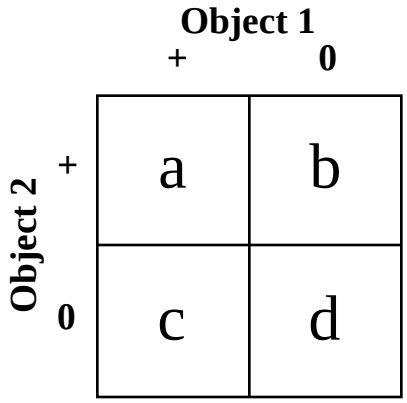
Some measures of resemblance

References: Sneath & Sokal (1973); Legendre & Legendre (1983, 1998)

1) Examples of binary similarity measures

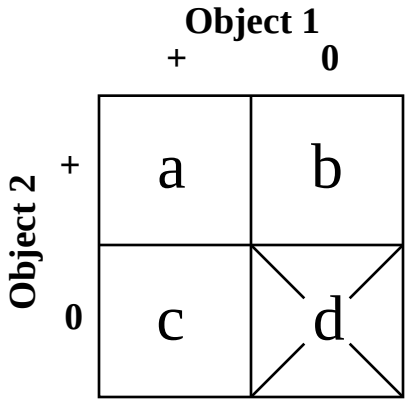
Coefficients including double-zeros (symmetrical coefficients)



Simple matching coefficient:

$$S(\mathbf{x}_1,\mathbf{x}_2) = \frac{a + d}{a + b + c + d}$$

Coefficients excluding double-zeros (asymmetrical coefficients)



Jaccard's coefficient of community:

$$S(\mathbf{x}_1,\mathbf{x}_2) = \frac{a}{a + b + c}$$

Example:	**Binary variables** (n = 7)							
Object $\mathbf{x}_1$	1	0	0	1	1	0	1	
Object $\mathbf{x}_2$	0	0	1	1	1	0	0	
Agreement		1		1	1	1		Sum = 4
Positive agreement				1	1			Sum = 2
Presence	1		1	1	1		1	Sum = 5

$S(\mathbf{x}_1,\mathbf{x}_2) = 4/7 = 0.57$

$S(\mathbf{x}_1,\mathbf{x}_2) = 2/5 = 0.40$

2) Examples of quantitative measures

Distance not designed for shared species
No upper limit

Euclidean distance (preserved in principal component analysis):

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2}$$

Distance designed for shared species
Bounded in the interval [0, 1]

Bray-Curtis distance:

$$D(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{2W}{A + B}$$

Example:	**Species abundances**					A	B	W
Object $\mathbf{x}_1$	7	3	4	5	1	20		
Object $\mathbf{x}_2$	2	4	7	6	3		22	
Minima	2	3	4	5	1			15
Differences	5	1	3	1	2			

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{5^2 + 1^2 + 3^2 + 1^2 + 2^2} = 6.325 \qquad D(\mathbf{x}_1, \mathbf{x}_2) = 1 - \frac{2 \times 15}{20 + 22} = 0.286$$

Distance designed for shared species
No upper limit

Chi-square (χ^2) distance (preserved in correspondence analysis):

$$D(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j} / y_{++}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

Example: $\left[y_{i+} \right]$

$$\mathbf{Y} = \begin{bmatrix} 7 & 3 & 4 & 5 & 1 \\ 2 & 4 & 7 & 6 & 3 \\ 12 & 8 & 5 & 14 & 6 \end{bmatrix} \begin{bmatrix} 20 \\ 22 \\ 45 \end{bmatrix} \rightarrow \left[\frac{y_{ij}}{y_{i+}} \right] = \begin{bmatrix} 0.350 & 0.150 & 0.200 & 0.250 & 0.050 \\ 0.091 & 0.182 & 0.318 & 0.273 & 0.136 \\ 0.267 & 0.178 & 0.111 & 0.311 & 0.133 \end{bmatrix}$$

$$\left[y_{+j} \right] = \begin{bmatrix} 21 & 15 & 16 & 25 & 10 \end{bmatrix} \quad 87$$

$D(\mathbf{x}_1, \mathbf{x}_2) = 0.653$

Compute the Gower distance

Gower coefficient: $S(\mathbf{x}_1,\mathbf{x}_2) = \sum_{j=1:p} s_j(x_1,x_2) \Big/ p$ p = number of variables

$D = 1 - S$

Partial similarity function: $s_j(x_1,x_2) = 1 - \frac{|y_{1j} - y_{2j}|}{R_j}$

Example –

	Variables						
	V1	V2	V3	V4	V5	V6	V7
Obj.1	2	2	2	2	4	2	6
Obj.2	1	3	1	2	2	2	5
[…]							
Obj.49	1	1	1	1	1	1	1
Obj.50	2	5	5	2	4	3	6
$\lvert y_{1j} - y_{2j} \rvert$	1	1	1	0	2	0	1
Range$_j$ in data matrix	1	4	4	1	3	2	5
$\lvert y_{1j} - y_{2j} \rvert / R_j$	1	0.25	0.25	0	0.667	0	0.20
$s_j = 1-(\lvert y_{1j}-y_{2j} \rvert / R_j)$	0	0.75	0.75	1	0.333	1	0.80

Sum(s_j) = 4.63333

p = 7

S($\mathbf{x}_1$, $\mathbf{x}_2$) = 4.63/7 = 0.66190

D($\mathbf{x}_1$, $\mathbf{x}_2$) = 1 – 0.66190 = 0.33810

Ecologists are, in principle, free to define and use any measure of association suitable for the ecological question under study; mathematics impose few constraints to this choice. This is why so many association coefficients are found in the literature. Some of them are of wide applicability whereas others have been developed to meet specific needs. Several coefficients have been rediscovered by successive authors and may be known under various names. Reviews of some coefficients can be found in Cole (1949, 1957), Goodman & Kruskal (1954, 1959, 1963), Dagnelie (1960), Sokal & Sneath (1963), Williams & Dale (1965), Cheetham & Hazel (1969), Sneath & Sokal (1973), Clifford & Stephenson (1975), Orlóci (1978), Daget (1976), Blanc *et al.* (1976), Prentice (1980), Gower (1985), and Gower & Legendre (1986).

1 — Similarity, distance, and dependence coefficients

In the following sections, *association* will be used as a general term to describe any measure or coefficient used to quantify the resemblance or difference between objects or descriptors, as proposed by Orlóci (1975). With *dependence* coefficients, used in the R mode, zero corresponds to no association. In Q-mode studies, *similarity* coefficients between objects will be distinguished from *distance* (or *dissimilarity*) coefficients. Similarities are *maximum* ($S = 1$) when the two objects are identical and *minimum* when the two objects are completely different; distances follow the opposite rule. Figure 7.2 shows the difference between the two types of measures: the length of the line between two objects is a measure of their distance, whereas its thickness, which decreases as the two objects get further apart, is proportional to their similarity. If needed, a similarity can be transformed into a distance, for example by computing its one-complement. For a similarity (S) measure, which takes values between 0 and 1, the corresponding distance (D) may be computed as:

Dependence
Similarity
Distance

$$D = 1 - S, \quad D = \sqrt{1 - S}, \quad \text{or} \quad D = \sqrt{1 - S^2}$$

We will see in Tables 7.2 and 7.3 and in Subsection 9.3.4 that the choice of a transformation instead of another may have consequences for the result of ordination analysis. Distances, which in several cases are not bound by a pre-determined upper value, can be normalized using eqs. 1.10 or 1.11:

$$D_{norm} = \frac{D}{D_{max}} \quad \text{or} \quad D_{norm} = \frac{D - D_{min}}{D_{max} - D_{min}}$$

where D_{norm} is the distance normalized in the interval [0, 1]; D_{max} and D_{min} are the maximum and minimum values taken by the distance coefficient, respectively. Normalized distances can be used to compute similarities by reversing the transformations given above:

$$S = 1 - D_{norm}, \quad S = 1 - D^2_{norm}, \quad \text{or} \quad S = \sqrt{1 - D^2_{norm}}$$

The following three sections describe the coefficients that are most useful with ecological data. Criteria to be used as guidelines for choosing a coefficient are