Welcome to Paradise
A quantitative measure of resemblance (i.e., similarity) between either objects (e.g., sites) and the variables describing them (e.g., species) can be an end in itself, or the standard precursor to subsequent ordination and classification procedures.

Association between objects: Q-mode analysis
Association between descriptors: R-mode analysis

Often times, an examination of the association hemi-matrix (derived from primary matrix) suffices to elucidate the basic structure of the data, and no additional analysis is needed.

Modes of Analysis
Catell (1952, 1966) was the first to recognize that EEB data could be studied from at least 6 viewpoints from 3-D data matrices composed of descriptors, objects, and times. He defined 6 possible "modes" of analysis.

The two main viewpoints used in EEB are R-mode or Q-mode.

The important point here is that these two modes of analysis are based on different measures of association.

Whether a particular analysis is R- or Q-mode is often confused in the literature & textbooks due to a lack of clarity.
Q- and R-Mode

In order to prevent confusion, any study starting with the computation of an association matrix among objects should be called a Q-mode analysis.

Any study starting with the computation of an association matrix among descriptors should be called an R-mode analysis.

Analytical Space

Following the terminology of Williams & Dale (1965), the space of descriptors (attributes) will be called "A-space". In this space, objects may be represented along axes that correspond to the descriptors.

Symmetrically, the space of reference in which the descriptors are positioned relative to axes corresponding to objects (or individuals) is called "I-space".

A-Space

- Example of 5 Objects with 2 Descriptors -

NB: The thickness of the lines joining objects is proportional to their degree of resemblance based on the two descriptors.
Analytical Space

The number of dimensions that can be represented on paper is obviously limited to two or three. However, we will see soon, that this simple analog can be extended to multiple dimensions.

The A- and I-spaces are called metric or Euclidean because the reference axes are quantitative and metric.

Ordination procedures by definition are restricted to these spaces, clustering procedures are not.

Ecological Resemblance

- Mode of analysis
- Analytical spaces
- Association (Resemblance) Coefficients
  - Q-mode similarity coefficients
    - Symmetrical binary coefficients
    - Asymmetrical binary coefficients
    - Symmetrical quantitative coefficients
    - Asymmetrical quantitative coefficients
    - Probabilistic coefficients
  - Q-mode distance coefficients
    - Metric distance
    - Semimetrics
  - R-mode coefficients of dependence
    - Non-abundance measures
    - Species abundance measures
- Choice of a coefficient
- ANOSIM

Association Coefficients

The most usual approach to assess the resemblance among objects or descriptors is to start with a rectangular data matrix and condense all of the information in to a square hemi-matrix of association values.

The structure resulting from the numerical association analysis may not necessarily reflect all of the information that was originally contained in the primary matrix.

Thus, there is considerable importance to choosing the appropriate measure of association.
Selection of Association Coefficient

The basic considerations for selecting an appropriate association coefficient fall under the following:

1. **nature of the study** determines the structure of the data to be evaluated with an association matrix,

2. the various measures available are subject to different **mathematical constraints** (depends upon whether one continues with ordination or clustering),

3. **computational** aspects such as what measures are available or can be programmed in particular software (less of a problem in recent years).

Selection of Association Coefficient

Because there are few mathematical constraints, biologists are free to define and use any measure of association suitable to the phenomenon under study—hence why there are so many coefficients in the literature.

We will use the general term "association coefficient" to describe any measure used to quantify resemblance; more specifically,

- R-mode studies generally use **dependence coefficients**.
- Q-mode studies typically use **similarity coefficients**.

**Q-Mode: Similarity Coefficients**

This is the largest group of coefficients in the literature. All of these coefficients are used to measure the association between objects.

Similarity measures are never metric since it is always possible to find two objects, A & B, that are more similar than the sum of their similarities with another more distant object C. Thus, similarities can NOT be used to position objects in metric space (they must be converted to distances) such as ordinations; however, they can be used in clustering analysis.
Similarity Coefficients

Similarity coefficients were first developed for binary (presence/absence) data and later became generalized for multi-state descriptors with the advent of computers.

NB: Similarity coefficients for binary data are used widely with current molecular data!

A major dichotomy in these descriptors exists regarding how the coefficient handles the double-negative or double-zero situation.

Double-Negative Problem

From classical niche theory, we know that species are distributed unimodally along environmental gradients. The abundance of a species reaches an optimum at some central set of conditions and is minimized near the minimum and maximum of the gradient sector.

The double-zero situation is a problem in this context because if a species is present at two sites, this is an indication of similarity of these sites. However, if a species is absent from both sites, it may be because both sites produce an environment that is above the optimal niche value, below the optimal, or one above and one below. One cannot tell.

Asymmetrical vs. Symmetrical Coefficients

Thus, it is [generally] preferable to abstain from drawing biological conclusions from double-negative situations (except under those conditions that permit accurate interpretation).

Numerically, this means that you should skip double zeros when computing similarity or distance coefficients when using binary data. Coefficients of this type are called asymmetrical because they treat zeroes in a different way than the rest of the data.

With symmetrical coefficients, the zero state for two objects is treated exactly the same way as for all other pairs of values, when computing similarity...
**Binary Coefficients**

In the simplest case, the similarity between two sites is based on presence-absence (binary) data. Observations are often summarized in a $2 \times 2$ table:

<table>
<thead>
<tr>
<th>Object $x_1$</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>0</td>
<td>$c$</td>
<td>$d$</td>
</tr>
</tbody>
</table>

Object $x_2$:

<table>
<thead>
<tr>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a$</td>
<td>$b$</td>
</tr>
<tr>
<td>$c$</td>
<td>$d$</td>
</tr>
</tbody>
</table>

Where:
- $a$ = the no. of descriptors for which the two objects are coded as 1
- $d$ = two objects coded as 0
- $b$ & $c$ = two objects coded differently
- $p = \text{the sum of all descriptors}$

An obvious way to compute the similarity between two objects is to count the number of descriptors that code the objects in the same way and divide this by the total:

$$S_{ij} = \frac{a + d}{p}$$

Coefficient $S_{ij}$ is called the simple matching coefficient (Sokal and Michener 1958). When using the coefficient, one assumes that there is no difference between double-0 or double-1.

Rogers and Tanimoto (1960) proposed a variant that gives more weight to differences:

$$S_{ij} = \frac{a + d}{a + 2c + 2d}$$

**Simple Matching Coefficients**

An obvious way to compute the similarity between two objects is to count the number of descriptors that code the objects in the same way and divide this by the total:

$$S_{ij} = \frac{a + d}{p}$$

Coefficient $S_{ij}$ is called the simple matching coefficient (Sokal and Michener 1958). When using the coefficient, one assumes that there is no difference between double-0 or double-1.

Rogers and Tanimoto (1960) proposed a variant that gives more weight to differences:

$$S_{ij} = \frac{a + d}{a + 2c + 2d}$$

**Simple Matching Coefficients**

Sokal & Sneath (1963), Sneath & Sokal (1973)

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>$\frac{2a + 2d}{2a + b + c + 2d}$</td>
<td>counts resemblances as being twice as important as differences</td>
</tr>
<tr>
<td>$S_2$</td>
<td>$\frac{a + d}{b + c}$</td>
<td>compares resemblances to differences, measure: $0$ to $\infty$</td>
</tr>
<tr>
<td>$S_3$</td>
<td>$1 + \frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}$</td>
<td>compares resemblances to marginal totals</td>
</tr>
<tr>
<td>$S_4$</td>
<td>$\frac{a}{\sqrt{(a+b)(a+c)}} \cdot \frac{d}{\sqrt{(b+d)(c+d)}}$</td>
<td>product of the geometric means of the terms relative to $a$ &amp; $d$ in coefficient $S_5$</td>
</tr>
</tbody>
</table>
Simple Matching Coefficients

Coefficients $S_1$ & $S_3$ have generally been the most popular; however, there may be times where the others are appropriate. Three additional measures are available in the NT-SYS computer software (popular among systematists):

\[
S = \frac{a + d - b - c}{p} \quad \text{Hamann coefficient}
\]

\[
S = \frac{ad - bc}{ad + bc} \quad \text{Yule coefficient}
\]

\[
\phi = \frac{ad - bc}{\sqrt{(a + b)(c + d)(a + c)(b + d)}} \quad \text{Pearson’s phi}
\]

MVSP is a popular software package that computes a large number (20) of similarity coefficients. http://www.kovcomp.co.uk/mvsp/

Similarity Coefficients - MVSP Software -

Consider the example of 5 Panamanian cockroach species scored for the presence/absence (0/1) in 6 habitats:
Simple Matching Coefficient
- Similarity Hemi-matrix -

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

NB: diagonal = 1

Simple Matching Coefficients
- MVSP Software -

So what formula does MVSP use to calculate the simple matching coefficient? \( SM_{ij} = S_j \) (Sokal & Michener 1958)

\[
SM_{ij} = \frac{a + d}{a + b + c + d}
\]

Confirm for yourself by comparing two samples:

BC1 vs. LC

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\[ SM = \frac{4}{5} = 0.800 \]
Use a spreadsheet and create a file named roach.csv

If you look around in R, you will find that it does not support an option for a simple matching coefficient like $S_1$.

Q: What to do?
A: Write it yourself slacker!

With a bit of poking around, you will see that there is a special function in the vegan package called designdist. This function is designed to use the abcd notation that we have developed thus far and write any coefficient you wish! Virtually all of the coefficients of Legendre & Legendre (1998) can be written using designdist.
Similarity Coefficients

- R -
Asymmetrical Binary Coefficients

Coefficients paralleling the ones just presented are available for comparing sites using species presence-absence data, where the comparison must exclude double-zeroes.

The best known measure is Jaccard’s (1900) coefficient of community, or more simply Jaccard’s coefficient:

$$ S_J(x, y) = \frac{a}{a + b + c} $$

Sørenson (1948) made an important modification by giving double weight to double presences. Sørenson’s coefficient:

$$ S_S(x, y) = \frac{2a}{2a + b + c} $$

Asymmetrical Binary Coefficients

Another variant of $S_S$ gives triple weight to double presences:

$$ S_D(x, y) = \frac{3a}{3a + b + c} $$

The weights seem to be most important when dealing with rare species. You may wish to explore response patterns over 2 or more weightings.

The asymmetrical analog to $S_S$ provides double weight to differences in the denominator:

$$ S_A(x, y) = \frac{a + d}{a + 2b + 2c} $$

Asymmetrical Binary Coefficients

Russell & Rao (1940) suggested a measure that allows the comparison of the number of double presences, in the numerator, to the total number of species found at all sites, including species that are absent ($d$) from the pairs of sites considered:

$$ S_{11}(x, y) = \frac{a}{p} $$
Asymmetrical Binary Coefficients

While Kulczynski (1928) proposed a coefficient opposing double-presences to differences:

\[ S_{k}(x, y) = \frac{d}{b + c} \]

Sokal and Sneath (1963) provide a modification of Kulczynski’s index where double-presences are compared to the marginal totals \((a + b)\) and \((a + c)\):

\[ S_{a}(x, y) = \frac{1}{2} \left[ \frac{d}{a + b} + \frac{d}{a + c} \right] \]

Asymmetrical Binary Coefficients

Ochiai (1957) used, as a measure of similarity, the geometric mean of the ratios of \(a\) to the number of species in each site, i.e., the marginal totals \((a + b)\) and \((a + c)\):

\[ S_{o}(x, y) = \frac{a}{\sqrt{(a + b)(a + c)}} \]

\(S_{o}\) is the same as \(S_{a}\) except for the part re double-zeros.

Faith (1983) suggested a coefficient in which disagreements (0/1s) are given a weight opposite that of double presences:

\[ S_{d}(x, y) = \frac{a + d/2}{b + c} \]

Symmetrical Quantitative Coefficients

EEB descriptors often have more than two states. The binary coefficients we just discussed can frequently be extended to accommodate multi-state descriptors. For example, the simple matching coefficient may be used as follows:

\[ S_{p}(x, y) = \frac{\text{agreements}}{p} \]

where the numerator contains the number of descriptors for which the two objects are in the same state.
Symmetrical Quantitative Coefficients

- Example: Simple Matching -

For example, if a pair of objects was described by the following 10 multi-state descriptors:

<table>
<thead>
<tr>
<th></th>
<th>Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object $x_1$</td>
<td>9 7 3 4 9 5 4 0 6</td>
</tr>
<tr>
<td>Object $x_2$</td>
<td>2 3 2 1 2 9 3 2 0 6</td>
</tr>
</tbody>
</table>

Agreements: 0 1 0 0 1 0 0 1 1

$S_{p}(x_1, x_2) = \frac{\text{agreements}}{p} = \frac{4}{10} = 0.4$

Symmetrical Quantitative Coefficients

In a similar fashion, it is possible to extend virtually all of the binary coefficients to create multi-state coefficients.

However, coefficients of this type often result in the loss of valuable information, especially in the case of ordered descriptors for which two objects can be compared on the basis of the amount of difference between states.

Gower’s Coefficient

Gower (1971) proposed a general coefficient of similarity which can combine different types of descriptors and process each according to its own mathematical type.

This coefficient can be VERY useful when you need to record data on binary, multi-state, and even quantitative variables all in the same primary matrix!

Gower’s coefficient takes the general form:

$S_{g}(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} h_{ij}$
The similarity between two objects is the average, over the $p$ descriptors, of the similarities calculated for all descriptors.

For each descriptor $j$, the partial similarity value $s_{12j}$ between objects $x_1$ and $x_2$ is computed as follows:

For binary descriptors, $s_{12j}=1$ (agreement) or 0 (disagreement); double-zeros are treated as agreement.

Qualitative and semi-quantitative descriptors are treated following the simple matching coefficient rule above.

Quantitative descriptors (real numbers) are treated in an interesting way...

Gower’s Coefficient

- Quantitative Descriptors -

For each descriptor, one first computes the difference between the states of the two objects $|y_{1j} - y_{2j}|$.

This value is then divided by the largest difference ($R_j$) found for the descriptor across all sites of the study—or if one prefers, in a reference population.

Since the ratio is actually a normalized distance, it is subtracted from 1 to transform it into a partial similarity:

$$s_{12j} = 1 - \left( \frac{|y_{1j} - y_{2j}|}{R_j} \right)$$

Gower’s Coefficient

Gower’s coefficient may be programmed to include an additional element of flexibility: no comparison is computed for descriptors where information is missing for one or the other object.

This is obtained by a value $w_j$, called Kronecker’s delta, describing the presence or absence of information: $w_j = 0$ when the information about $y_j$ is missing one or the other object, or both; $w_j = 1$ when info is present for both objects.

Thus, the final form of Gower’s coefficient is:

$$S_{12}(x_1, x_2) = \frac{\sum w_{1j} x_{1j}}{\sum w_{1j}}$$

(ranges from 0 to 1)
Gower’s Coefficient - Worked Example -

Two sites, eight quantitative descriptors of the environment.

<table>
<thead>
<tr>
<th>Descriptors $j$</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object $x_1$</td>
<td>1 3 3 1 2 2 5</td>
</tr>
<tr>
<td>Object $x_2$</td>
<td>1 4 2 4 1 3 2 5</td>
</tr>
<tr>
<td>$w_{ij}$</td>
<td>1 1 0 1 1 1 1 1</td>
</tr>
<tr>
<td>$R_j$</td>
<td>1 0.25 0 0.75 1 0.33 1 0.80</td>
</tr>
<tr>
<td>$y_{ij}-y_{2j}$</td>
<td>1 1 0 2 0 1</td>
</tr>
<tr>
<td>$</td>
<td>y_{ij}-y_{2j}</td>
</tr>
<tr>
<td>$w_{ij}/R_j$</td>
<td>0 0.75 0 0.75 1 0.33 1 0.80</td>
</tr>
</tbody>
</table>

Thus, $S_d(x_1, x_2) = 4.63/7 = 0.6614$

(NB: $R_j$, range of values among all objects for each $y_j$, was pre-calculated.)

Gower’s Coefficient - Using R -

Use a spreadsheet to create a comma-delimited data file (CSV) containing the 8 observations for each of 2 objects.

In this case, I used CALC in openoffice.org and saved the files as: gower.csv

Gower’s Coefficient - Using R (Gower Script File) -
So, R gives us 0.75, but hand calculation gives 0.66. WHY?

Modification

Estabrook and Rogers (1966) made a modification \( S_{16} \) of the Gower coefficient. The general equation is the same as \( S_{15} \), but differs in the computation of the partial similarities \( s_j \).

Here the partial similarity between two objects for a given descriptor \( j \) is computed using a monotonically decreasing function of partial similarity. The following function can be used for two numbers \( d \) and \( k \):

\[
\begin{align*}
    s_{ij} &= f(d_{ij}, k_j) = \frac{2(k+1-d)}{2k + 2 + d} \\
    &\quad \text{when } d \leq k \\
    s_{ij} &= f(d_{ij}, k_j) = 0 \\
    &\quad \text{when } d > k
\end{align*}
\]
Modification

For the two values \( d \) and \( k \), \( d \) is the distance between the states of the two objects \( x_1 \) and \( x_2 \) for descriptor \( j \) (i.e., the same value \( |x_1^j x_2^j| \) as in Gower’s coefficient) and \( k \) is a parameter determined \textit{a priori} by the users for each descriptor.

Parameter \( k \) is equal to the largest difference \( d \) for which the partial similarity \( s_{12}^j \) (for descriptor \( j \)) is allowed to be different from 0.

Values of \( k \) for the various descriptors may be quite different from each other. For example, for a descriptor coded 1 to 4, one might use \( k = 1 \); for a descriptor coded 1 to 50, perhaps \( k = 10 \) could be used.

Similarity Calculation \( S_{16} \)

- Modification Example -

<table>
<thead>
<tr>
<th>Descriptors ( j )</th>
<th>( S_{16}(x_1, x_2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Object ( x_1 )</td>
<td>2 1 3 4 2 1</td>
</tr>
<tr>
<td>Object ( x_2 )</td>
<td>2 2 4 3 2 3</td>
</tr>
<tr>
<td>( k_j )</td>
<td>1 0 1 2 1 1</td>
</tr>
</tbody>
</table>

\( s_{12}^j = f(d_{12}^j, k_j) \)

\( \begin{align*} &1.0+ \quad 0+ \quad 0.4+ \quad 0.5+ \quad 1.0+ \quad 0 \quad 2.9 / 6 \quad 0.483 \end{align*} \)

Possible values of \( k \) are used for example. NB: if \( k = 0 \) for all descriptors, \( S_{16} \) is identical to the SM coefficient.

Asymmetrical Quantitative Coefficients

Just as we did in the previous section, we extend asymmetrical binary coefficients to accommodate multi-state descriptors.

For example, Jaccard’s coefficient becomes:

\[
S_{j}(x_1, x_2) = \frac{\text{agreements}}{p - \text{double-zeros}}
\]

Where the numerator is the number of species with the same abundance state at the two sites. This coefficient is useful when species abundances are coded to a small number of classes and you wish to strongly contrast the abundances.
Some coefficients lessen the effect of the largest differences and may therefore be used with raw species abundances.

The best known coefficient here is one frequently referred to as the Bray-Curtis coefficient. The origin is a bit unclear as the coefficient has been “discovered” numerous times in the literature.

This coefficient compares two sites \((x_1, x_2)\) in terms of the minimum abundance of each species:

\[
S_{BC}(x_1, x_2) = \frac{2W}{(A + B)/2}
\]

In this formula, \(W\) is the sum of the minimum abundances of the various species; \(A\) & \(B\) are the sums of the abundances of all species at each of the two sites. For example:

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site 1</td>
<td>7</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>Site 2</td>
<td>2</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>Minimum</td>
<td>2</td>
<td>3</td>
<td>11</td>
</tr>
</tbody>
</table>

\[
S_{BC}(x_1, x_2) = \frac{2W}{(A + B)/2} = \frac{2 \times 11}{16 + 22} = 0.579
\]

Obviously, Excel can be used to calculate all of these similarity measures (many are not available in pre-packaged statistical software apps).
Building Excel Macros
For real data sets, it is usually much easier to write a macro

May need to edit VB macro
Kulczynski Coefficient

Kulczynski’s coefficient (1928) also belongs to this group of measures that are suited to raw abundance data. The sum of the minima is first compared to the grand total at each site; then the two values are averaged:

$$S_{K}(x, x') = \frac{1}{2} \left( \frac{W}{W + W} \right)$$

For binary data, $S_{K}$ becomes $S_{13}$. For the numerical example just completed (BC):

$$S_{K}(x, x') = \frac{1}{2} \left( \frac{1 + 1 + 1 + 1 + 1}{1 + 1 + 1 + 1 + 1} \right) = 0.594$$

Normalized Data

Often, species distributed across an ecological gradient are strongly skewed.

Several coefficients have been adapted to handle “normalized” abundance data.

Data can be normalized through transformation or scaling (e.g., 0-7, rare-abundant).

For example, Gower’s coefficient ($S_{15}$) can be easily adapted:

Normalized Data

$$S_{15}(x, x') = \sum_{j=1}^{p} \frac{w_{1j} \cdot s_{1j}}{\sum_{j=1}^{p} w_{1j}}$$

where

$$s_{1j} = 1 - \frac{|y_{1j} - y_{2j}|}{R_{j}}$$

as in $S_{15}$,

and $w_{1j} = 0$

when $y_{1j}$ or $y_{2j}$ = absence of information, or

when $y_{1j}$ or $y_{2j}$ = absence of species ($y_{1j}$ or $y_{2j} = 0$),

while $w_{1j} = 1$ in all other cases.

Modified Gower’s Coefficient

$$S_{15}(x, x') = \frac{\sum_{j=1}^{p} w_{1j} \cdot s_{1j}}{\sum_{j=1}^{p} w_{1j}}$$

where

$$s_{1j} = 1 - \frac{|y_{1j} - y_{2j}|}{R_{j}}$$

as in $S_{15}$,

and $w_{1j} = 0$

when $y_{1j}$ or $y_{2j}$ = absence of information, or

when $y_{1j}$ or $y_{2j}$ = absence of species ($y_{1j}$ or $y_{2j} = 0$),

while $w_{1j} = 1$ in all other cases.
χ² Similarity

The last quantitative coefficient that excludes double-zeros is called χ² Similarity.

It is the complement of the chi-square metric ($D_{13}$) discussed subsequently (and we will defer a fuller discussion of how to apply this measure until then).

$$S_{21}(x_1, x_2) = 1 - D_{13}(x_1, x_2)$$

Ecological Resemblance

- Outline -

- Ecological Resemblance
- Mode of analysis
- Analytical spaces
- Association Coefficients
  - Q-mode similarity coefficients
  - Symmetrical binary coefficients
  - Asymmetrical binary coefficients
  - Symmetrical quantitative coefficients
  - Asymmetrical quantitative coefficients
  - Probabilistic coefficients
- Q-mode distance coefficients
- Metric distance
- Semimetrics
- R-mode coefficients of dependence
- Non-abundance measures
- Species abundance measures
- Choice of a coefficient

Probabilistic Coefficients

Probabilistic coefficients form a special category. These coefficients are based on statistical estimation of the significance of the relationship between objects.

One of the better known coefficients in this category is Goodall’s probabilistic coefficient. This coefficient takes into account the frequency distribution of the various states of each descriptor in the whole set of objects.

This coefficient was originally developed for plant taxonomy, has been used in paleontology, and has good application in ecology.
Goodall’s Probabilistic Coefficient

This coefficient is nice because like Gower’s coefficient, it permits the use of binary and quantitative descriptors together. Goodall (1966) first conceived of this index and it was later improved by Orlóci (1978).

NB: the use of this coefficient is limited to cluster analysis only!

There are five computational steps which we will review first, then do an example.

Goodall’s Coefficient
- Step 1 -

A partial similarity coefficient $s_j$ is first calculated for all pairs of sites and for each species $j$. With $n$ sites, there are $n(n-1)/2$ calculations.

If the abundances have been normalized, choose the $s_{12}$ function of $S_{19}$. Double-zeros MUST be excluded. This is accomplished by multiplying the partial similarities $s$ by the Kronecker delta $w_{12}$ whose value is 0 upon occurrence of double-zero.

For raw abundance data, $S_{17}$ may be used, computed for a single species at a time.

The result is a partial similarity matrix containing as many rows as there is species, and $n(n-1)/2$ columns.

Goodall’s Coefficient
- Step 2 -

In a second matrix of the same size, for each species $j$ and each of the $n(n-1)/2$ pairs of sites, one computes the proportion of partial similarity values belonging to species $j$ that are larger than or equal to the partial similarity of the pair of sites being considered; the $s_j$ value under consideration is itself included in the calculation of the proportion.

The larger the proportion, the less similar are the two sites with regards to the given species.
Goodall’s Coefficient
- Step 3 -

The proportions (probabilities) of step 2 are combined into a site x site similarity matrix, using Fisher's method, i.e.,
by computing the product $\Pi$ of the probabilities relative to the various species.

Since none of the probabilities can be zero (from previous step) there is no problem in finding product.

But, there is an assumption that species are non-correlated. If they are not, the procedure requires that you use principal component scores instead of species abundances.

Goodall’s Coefficient
- Step 4a -

There are two ways to define Goodall’s similarity index.

In the first approach (4a), the products $\Pi$ are put in increasing order. Following this the similarity between the sites is calculated as the proportion of the products that are larger than or equal to the product of the pair of sites considered:

$$S_{ij} (x_i, x_j) = \frac{\sum_{d=n}^{d=n/2} d}{n(n-1)/2} \text{ where } \begin{cases} d=1 \text{ if } \Pi_i \geq \Pi_{ij} \\ d=0 \text{ if } \Pi_i < \Pi_{ij} \end{cases}$$

Goodall’s Coefficient
- Step 4b -

In the second approach (4b), the $\chi^2$ value corresponding to each product is computed under the hypothesis that the probabilities of the different species are independent vectors:

$$\chi^2_{ij} = -2 \ln \Pi_{ij}$$

which has $2p$ degrees of freedom ($p$ = no. of species). The similarity index is the complement of the probability associated with this $\chi^2$:

$$S_{ij} (x_i, x_j) = 1 - \text{prob} (\chi^2_{ij})$$
**Goodall's Coefficient**

- Example -

5 ponds, 8 phytoplankton species, rel. abundance 0-5

<table>
<thead>
<tr>
<th>Pond</th>
<th>Spp 212</th>
<th>Spp 214</th>
<th>Spp 233</th>
<th>Spp 431</th>
<th>Spp 432</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
</tbody>
</table>

**Goodall's Index**

- Example: Step 1 -

Gower’s matrix of partial similarities

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>212</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>214</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>212</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>214</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>233</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>431</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>432</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>433</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>434</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>435</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>436</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>437</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>438</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

For example, for the pond pair (214,233), the sp-3 has a S of 0.67. In the third row, there are 3 values out of ten (including the value itself) that are ≥ 0.67. Thus, the associated ratio for the new table is 0.3.

**Goodall's Index**

- Example: Step 2, Part a -

Determine the proportion of partial similarity in each row that are ≥ of the pair of sites being considered

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>212</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>214</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>212</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>214</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>233</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>431</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>432</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>433</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>434</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>435</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>436</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>437</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>438</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Goodall’s Index
- Example: Step 2, Part b -
Build a new matrix based upon the proportions of partial similarity ratios determined in Part-a

<table>
<thead>
<tr>
<th></th>
<th>212</th>
<th>212</th>
<th>212</th>
<th>214</th>
<th>214</th>
<th>223</th>
<th>223</th>
<th>223</th>
<th>431</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spp</td>
<td>214</td>
<td>233</td>
<td>431</td>
<td>432</td>
<td>233</td>
<td>431</td>
<td>432</td>
<td>431</td>
<td>432</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0.7</td>
<td>1</td>
<td>1</td>
<td>0.7</td>
<td>0.3</td>
<td>0.7</td>
<td>0.1</td>
<td>1</td>
<td>0.3</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.7</td>
<td>0.3</td>
<td>0.3</td>
<td>0.3</td>
</tr>
<tr>
<td>5</td>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>0.3</td>
<td>0.3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.4</td>
<td>1</td>
<td>0.4</td>
<td>1</td>
<td>0.4</td>
</tr>
<tr>
<td>8</td>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Goodall’s Index
- Example: Step 3 -
Assemble a symmetrical site × site hemi-matrix
(products of the terms in each column from previous matrix)

<table>
<thead>
<tr>
<th></th>
<th>212</th>
<th>214</th>
<th>233</th>
<th>431</th>
<th>432</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ponds</td>
<td>212</td>
<td>214</td>
<td>233</td>
<td>431</td>
<td>432</td>
</tr>
<tr>
<td>212</td>
<td>0</td>
<td>0.00035</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>214</td>
<td>1.0</td>
<td>1.0</td>
<td>0.15000</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>233</td>
<td>0.00000</td>
<td>0.05880</td>
<td>0.01200</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>431</td>
<td>0.28000</td>
<td>0.02100</td>
<td>0.09000</td>
<td>0.00280</td>
<td>–</td>
</tr>
<tr>
<td>432</td>
<td>0.49000</td>
<td>0.09000</td>
<td>0.00280</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

e.g., 0.01200 = (1) (0.1) (1) (0.3) (1) (1) (0.4) (1)

Goodall’s Index
- Example: Step 4a -
Construct site × site similarity hemi-matrix
(based on the proportions of the products that are larger ≥ the product corresponding to each pair of sites)

<table>
<thead>
<tr>
<th></th>
<th>212</th>
<th>214</th>
<th>233</th>
<th>431</th>
<th>432</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ponds</td>
<td>212</td>
<td>214</td>
<td>233</td>
<td>431</td>
<td>432</td>
</tr>
<tr>
<td>212</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>214</td>
<td>1.0</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>233</td>
<td>0.1</td>
<td>0.4</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>431</td>
<td>0.3</td>
<td>0.6</td>
<td>0.8</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>432</td>
<td>0.2</td>
<td>0.7</td>
<td>0.5</td>
<td>0.9</td>
<td>–</td>
</tr>
</tbody>
</table>

e.g., Product of (212, 431) is 0.28; 3 of 10 values are >, hence the similarity S_{212}^{431} = 0.3.
Goodall’s Index
- Example: Step 4b -
If the chosen similarity measure is the complement of the probability assoc. with $\chi^2$ (alternative approach):

<table>
<thead>
<tr>
<th>Ponds</th>
<th>212</th>
<th>214</th>
<th>233</th>
<th>431</th>
<th>432</th>
</tr>
</thead>
<tbody>
<tr>
<td>212</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>212</td>
</tr>
<tr>
<td>214</td>
<td>0.54110</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>233</td>
<td>0.00000</td>
<td>0.00079</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>431</td>
<td>0.00000</td>
<td>0.00869</td>
<td>0.00037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>432</td>
<td>0.00000</td>
<td>0.04340</td>
<td>0.00340</td>
<td>0.23942</td>
<td></td>
</tr>
</tbody>
</table>

E.g., For (212, 431), $\chi^2 = -2\ln(0.28) = 2.5459$, $df = 2p = 16$ with a corresponding $P = 0.99994$; $S^2_{23}(212, 431) = 1 - 0.99994 = 0.00006$

- Ecological Resemblance
- Mode of analysis
- Analytical spaces
- Association Coefficients
- Q-mode similarity coefficients
  - Symmetrical binary coefficients
  - Asymmetrical binary coefficients
  - Symmetrical quantitative coefficients
  - Asymmetrical quantitative coefficients
  - Probabilistic coefficients
- Q-mode distance coefficients
  - Metric distance
  - Semimetrics
- R-mode coefficients of dependence
  - Non-abundance measures
  - Species abundance measures
- Choice of a coefficient

Q-mode Distance Coefficients

Distance coefficients are functions that take their maximum values (usually 1) for two objects that are completely different, and 0 for objects that are identical over all descriptors.

Note that all of the similarity coefficients that we just reviewed can be transformed into distances, usually as the complement; i.e., $D = (1 - S)$. Some simple transforms include, $D = \sqrt{1 - S}$.

Distances, like similarities, are used to measure the association between objects.

Distance coefficients can be divided into 3 groups:
1) metrics
2) semimetrics
3) nonmetrics
“Metric” Distance Coefficients

Metric distance coefficients share the following properties:

1) minimum 0: if \( a = b \), then \( D(a, b) = 0 \);
2) positiveness: if \( a \neq b \), then \( D(a, b) > 0 \);
3) symmetry: \( D(a, b) = D(b, a) \)
4) triangle inequality: \( D(a, b) + D(b, c) \geq D(a, c) \)

“Semimetric” & Nonmetric Distance Coefficients

Semimetric:
These measures 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 \((a, b, \text{ and } c)\), the sum of the distances from \(a\) to \(b\) and from \(b\) to \(c\) may be smaller than the distance from \(a\) to \(c\).

Nonmetric:
These coefficients can take negative values, thus violating the property of positiveness of metrics.

Metric Distances

The most common metric measure is the Euclidean distance. It is computed using Pythagora’s formula, from site-points positioned in a \(p\)-dimensional space called a metric or Euclidean space:

\[
P_1(x_1, x_2) = \sqrt{\sum (x_{ij} - y_{ij})^2}
\]

When there are only two descriptors, this expression becomes the measure of a right-angled triangle:

\[
P_2(x_1, x_2) = \sqrt{(x_1 - y_{12})^2 + (x_2 - y_{22})^2}
\]
Euclidean Distance

The square of $D_1$ may also be used for clustering purposes. One should notice though that $D_1^2$ is a semimetric, which makes it less appropriate than $D_1$ for ordination.

$$D_1(x_i, x_j) = \sum_{j=1}^{p} (y_{ij} - \bar{y}_j)^2$$

Note that ED does not have an upper limit, its value increases indefinitely with the number of descriptors. The value also depends upon the scale of the descriptors. Standardization may be used to reduce scale effects (instead of using raw data).

Euclidean Distance

The Euclidean distance, used as a measure of resemblance among sites on the basis of species abundances, may lead to the following paradox: two sites without any species in common may be at a smaller distance than another pair of sites sharing species.

Orloci (1978) provides an example:

<table>
<thead>
<tr>
<th>Sites</th>
<th>Species</th>
<th>Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>0 1 1</td>
<td>$x_1$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>1 0 0</td>
<td>$x_2$</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0 4 4</td>
<td>$x_3$</td>
</tr>
</tbody>
</table>

From the previous example, we see that the ED between $x_1$ and $x_2$, which have no species in common, is smaller than $x_1$ and $x_3$, which share species $y_2$ and $y_3$.

In general, double-zeros lead to reduction in distances. This situation must be avoided (but occurs frequently with community data, less so with morphometric data). Most argue that ED should NOT be used with species abundance data.

The main difficulty in ecology is that a major method (PCA) orders objects in multidimensional space using $D_1$. 

Euclidean Distance
Average Euclidean Distance

Various modifications have been proposed to deal with the drawbacks of the Euclidean distance applied to species abundances.

The effect of the number of descriptors may be tempered by computing an average distance:

\[ D_{av}^2 (y, y') = \frac{1}{p} \sum_{j=1}^{p} (y_j - y'_j)^2 \] or \[ D_{av} (y, y') = \sqrt{D_{av}^2} \]

Chord Distance

Another modification of ED was proposed by Orloci (1967) and named the chord distance, which has a maximum value of \( \sqrt{2} \) for sites with no species in common and 0 when two sites share the same proportions (without it being necessary for the same absolute abundances).

This measure is the ED computed after scaling the site factors to length 1 (vector normalization).

The chord distance may also be calculated directly from non-normalized data:

\[ D_c (y, y') = \sqrt{\sum_{i=1}^{p} \left( \frac{y_i}{\sqrt{\sum_{j=1}^{p} y_j^2}} - \frac{y'_i}{\sqrt{\sum_{j=1}^{p} y'_j^2}} \right)^2} \]

This solves the problem of using spp. abundance data.

Geodesic Metric

The geodesic metric is a transformation of the chord procedure. It measures the length of the arc at the surface of the hypersphere of unit radius:

\[ D_g (y, y') = \arccos \left( 1 - \frac{D_c^2 (y, y')}{2} \right) \]

In the numerical example we did, pairs of sites \((y_1, y_2)\) and \((y_2, y_3)\), with no species in common are at an angle of 90°, whereas pairs of sites \((y_1, y_3)\), which share two of the three species, are at a smaller angle (88°).
Mahalanobis Distance

Mahalanobis (1936) developed a generalized distance that takes into account the correlations among descriptors and is independent of the scales of the various descriptors.

This measure computes the distance between two points in a space whose axes are not necessarily orthogonal.

In practice, the *Mahalanobis generalized distance* is only used for comparing groups of sites.

For two groups of sites, \( w_1 \) and \( w_2 \), containing \( n_1 \) and \( n_2 \) sites, respectively, and described by the same \( p \) variables, the square of the generalized distance is given by the following matrix formula:

\[
D^2(w_1, w_2) = \mathbf{d}_{12}' \mathbf{V}^{-1} \mathbf{d}_{12}
\]

Where, 

\[ \mathbf{d}_{12} \] is the vector (length = \( p \)) of the differences between the means of the \( p \) variables in the two groups of sites. \( \mathbf{V} \) is the pooled within-group dispersion matrix of the two groups of sites estimated from the matrices of SS&CP between group-centered descriptors for each of the two groups, added up term by term (as in MANOVA).

In other words,

\[
\mathbf{V} = \frac{1}{n_1 + n_2 - 2} \left[ (n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2 \right]
\]

Where \( \mathbf{S}_1 \) and \( \mathbf{S}_2 \) are the dispersion matrices for each of the two groups.

Whereas \( \mathbf{d}_{12} \) measures the difference between the \( p \)-dimensional means of the two groups (\( p \) descriptors), \( \mathbf{V} \) takes into account the covariance among descriptors.
Mahalanobis Distance

The nice feature of this latter method is that the result can be tested for significance.

One must first meet the assumption of matrix homogeneity by applying Kullback’s test:

\[ \chi^2 = \sum_{j=1}^{g} \frac{(n_j - 1)}{2} \ln \left( \frac{\left| \Sigma_j \right|}{\left| V \right|} \right) \]

with df = (g-1)m(m+1)/2, \( n_j \) the number of objects in the group, and \( V \) is the determinant of the pooled within-group dispersion matrix of group \( j \).

---

Mahalanobis Distance

- Testing for Significance -

To perform the test of significance, the generalized distance is transformed into Hotelling’s \( T^2 \) (1931) statistic:

\[ T^2 = \frac{n_1 n_2}{(n_1 + n_2)} D^2 \]

Then compute the appropriate \( F \)-statistic as:

\[ F = \frac{n_1 + n_2 - (p + 1)}{(n_1 + n_2 - 2p)} T^2 \]

With df = \( p \), \( [n_1 + n_2 - (p + 1)] \)

---

Manhattan Metric

\[ D_1 (x_i, x_j) = \sum_{j=1}^{p} |y_{ij} - y_{lj}| \]

The Manhattan metric, city-block metric, or taxicab metric all refer to the same distance measure. It refers to the fact, that for two descriptors, the distance between two sites is the distance on the abscissa plus the distance on the ordinate (much like the orthogonal distances traveled by taxicabs in NYC).

This metric presents the same problems with double zeros as in ED and leads to the same paradox.
Mean Character Difference

The mean character difference was originally proposed by Czekanowski, an anthropologist, in 1909:

\[ D_k(x_1, x_2) = \frac{1}{p} \sum_{j=1}^{p} |y_{1j} - y_{2j}| \]

It has the advantage over \( D_s \) of not increasing with the number of descriptors \( (p) \).

It may be used for species abundance analysis if you exclude double zeros from the absolute value of the differences in \( y \) by replacing with \( (p \cdot \text{no. double-zeros}) \).

Whittaker’s Index of Association

WIA is well adapted to species abundance data, because each species is first transformed into a fraction of the total number of individuals at the site, before the subtraction:

\[ D_i(x_1, x_2) = \frac{1}{2} \sum_{j=1}^{p} \left| \frac{y_{1j}}{\sum_{j=1}^{p} y_{1j}} - \frac{y_{2j}}{\sum_{j=1}^{p} y_{2j}} \right| \]

The difference is zero for a species when its proportions are identical in the two sites.

Alternative to Whittaker’s Index

An identical result to the WIA is obtained by computing, over all species, the sum of the smallest fractions calculated for the two sites:

\[ D_o(x_1, x_2) = 1 - \sum_{j=1}^{p} \min \left( \frac{y_{1j}}{\sum_{j=1}^{p} y_{1j}}, \frac{y_{2j}}{\sum_{j=1}^{p} y_{2j}} \right) \]
Likewise, Australians Lance & Williams (1967) provide the Canberra metric as an alternative to the Manhattan metric:

\[ D_{10}(x_i, x_j) = \sum_{j=1}^{p} \frac{|y_{ij} - y_{ij}|}{(y_{ij} + y_{ij})} \]

A scaled version of \( D_{10} \) was devised by Clark (1952):

\[ D_1(x_i, x_j) = \sqrt{\frac{1}{p} \sum_{j=1}^{p} \left( \frac{y_{ij} - y_{ij}}{y_{ij} + y_{ij}} \right)^2} \]

Another index with some good properties, which is related to \( D_{10} \), was developed by an anthropologist under the name Coefficient of Racial Likeness.

Using this coefficient, it is possible to measure a distance between groups of sites, like with the Mahalanobis distance (\( D_2 \)), but without eliminating the effect of correlations among descriptors:

\[ D_2(w_i, w_j) = \sqrt{\frac{1}{p} \sum_{j=1}^{p} \left( \frac{y_{ij} - y_{ij}}{s_{ij}^2 + (s_{ij}^2/n_i)} \right)^2} - \frac{2}{p} \]

where \( w_i \) & \( w_i \) contain \( n_i \) & \( n_j \), \( y_{ij} \)-bar is mean of descriptor \( j \) in group \( i \), \( s_{ij} \) is the variance.

The last group of common metrics are the \( \chi^2 \) distance measures. The most general form is known as the \( \chi^2 \) metric.

In order to calculate the \( \chi^2 \) metric, the data matrix must first be transformed into a matrix of conditional probabilities. The elements of the matrix become the new terms \( y_{ij}/y_{ij} \), where \( y_{ij} \) is the sum of the frequencies in row \( i \).

An example may be the easiest way to understand...
The distance between the first two rows of the right-hand matrix could be computed using the formula for Euclidean distance ($D_1$), but, the most abundant species would contribute predominantly to the sum of squares.

Instead, the $\chi^2$ metric is computed using a weighted expression:

$$
D_1(x_1, x_2) = \sqrt{\sum_i \frac{1}{y_{i2}} \left( \frac{y_{i1}}{y_{i2}} - \frac{y_{j1}}{y_{j2}} \right)^2}
$$

Where $y_{ij}$ is the sum of the frequencies of the column $j$. While this measure has no upper limit, most values < 1.

For the numerical example, computation of $D_1$ between the first two sites yields:

$$
D_{15}(x_1, x_2) = \sqrt{\frac{(0.563 - 0.543)^2}{77} + \frac{(0.125 - 0.174)^2}{33} + \frac{(0.188 - 0.217)^2}{14} + \frac{(0.125 - 0.065)^2}{25}} = 0.015
$$

NB: The 4th species, which is absent from the first two sites, cancels itself out; thus how $\chi^2$ metric deals with double-zeros.
Chi-Square Distance

The $\chi^2$ distance ($D_{\chi^2}$) is related to the $\chi^2$ metric ($D_{\chi}$). It differs from the metric in that the terms of the sum of squares are divided by the probability of each row in the table instead of its absolute frequency. Thus,

$$D_{\chi^2}(x_i, x_j) = \sqrt{\sum_{j=1}^{n} \frac{1}{p_{j,j}} \left( \frac{y_{ij}}{y_{ij}} - \frac{y_{ij}}{y_{ij}} \right)^2}$$

The $\chi^2$ distance is the distance preserved in correspondence analysis (CA), when computing similarity between sites (as we’ll see later).

Hellinger Distance

The last distance measure in this category is the Hellinger distance. This is often recommended prior to a principal coordinates analysis (PCO):

$$D_H(x_i, x_j) = \sqrt{\sum_{j=1}^{n} \left( \frac{y_{ij}}{\sqrt{y_{ii} \cdot y_{jj}}} - \frac{y_{ij}}{\sqrt{y_{ii} \cdot y_{jj}}} \right)^2}$$

Q-mode Distance Coefficients: Semimetrics

Some distance measures do not follow the fourth property of metrics, i.e., the triangle inequality axiom.

As a consequence, they do not permit a proper ordination of points in Euclidean space. They may, however, be used for ordination by PCO after correction for negative eigenvalues.

One of the first semimetrics was derived from the Sørenson coefficient ($S_{\text{so}}$) which was used to form the nonmetric coefficient:

$$D_{\text{so}}(x_i, x_j) = 1 - \frac{2a}{2a + b + c} = \frac{b + c}{2a + b + c}$$
Among the measures for species abundance data, the coefficients of Steinhaus ($S_{17}$) and Kulczynski ($S_{18}$) are semimetrics when transformed into distances. In particular, $D_{14} = 1 - D_{17}$ was first described by Odum (1950) and later by Bray and Curtis (1957) who called it the percentage difference:

$$D_{14}(x, y) = \frac{\sum |x_i - y_i|}{\sum (x_i + y_i)} = 1 - \frac{2W}{A + B}$$

Contrary to the Canberra metric ($D_{9a}$), differences between abundant species contribute the same as rare species. This is often a desirable property, particularly when using normalized data.

---

% Ecological Resemblance
% Mode of analysis
% Analytical spaces
% Association Coefficients
% Q-mode similarity coefficients
% Symmetrical binary coefficients
% Asymmetrical binary coefficients
% Symmetrical quantitative coefficients
% Asymmetrical quantitative coefficients
% Probabilistic coefficients
% Q-mode distance coefficients
% Metric distance
% Semimetrics
% R-mode coefficients of dependence
% Non-abundance measures
% Species abundance measures
% Choice of a coefficient

---

**R-mode: Coefficients of Dependence**

The main purpose of R-mode analysis is to investigate the relationships among descriptors, and are sometimes used in PCA or DA to order objects.

Most dependence coefficients are amenable to statistical testing. For such coefficients, it is thus possible to associate a matrix or probabilities with the R-matrix, if required by subsequent analyses.

If you do statistical testing, the data must follow all of the regular assumptions for the data (e.g., normality, etc.).
Descriptors Other Than Species Abundances

The resemblance between quantitative descriptors can be computed using parametric measures of dependence; i.e., measures based on parameters of the frequency distributions of descriptors.

These measures are the covariance and the Pearson correlation coefficient. They can ONLY be adapted to descriptors whose relationships are linear.

Covariance

Recall that the covariance between descriptors \( j \) and \( k \) is computed from centered variables \((y_j - \bar{y}_j)\) and \((y_k - \bar{y}_k)\).

The range of values of the covariance has no a priori upper or lower limits.

The variances and the covariances among a group of descriptors form their dispersion matrix \( S \):

\[
S = \frac{1}{n-1} \begin{bmatrix} (y - \bar{y}) \end{bmatrix} \begin{bmatrix} (y - \bar{y}) \end{bmatrix}
\]

Recall: multiply matrix of centered data w/its transpose.

Correlation

Pearson’s correlation coefficient \( r_{jk} \) is their covariance of descriptors \( j \) and \( k \) computed from standardized variables.

The coefficients of correlations among a group of descriptors form the correlation matrix \( R \).

Correlation coefficients range in value from -1 to +1.

The significance of individual coefficients (i.e., \( H_0: r = 0 \)) can be statistically tested.
Correlation: R vs. Q

Some authors have used Pearson’s $r$ for Q-mode analyses after transposing the primary matrix. There are, however, a number of objections to doing this, some of which include:

1. $r$ is dimensionless and may be hard to interpret
2. In R-mode, the value of $r$ remains unchanged after rescaling, but may change dramatically in Q-mode

UPSHOT: measures that are designed for one mode of analysis should not be analyzed in the other mode!

Nonparametric Correlation

The resemblance between semi-quantitative descriptors, and more generally between any pair of ordered descriptors whose relationship is monotonic may be determined using nonparametric measures of dependence.

Spearman’s $r$ (continuous or ordinal variables) and Kendall’s $\tau$ (ordinal variables) are appropriate to use under these circumstances, and like Pearson’s $r$, can be subjected to statistical testing.

Species Abundances: Biological Associations

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

This problem surfaces regularly in community data because biological assemblages usually contain a small number of dominant species and a large number of rare species.

The literature is replete with incorrect approaches to this problem. Double-zeros need to be neutralized or not included in the analysis.
Approaches to Minimizing the Double-Zero Problem

1) Eliminate less frequent species from the primary data matrix. They will be of little use in assessing ecological species associations.

2) Eliminate all zeros from the comparisons by declaring that zeros are missing values.

3) Eliminate double-zeros only from the computation of the correlation or covariance matrix (this must generally be programmed separately). The resulting dispersion matrix can then be directly analyzed (e.g., PCA).

Note that this is not a full list of options.

For example, Correspondence Analysis (CA) is a special form of PCA which preserves the $\chi^2$ distance ($D_{16}$) instead of the Euclidean distance ($D_1$).

Because $D_{16}$ excludes double-zeros, whereas $D_1$ includes them, CA is usually better adapted to the study of species associations than is PCA.

Other Approaches

Biological associations may also be defined on the basis of co-occurrence of species instead of the relationships between fluctuations in abundances.

In fact, quantitative data may not accurately reflect the proportions of the various species in the environment (usually because of sampling or identification problems).

There are many approaches to this in the literature, but by far, the most common is the $2 \times 2$ frequency table.
2 × 2 Frequency Table

<table>
<thead>
<tr>
<th></th>
<th>Presence</th>
<th>Absence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species y₁</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Presence</td>
<td>a+b</td>
<td></td>
</tr>
<tr>
<td>Absence</td>
<td>c+d</td>
<td></td>
</tr>
<tr>
<td>Species y₂</td>
<td>a+c</td>
<td>b+d</td>
</tr>
<tr>
<td>Presence</td>
<td>n=a+b+c+d</td>
<td></td>
</tr>
<tr>
<td>Absence</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Where a and d are numbers of sites in which the two species are present and absent, respectively; whereas b and c are the numbers of sites in which only one of the two species is present; n is the total number of sites.

Binary Coefficients

As already discussed, many binary coefficients exclude double-zeros. Jaccard’s coefficient of community ($S_j$) has been popular

$$S_j(x, y) = \frac{a}{a + b + c}$$

along with its corresponding distance measure:

$$D = 1 - S_j(x, y) = \frac{b + c}{a + b + c}$$

Binary Coefficients

Dice’s coincidence index ($S_d$) (a.k.a. Sørenson’s coefficient) was in fact originally designed to specifically study species associations:

$$S_d(x, y) = \frac{2a}{2a + b + c}$$

A more elaborate coefficient was proposed by Fager & McGowan (1963) to make minor corrections to $S_d$ (esp. for small sample sizes):

$$S_{f}(x, y) = \frac{a}{\sqrt{(a+b)(a+c)}} - \frac{1}{2\sqrt{a+c}}$$
Choice of a Coefficient

Given that multivariate statistics is exploratory in nature, there are not the same "hard and fast" rules as one might often see in inferential statistics.

There are, however, important guidelines. We have seen how the choice of coefficient can have a major influence on the outcome and interpretation of resemblance. Thus, considerable care should be exercised in choosing a resemblance coefficient.

Please refer to class handout for selection criteria.

R supports a variety of coefficients used in different libraries.

The R STATS library contains the `dist()` function which contains euclidean, manhattan, binary, and Canberra.

The vegan library supports the `vegdist()` function and LabDSV provides the `dsvdis()` function, both of which support many more coefficients.

In all cases, the first argument to the function refers to the primary matrix and the second argument is the index. For example:

```r
dis.bray <- vegdist(veg, method="bray")
```

Which creates a dissimilarity hemi-matrix called `dis.bray` using the `vegdist()` procedure and `bray-curtis` coefficient.

OCCAS Analysis

One way of helping to assess resemblance coefficients is to construct artificial data representing contrasting situations that a S or D value should be able to discriminate.

OCCAS (ordered comparison case series; Hajdu 1981) involves constructing just such a series, corresponding to linear changes in the abundance of two species along a simulated gradient.

The method is straightforward and easy to apply. In order for a coefficient to perform well, it MUST provide a linear result. Gower and Legendre (1986) used this approach to evaluate 15 binary coefficients and 10 coefficients for quantitative data.
OCCAS Analysis

-Example-

Consider two species.

Site 1 has frequencies were \( y_{11} = 100 \) and \( y_{12} = 0 \).

Site 2 has frequencies \( y_{21} = 50 \) and then \( y_{22} \) was varied from 10 to 120, in steps of 10.

The results for three coefficients are:

ANOSIM

Analysis of similarities (ANOSIM) provides a way to directly test whether there is a significant difference between two or more groups of sampling units.

The function ANOSIM operates directly on a dissimilarity matrix produced by VEGDIST.

The test statistic is \( R \). The statistical significance of \( R \) is assessed by permuting the grouping vector to obtain the empirical distribution of \( R \) under a null-model.

ANOSIM

- Example -

Create two data sets that share a common dimension. One data set contains the species/stand info, the other is an indexing set.