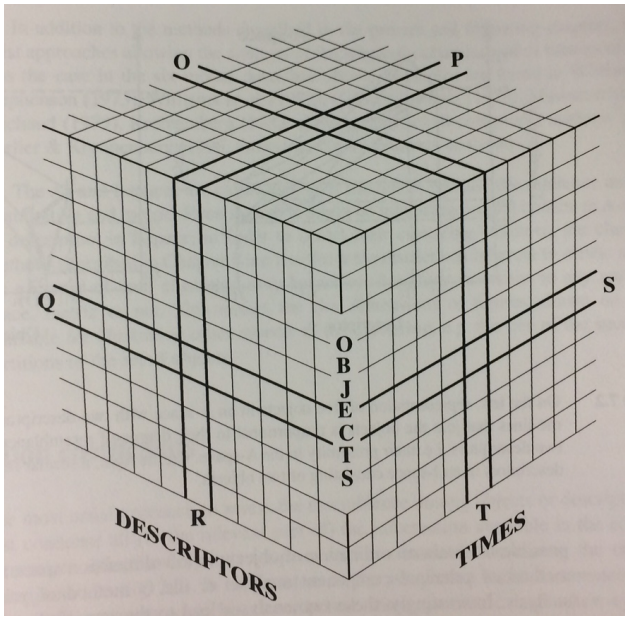
Association metrics and implementation in R

Q mode vs. R mode

A common format for a multivariate data set is the observation of a number of "descriptors" across a number of "objects". In biogeochemistry or biochemistry studies, the descriptors might be the chemical concentrations of a number of compounds of interest across ecosystems, individuals, or species. In community ecology this would be the abundance or presence-absence of species across a number of communities. It is useful to distinguish situations when we are making inference about the objects, known as Q mode, or making inference about the descriptors (R mode). One important reason to distinguish between these two possibilities is that different association metrics perform better in one mode or the other.

As a point of context, Q and R come from a broader conceptualization of data (Cattell 1966; Legendre & Legendre 1998).



This "data box" (objects x descriptors x time) can be considered from many viewpoints:
O – among time instances, based on all observed descriptors for a single object
P – among descriptors, based on all observed times for a single object
Q – among objects, based on all observed descriptors for a singe instance
R – among descriptors, based on all observed objects for a single instance
S – among objects, based on all observed times for a single descriptor
T – among time instances, based on all observed objects for a single descriptor

Pair-wise associations

Association metrics are applied in a pair-wise manner for most multivariate approaches, including dimension reduction, visualization, classification, and hypothesis testing. This means that when conducting an analysis in Q mode we will generate a symmetrical, square matrix with the number of rows and columns equal to the number of objects in our data set. The diagonal of this matrix will contain the same value, which indicates an identical match as the diagonal contains all of the self-to-self pair-wise comparisons. If in R mode, a similar square matrix will be generated, but the number of rows and columns will be the number of descriptors included in the analysis.

Association Metrics
Association is a general term to describe any measure used to quantify resemblance or difference. This can be thought of in at least three ways:

Similarity – Similarities are maximum when the two objects are identical and minimum when the objects are completely different. Similarities are most commonly used when comparing objects (Q mode).

Distance – Distances work opposite to similarities. They are maximum when completely different and minimum when identical. These are also used when comparing objects (Q mode).

Similarities (S) and distances (D) can be transformed between each other. If a similarity varies between 0 and 1 it is often transformed as $D = 1 - S$, $D = \text{sqrt}(1-S)$, or $D = \text{sqrt}(1-S^2)$. Distances can sometimes vary beyond some pre-determined upper bound, but they can be normalized as $D_{norm} = D/D_{max}$ or $D_{norm} = (D-D_{min})/(D_{max}-D_{min})$. The previously mentioned transformations for S and D can then be applied to D_{norm} as well.

Dependence – Dependence metrics are commonly used in R mode. These are often correlation coefficients or some similar metric.

Symmetry and Association Metrics
All or nearly all association metrics we will use in class are symmetrical in the sense that the similarity between n_1 and n_2 is the same as the similarity between n_2 and n_1. As a result the concept and term "symmetrical" is used to describe another aspect of association metrics. In this case, a symmetrical association metric is a metric that considers the shared absence (a value of zero) of a descriptor to be informative or meaningful. In contrast, an asymmetrical association metric ignores cases where a descriptor is zero in two objects. Symmetric association metrics are often useful when considering concentrations of some chemical constituent, whereas asymmetric association metrics are commonly used when thinking about species counts.

Common Association Metrics for Q mode
-Quantitative data

*Euclidean distance: not commonly used for species data as double zeroes are included (this is a symmetric distance). There is no upper limit to this distance and it is sensitive to the range of the descriptors. Often a z-score transformation (subtract the mean and divide by the standard deviation) within each descriptor is applied to data prior to use of this distance.
*Bray-Curtis dissimilarity: often used for species data. Can be used on raw abundances, log-transformed abundances, or relative abundances.
*Chord distance: often used for species data. Euclidean distance after the chord transformation (site vectors normalized to a length 1)
*Hellinger distance: often used for species data. Euclidean distance after the Hellinger transformation (species abundances scaled by object total abundance and the square-root transformed). This distance reduces the effect of rare descriptors.

-Binary data

*Simple matching coefficient: This is the simplest symmetrical similarity available for binary data.
*Jaccard similarity: commonly used with species presence-absence data. This is the ratio between the number of double 1s and the number of species (excluding double zero species).

*Sørensen similarity: commonly used with species presence-absence data. This more heavily weights shared presences and is equivalent to one minus the Bray-Curtis distance computed on presence-absence data.

-Qualitative data

*Gower's similarity: This similarity is not commonly used, but can handle categorical variables.

R mode

-Quantitative data

*Covariance and parametric and non-parametric correlation coefficients are commonly used in this case. These metrics are sensitive to double zeros as well as differences in total individuals across objects. These issues can be reduced by transforming the data prior to use. Common transformations include using relative abundance, the chord transformation, and the Hellinger transformation. Rank-based correlation coefficients can help as well.

*The Chi-square distance is also used for species data in R mode.

-Binary data

*The Jaccard and Sørensen similarities can be used in R mode.

-Qualitative data

*Correlation coefficients can work here, but these are linear and may run into issues depending on the distribution of data across categories. Rank-based correlation coefficients can help as well.