2. Association matrices and coefficients

2.1 Association matrices

A large majority of the methods of multivariate analysis, especially ordination and most clustering techniques, are explicitly or implicitly based on a **comparison of all possible pairs** of objects or descriptors.

When the pairs of **objects** are compared the analysis is said to be of **Q-mode**. When the pairs of **descriptors** are compared the analysis is said to be of **R-mode**.

This distinction is important because the comparison is based on **association coefficients**, and the coefficients of Q and R analyses are not the same.

In **Q-mode**, the coefficients measure the **distance** or the **similarity** between pairs of objects. Example: Euclidean distance, Jaccard similarity. In **R-mode**, one rather uses coefficients of **dependence** among variables, like for instance covariance or correlation.

Computing all possible comparisons among pairs of objects produces a **square and symmetrical association matrix**, of dimensions $n \times n$ (Q-mode) or $p \times p$ (R-mode):

Q-mode

$$A_{nn} = \begin{bmatrix} a_{11} & \cdots & a_{1i} & \cdots & a_{1n} \\ \vdots & \ddots & & & \\ a_{i1} & & a_{ii} & & a_{in} \\ & & & \ddots & \vdots \\ a_{n1} & & a_{ni} & & a_{nn} \end{bmatrix}$$

R-mode

$$A_{pp} = \begin{bmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1p} \\ \vdots & \ddots & & & \\ a_{j1} & & a_{jj} & & a_{jp} \\ & & & \ddots & \vdots \\ a_{p1} & & a_{pj} & & a_{pp} \end{bmatrix}$$

Figure 5 - Association matrices.

Every value in these matrices yields a comparison between two objects or descriptors whose location in the raw data matrix is given by the subscripts: a_{in} is the comparison measure between object *i* and object *n*. Ecological association matrices are usually symmetrical since $a_{in} = a_{ni}$. The values located on the diagonal compare the objects or variables with themselves. In Q-mode, the diagonal is usually made of 0 (when the coefficient is a distance) or 1 (when the coefficient is a similarity). In R-mode the diagonal gives a measure of dependence of a variable with itself: for instance this value equals 1 if the dependence measure is a Pearson correlation coefficient, or it equals the variance of the variable if the dependence measure is a covariance.

All the useful information of an association matrix is thus given in the **triangle** located above or below the diagonal (without the diagonal itself). The number of comparisons of all possible pairs of *n* objects is thus equal to $n(n–1)/2$.

2.2. Types of descriptors

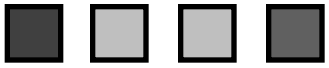
Before reviewing the available categories of coefficients of association, one must specify the mathematical type of variables to which these coefficients will be applied. Figure 6 (below) summarises these types in the form of a hierarchy of complexity starting with the binary type (the simplest one: 1-0, yes-no, present-absent, open-closed...) and proceeding to the continuous quantitative type. In data analysis, one can simplify the information at hand (e.g. recode species abundance data into presence-absence data), but usually not the reverse. Note that it often happens that the information required from an analysis can be obtained without the variables being measured with the maximum possible precision. Frequently, a large amount of objects characterised by measurements made with a limited precision is preferred over a small number of objects whose variables are measured with a very high precision.

Binary: 1 - 0 present - absent

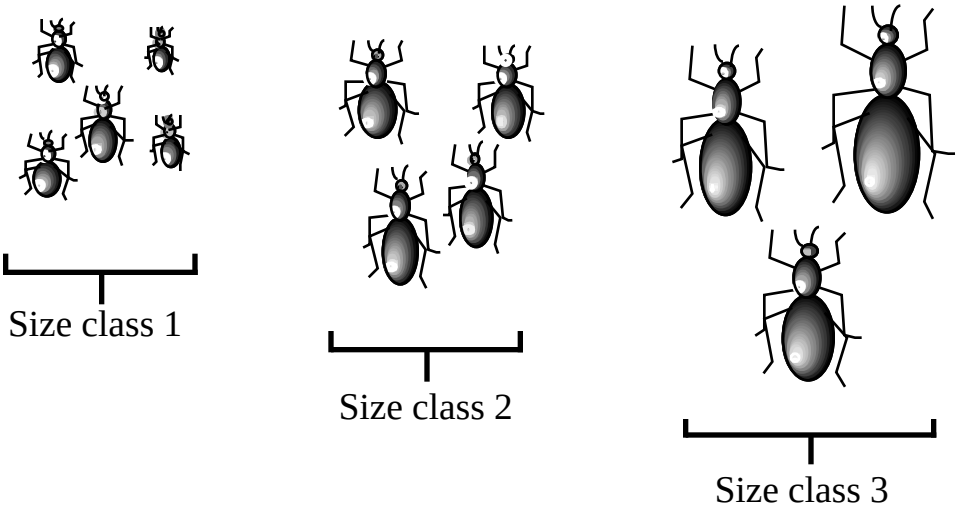
Descriptor "species" — Description "species present": 1 / Description "species absent": 0

	Spec. 1	Spec. 2	Spec. 3
Site 1	1	0	1
Site 2	1	1	0
Site 3	0	0	1

Multi-state: - **Unordered**, nominal : ex. colors, type of soil...



- **Ordered**: - **Semiquantitative**, ordinal, rank-ordered, : ex. size classes (0-10 cm, 10-50 cm, more than 50 cm...), rank in a race.



- **quantitative**: - **discontinuous**, meristic, discrete (ex.: number of persons in this room, nb. of individuals per species...

	Spec. 1	Spec. 2	Spec. 3
Site 1	12	0	18
Site 2	3	56	0
Site 3	0	0	1

- **continuous** (ex.: temperature, length, ...)

Figure 6 - Mathematical types of descriptors used in ecology.

2.3. The double-zero problem

In the following sections, the association coefficients will be grouped into categories depending on the type of objects or descriptors to which they are applied. Before this review, it is necessary to bring up a problem pertaining to the comparison of objects when a descriptor has the value "zero" in a pair of objects.

In certain cases, the zero value has the same meaning as any other value on the scale of the descriptor. For instance, the absence (0 mg/L) of dissolved oxygen in the deep layers of a lake is an ecologically meaningful information.

On the contrary, the zero value in a matrix of species abundances (or presence-absence) is much more tricky to interpret. The *presence* of a species at a given site generally implies that this site provides a set of minimal conditions allowing the species to survive (the dimensions of its ecological niche). The *absence* of a species from a relevé or site, however, may be due to a variety of causes: the species' niche may be occupied by a replacement species, or the absence of the species is due to adverse conditions on *any* of the important dimensions of its ecological niche, or the species has been missed because of a purely stochastic component of its spatial distribution, or the species does not show a regular distribution on the site under study. The key here is that the absence from two sites cannot readily be counted as an indication of resemblance between the two sites, because this double absence may be due to completely different reasons.

The information "presence" has thus a clearer interpretation than the information "absence". This is why one can distinguish two classes of association coefficients based on this problem: the coefficients that consider the double zero (sometimes also called "negative match") as a resemblance (as any other value) are said to be **symmetrical**, the others, **asymmetrical**. It is preferable to use asymmetrical coefficients when analysing species data.

The following sections review the main categories of coefficients with several examples. For a comprehensive review and keys to help choose the appropriate coefficient, see Legendre & Legendre (1998). All the indices listed in that book are available in the R package for Macintosh of Legendre, Casgrain and Vaudor at following web address: <http://www.bio.umontreal.ca/legendre/> (not to be confused with the R language!)

The choice of an appropriate coefficient is fundamental, because nearly all the subsequent analyses will be done on the resulting association matrix. Therefore, the structures revealed by the analyses will be those of the association matrix.

2.4. Q mode: resemblance between objects

The coefficients most frequently used for the comparison of objects are **similarity** or **distance** measures. Depending on the above-mentioned characteristics of the variables in the data table, these coefficients can be classified as follows (Figure 7):

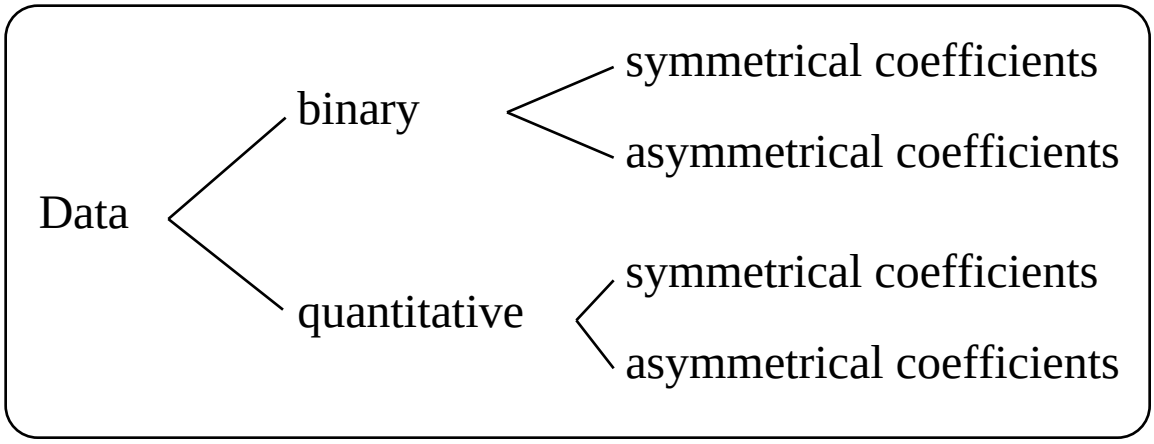


Figure 7 - Types of association coefficients in Q-mode analysis.

2.4.1. Symmetrical binary similarity coefficients

This expression means that these coefficients are made for binary data (and not that the values of the index itself are binary!) and that these coefficients treat a double zero in the same way as a double 1.

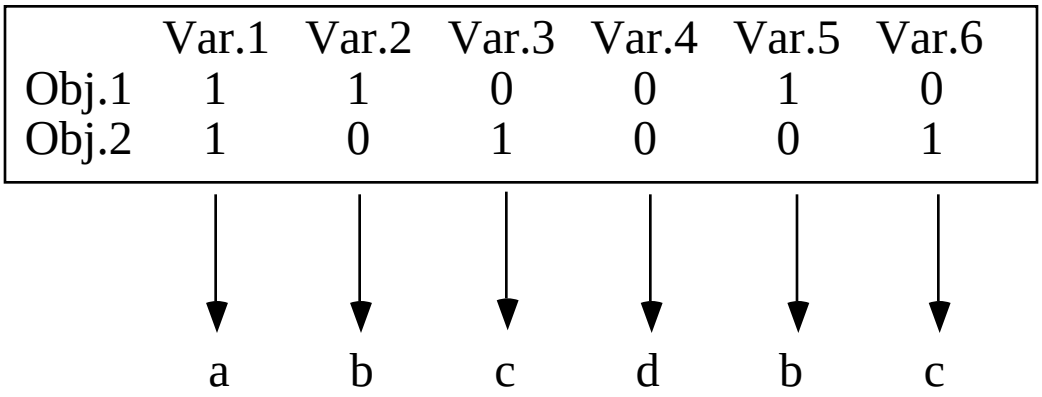
For binary coefficients, depending on the value taken by a variable in a pair of objects, one can represent the observations in a 2 × 2 contingency table (Figure 8):

Object 2 \ Object 1	1	0
1	a	c
0	b	d

Figure 8 - Contingency table showing the descriptors used to compare two objects: a = nb. of descriptors coding the two objects 1, d = nb. of descriptors coding the two objects 0, b and c = number of descriptors coding the two objects differently.

(a + b + c + d) is the total number of descriptors.

The most typical index of this category is the **simple matching coefficient** S_1 [the numbering of the coefficients is the one of Legendre & Legendre (1998)]. It is the number of variables that take the same value in both objects (i.e. double 1s + double 0s) divided by the total number of variables in the matrix. It is thus built as follows (Figure 9):



$$S_1 = \frac{a + d}{a + b + c + d}$$

Figure 9 - Computation of the simple matching coefficient

In this example, the simple matching coefficient is equal to:

$S_1 = (1+1)/(1+2+2+1) = 2/6 = 0.333$

which means that two of the six descriptors have the same value (0 or 1) in the two objects considered.

This coefficient, as well as the others of this category, are used to compare objects described by binary variables other than species presence-absence.

Note: a form of the simple matching coefficient exists, where the variables are multiclass nominal instead of "only" binary. The index is the number of descriptors having the same state in both objects, divided by the total number of descriptors.

2.4.2. Asymmetrical binary similarity coefficients

This category has the same role as the previous one, but for **presence-absence species data**. The formulas are the same as those of the category above, but ignore the d (double zero). The best known

coefficients of this category are the Jaccard community index (S_7) and the Sørensen index (S_8).

$$S_7 = \frac{a}{a+b+c}$$

$$S_8 = \frac{2a}{2a+b+c}$$

The use of these two coefficients is widespread in botany as well as zoology.

2.4.3. Symmetrical quantitative similarity coefficients

Coefficients of this category are interesting because they allow one to compare, within a single coefficient, descriptors of different mathematical types. The trick consists in computing partial similarities for each descriptor, and then to take the average of these partial similarities. Among the coefficients of this kind, let us mention Estabrook & Rogers (S_{16}) and Gower (S_{15}).

2.4.4. Asymmetrical quantitative similarity coefficients

This category, adapted to species abundance data, comprises among the most frequently used coefficients. Let us mention two of them: the Steinhaus index S_{17} (also well-known in its distance form as the Bray-Curtis index, D_{14}), and the χ^2 similarity S_{21}.

The S_{17} index (Figure 10) compares for each species the smallest abundance to the mean abundance in the two objects:

$$S_{17} = \frac{W}{(A+B)/2} = \frac{2W}{A+B}$$

Example:

	Species abundances					A	B	W
Object 1	70	3	4	5	1	83		
Object 2	64	4	7	4	3		82	
Minima	64	3	4	4	1			76

$$S_{17} = \frac{2 \times 76}{83 + 82} = 0.921$$

Figure 10 - Computation of the Steinhaus coefficient S_{17}.

A caveat about S_{17} is that, by construction, it gives the same importance to a difference of, say, 10 individuals, whether this means a difference between 1 and 11 individuals or between 301 and 311 individuals. This goes against intuition (and, in many cases, against ecological theory), and many users prefer to log-transform their data prior to an S_{17}-based analysis.

Another similarity measure adapted to species data, the χ^2 **similarity** is related to the χ^2 measure used to study contingency tables. The species abundances are transformed into profiles of conditional probability; thereafter one computes a weighted Euclidean distance matrix among sites, yielding the χ^2metric D_{15}. S_{21} is the reciprocal (S_{21} = 1–D_{15}) of this distance. The formula for D_{15} is given in Section 2.4.5.2 below.

2.4.5. Distance measures in Q-mode

2.4.5.1 Distance measures for qualitative binary or multiclass descriptors

All similarity coefficients can be converted into distances by one of the following formulas:

$$D = 1 - S \qquad D = \sqrt{1 - S^2}$$

$$D = \sqrt{1 - S} \qquad D = 1 - S/S_{\max}$$

These formulas provide appropriate conversions of *S* indices for qualitative binary or multiclass, and sometimes quantitative descriptors.

2.4.5.2 Distance measures for quantitative descriptors

Contrary to similarity coefficients, distances measures give a maximum value to two completely different objects, and a minimum value (0) to two identical objects. One can define three categories of indices depending on their geometrical properties:

• The *metrics*, which share the following four properties:

1. Minimum 0: if a = b then D(a,b) = 0
2. Positiveness: if a ≠ b then D(a,b) > 0
3. Symmetry: D(a,b) = D(b,a)
4. Triangle inequality: D(a,b) + D(b,c) ≥ D(a,c)

• The *semimetrics* (or *pseudometrics*), that do not follow the triangle inequality axiom. These measures cannot directly be used to order points in a metric or Euclidean space because, for three points (*a*, *b* and c*)*, the sum of the distances from *a* to *b* and from *b* to *c* may be smaller than the distance between *a* and *c*.

• The *nonmetrics*, a group of measures that can take negative values, thus violating the second principle above (positiveness).

Among the metric distance measures, the most obvious is the **Euclidean distance** (D_1). Every descriptor is considered as a dimension of a Euclidean space, the objects are positioned in this space according to the value taken by each descriptor, and the distance between two objects x_1 and x_2 is computed using **Pythagora's formula**:

$$D_1(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left(y_{1j} - y_{2j}\right)^2}$$

When there are only two descriptors, this expression becomes the measure of the hypotenuse of a right-angled triangle (Figure 11):

$$D_1(x_1, x_2) = \sqrt{\left(y_{11} - y_{21}\right)^2 + \left(y_{12} - y_{22}\right)^2}$$

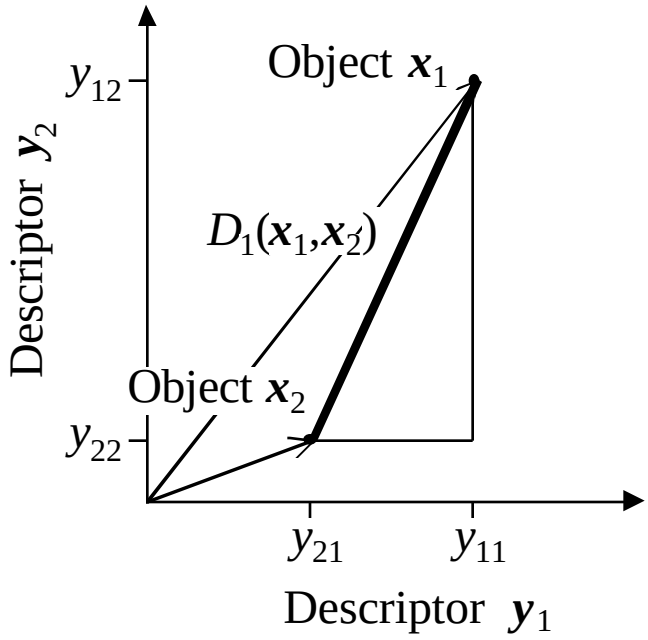


Figure 11 - Graphical representation of the Euclidean distanceD_1.

This measure has no upper limit; its value increases indefinitely with the number of descriptors, and, an important point, the value depends

on the *scale* of measurement of each descriptor. The problem may be avoided by computing the Euclidean distance on **standardized variables** instead of the original data. Standardization is not necessary when D_1 is applied to a group of dimensionally homogeneous variables.

For clustering purposes, the square of D_1 is sometimes used. ${D_1}^2$ is semimetric, however, making it less suitable for ordination.

D_1 is the essential ***linear*** measure! It is linked to the Euclidean space where a large majority of the usual statistical techniques are defined: regression, ANOVA... One consequence is that this measure is **not adapted to species data**: in Euclidean space, zero is a value like all others. Two objects with zero abundances of a given species will be as close to one another as if the species had, for instance, 50 individuals in each object, all other values being equal. Therefore, methods respecting the Euclidean distance among objects cannot generally be used on species data without proper adaptations. Some of these adaptations are *pre-transformations* of species data (see chapter 4: ordination); some adaptations can be imbedded into the Euclidean distance itself.

D_3, the **chord distance**, for instance, is a Euclidean distance computed on site vectors scaled to length 1 (=normalized vectors). It can be computed as D_1 after normalizing the site vectors to 1, or directly on the raw data through the following formula:

$$D_3(x_1, x_2) = \sqrt{2\left(1 - \frac{\sum_{j=1}^{p} y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^{p} y_{1j}^2 \sum_{j=1}^{p} y_{2j}^2}}\right)}$$

This trick provides a distance measure that is insensitive to the double zeros, making it suitable for species abundance data.

The chord distance is equivalent to the length of a chord joining two points within a segment of a sphere or hypersphere of radius 1. If only two descriptors are involved, the sphere becomes a circle and the chord distance can be represented as follows:

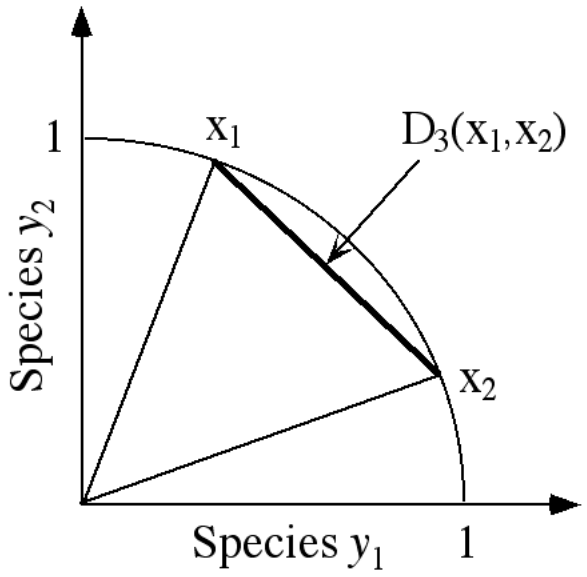


Figure 12 - Graphical representation of the chord distance D_3.

The chord distance is maximum when the species at the two sites are completely different (no common species). In this case, the normalised site vectors are at 90° from each other, and the distance between the two sites is √2. The chord distance is metric.

In Section 2.4.4, devoted to asymmetrical quantitative similarity coefficients, we mentioned S_{21}, the χ^2 similarity. This coefficient is actually the reciprocal of the χ^2 **metric** D_{15}. Its computation is done using following equation:

$$D_{15}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

where y_{+j} is the sum of abundances in all sites for species j, and y_{1+} and y_{2+} are the sums of species abundances in sites 1 and 2 respectively.

A related measure is called the χ^2 **distance** D_{16}, where all the terms of the sums of squares are divided by the *relative* frequency of each column in the overall table instead of the absolute frequency. In other words, it is identical to the χ^2 metric multiplied by $\sqrt{y_{++}}$, where y_{++} is the grand total of the data table:

$$D_{16}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}/y_{++}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2} = \sqrt{y_{++}} \sqrt{\sum_{j=1}^{p} \frac{1}{y_{+j}} \left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}} \right)^2}$$

The χ^2 distance is the distance preserved in correspondence analysis (CA, chapter 4). This measure has no upper limit.

A coefficient related to D_{15} and D_{16} is the **Hellinger distance** D_{17}, for which the formula is:

$$D_{17}(x_1, x_2) = \sqrt{\sum_{j=1}^{p} \left(\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right)^2}$$

We shall mention interesting uses of this distance measure, as well as of the chord distance D_3, in Chapter 4.

Finally, among the semimetric distance measures, the most frequently used is D_{14}, the Bray and Curtis distance, which is the reciprocal of the Steinhaus similarity coefficient: $D_{14} = 1 - S_{17}$, and is therefore adapted to species data.

2.5 R mode: coefficients of dependence

When one compares descriptors on the basis of their values in a series of objects, one generally wants to describe the way **these descriptors co-vary**. Once again we have to distinguish between the case where the descriptors are species abundances and the other cases.

*2.5.1 Descriptors **other** than species abundances*

Qualitative descriptors: their comparison can be done using two-way contingency tables and their χ^2 statistic.

Semi-quantitative descriptors: if a pair of such descriptors is in monotonic relationship, its resemblance can be measured using Spearman's ρ and Kendall's τ nonparametric correlation coefficients. If the relationship is not expected to be monotonic, then it may be preferable to use the χ^2 statistic for contingency tables. The semi-quantitative information is lost, but the relationship can be detected.

Quantitative descriptors: their relationship is generally measured by common parametric dependence measures like covariance or Pearson's correlation. Remember also that Pearson's correlation is equal to covariance measured on standardized variables. Note that covariance and correlation are only adapted to descriptors whose relationships are *linear*.

2.5.2 Species abundances: biological associations

Analyzing species abundances in R mode causes the same problem as in Q mode: what to do with double zeros?

Double absences are frequent in ecological communities because these contain many rare species and only a few dominant ones. Since one generally wants to define biological associations on the basis of all (or most) species present, the data matrix contains a large number of zeros. However, we know that the zeros do not have a nonequivocal interpretation. Therefore, it is not recommended to use the covariance or correlation coefficients mentioned above (including the nonparametric ones!), since these use the zero as any other value. Furthermore, correlation or covariance coefficients measure linear relationships, so that species that are always found together but whose abundances are not in linear relationship would not be recognised as belonging to the same association by these coefficients. The same

holds for nonparametric correlation coefficients, that detect monotonic relationships only.

If one has only access to such coefficients, several options are available to minimize their adverse effects:

- eliminate from the study the less frequent species, so as to reduce the number of double zeros;

- eliminate the zeros (by declaring them as missing values);

- eliminate double zeros only from the computation of the correlation or covariance matrix; this must generally be programmed separately;

Another method is to use the S_{21} coefficient among variables (species): as an exception, this coefficient can be applied in R mode as well as in Q mode.

Yet another approach is to apply Goodall's probabilistic coefficient (S_{23}) to species. This allows one to set an "objective", probabilistic limit to associations, such as: "all species that are related at a probability level $p \geq 0.95$ are members of the association".

Alternately, one can also define species groups by clustering the species scores of an ordination.

Presence-absence data: in several instances it may be preferable to define species associations on the basis of presence-absence data, for instance in cases where quantitative data do not reflect the true proportions among species (because of sampling biases, identification problems, and so on). Biological associations are then defined on the basis of the *co-occurrence of species* instead of the relationships between fluctuations in abundances. In this case there is another exception to the rule that Q-mode coefficients cannot be used in R mode: the Jaccard community coefficient S_7 or the Sørensen coefficient S_8 can be applied to species vectors (in R mode). Otherwise, Fager's coefficient (S_{24}) or Krylov's probabilistic coefficient (S_{25}) can be used. See Legendre & Legendre (1998) for more details.

Legendre (2005)[1] proposed to use Kendall's W coefficient of concordance, together with permutation tests, to identify species associations: "An overall test of independence of all species is first carried out. If the null hypothesis is rejected, one looks for groups of correlated species and, within each group, tests the contribution of each species to the overall statistic, using a permutation test." The simulations accompanying the paper show that "when the number of judges [= species] is small, which is the case in most real-life applications of Kendall's test of concordance, the classical χ^2 test is overly conservative, whereas the permutation test has correct Type I error; power of the permutation test is thus also higher."

Permutation tests are addressed in Chapter 5.

2.6 Choice of a coefficient

Legendre & Legendre (1998) p. 299-301 provide tables to help choose an appropriate similarity, distance or dependence coefficient. These tables are extremely helpful because of the many criteria to consider and the vast number of available coefficients.

[1] Legendre, P. 2005. Species Associations: The Kendall Coefficient of Concordance Revisited. Journal of Agricultural, Biological, and Environmental Statistics 10 (2): 226–245.