical to the chord distance, (2) known as Whittaker-distance, (3) product moment correlation, (4) Renkonen index, (5) half of Manhattan metric standardized similarly, (6) agrees with the Manhattan metric standardized by the same way. Combining these with another standardization method is not recommended.

	By variables					By objects			
	Range	St. deviation	Total	Maximum	Norm	Range	Total	Maximum	Norm
Euclidean distance							2		1
Manhattan metric									
Canberra metric	E	?	N	N	N		E		
Clark	E	E	N	N	N		E		
Bray-Curtis				E				5	
Marczewski-Steinhaus						E			
Kulczynski				E				6	
Pinkham-Pearson	E	E	N	N	N		E		
Gleason						E			
Ellenberg						E			
Pandeya						E			
covariance						3			
similarity ratio						E			
Kendall				E			4		+

ways retained after standardization!

Q: *Is it really a serious disadvantage that my favourite coefficient is not Euclidean?*

A: Not necessarily, because non-Euclidean and even non-metric measures may often lead to distance matrices that can be perfectly represented in the Euclidean space. Some functions do not violate the metric axioms in actual situations; you have to fabricate unrealistic data in order to show the non-metric property. What is more interesting is that violation of conditions usually does not prevent interpretability of results...

Q: *How can I tell if a distance matrix has a corresponding Euclidean representation? How do you measure the extent to which the conditions are not met?*

A: This is a good question, but I am afraid it is a little early to give you a fully understandable answer. You can use principal coordinates analysis to see if it is possible to arrange the objects into some Euclidean space such that their distances are preserved (Subsection 7.4.1). If you get negative eigenvalues, then you can be sure that your coefficient is not Euclidean. For example, the eigenvalues of matrix (3.2), which is metric but not Euclidean, are 4.5, 4.5, 0.0 and –0.33. If the 1.6s of the matrix are replaced by 1.732, then you get 4.5 twice again, and two zeros. Why two? - you could ask. Because the four points fall onto the plane, but let me reserve the right to expand this topic later...

Q: *Your examples in the previous chapters were quite convincing that successive application of small changes to sampling or to the data provides a series, the study of which reveals much more on the properties of the objects than when a particular stage of the series were investigated only. I still remember quadrat size changes and the flexible parameter of Clymo's transformatiom. Can you define similar series for resemblance functions as well?*

A: I think you are simply provoking me to give an answer that you almost know already. Yes, there are such series pertaining to distances. The family of Minkowski metrics represents one example, even though only two stages of the series, the Manhattan metric and Euclidean distance are of practical importance, because higher powers overemphasize large differences. A series is formed by another equation, the "*intermediate coefficient*" (3.75) suggested by Faith, if modified as follows:

$$INT_{jk} = \Sigma\ [\alpha|x_{ij}–x_{ik}| + (1–\alpha)(\max_h[x_{ih}] – \min[x_{ij}, x_{ik}])] \quad 0 \leq a \leq 1 \qquad (3.118)$$

By changing the value of α, we can define a full continuum between two extremes, the Manhattan-metric (when $\alpha = 1$) and the dissimilarity version of the Kendall measure (when $\alpha = 0$). I can imagine a similar transitional formula for the Euclidean and the chord distance, to change balance between absolute and relative differences...

Q: *You admitted somewhere that some important functions may have been omitted inadvertendly in this chapter. There is at least one: I have heard of Pearson's contingency coefficient several times. If you have space, then I would be happy to see it again....*

A: The contingency coefficient is similar to Cramér's formula (3.37) in that the sample size dependence of the maximum of χ^2 is considered:

$$KK = \left(\frac{\chi^2}{f_{..} + \chi^2} \right)^{1/2} \qquad (3.119)$$

If we assume that the range of both variables being compared can be subdivided into many categories (*p* and *q* are large), and that the frequency distribution derived from many observations approaches the bivariate normal distribution, then *KK* squared will approximate the square of the correlation coefficient (3.7). This is of theoretical importance, however, because these criteria are rarely met (Anderberg 1973); this is why I did not mention this possibility earlier.

There is yet another measure resembling the Cramér index, in which relativization is achieved by the geometric mean of p–1 and q–1:

$$CS = \left(\frac{\chi^2 / f_{..}}{[\,(p-1)\,(q-1)\,]^{0,5}} \right)^{0.5} \qquad (3.120)$$

(Chuprow formula, cf. Anderberg 1973). The difference from the Cramér index increases as the gap between p and q increases.

Q: *You mentioned, without details, that biological interpretability is crucial in case of genetic distance measures. How is it elsewhere, for example, in ecology?*

A: On the analogy of genetic distances, we can talk about ecological or taxonomic distances. The fundamental question is whether a geometrically impressive and understandable distance measure is also interpretable in the context of the given research field. Let us remain in the domain of ecology. Imagine, for example, that somewhere in the temperate region we start from the coast and ascend in a coastal mountain, up to an elevation of 2500 m above sea level. On the coast, we have a species-poor salt marsh assemblage. On the other end of the gradient, in the alpine zone beyond 2000 m, the flora is just as well poor in species, whereas the montane vegetation, around 1000 m, is much more diverse than anywhere. If Euclidean distances are calculated from presence/absence data then we find that the alpine vegetation is closer to the coastal than to the montane, which is ecologically nonsensical. Geometric interpretabaility is not enough by itself, in addition we must always examine the biological meaning of our measures.

Q: *How could we enforce the different importance of our variables in constructing a resemblance function?*

A: What you mean is weighting variables, which is possible with many of our formulae. There have been many attempts to incorporate external or internal weighting in the distance functions. For example, Gordon (1990) proposed a method that relies on the judgment of an assessor in defining weights, an example of external weighting that is not derived from the data themselves. To illustrate the problem of data-derived weighting, we can declare that in presence/absence ecological data differences in common species convey much more information than differences in rare species (weighted dissimilarity index, Podani 1978):

$$WDI_{jk} = \frac{\sum_{i=1}^{n} p_i \,|\, x_{ij} - x_{ik} \,|}{\sum_{i=1}^{n} p_i} \qquad (3.121)$$

The weight, p_i, is the probability of the presence of species i as estimated from the sample. You can of course apply other weighting systems, such as the entropy of the species, which emphasizes species of intermeditate frequency in the data (Tóthmérész 1996b). If importance of species is determined from their relationships with the others, then you have two alternatives: you can assign higher importance to species that are highly associated with the others using the sum of chi-square values calculated for each species with all the others (MacNaughton-Smith 1965, Feoli & Lagonegro 1983) or to species that are most independent from the others, using the reciprocal value of that sum (Burnaby 1970). You can then compare how these weights are related to the generalized distance approach, for example.