Chapter 4 - Multidimensional quantitative data

Multidimensional Variables

Simple analysis of descriptors is not enough becase it doesn't take into account the covariance among descriptors. *Remember, objects are set a priori and descriptors "describe" each object.* Lets assume that in this case the objects are sites and the descriptors are species. As the number of descriptors increases, the number of dimensions of the random variable increases. Therefore more axes are necessary to construct the space in which the objects are plotted.

This chapter focuses on the *dependence* among descriptors.

To sum up:

1. The p descriptors in ecological data matrices are the pdimensions of a random variable "descriptors". As the number of species increases, so do the dimensions of the sites.
2. The p descriptors (species) are not independent of one another. That's why we can't use unidimensional analysis.

Variance

Variance a measure of the dispersal of a random variable y around it's mean. ie. how much does a variable deviate from it's mean.

$$s_j^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \overline{y_j})^2 \tag{1}$$

Covariance

Is the extension to two descriptors of variance. It measures the joint dispersion of two random variables y_i and y_k around their means.

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(y_{ij} - \overline{y_j})^2(y_{ik} - \overline{y_k})^2 \tag{2}$$

When the covariance is positive it means that both descriptors have a positive relationship. A negative covariance means that the descripts have a negative relationship.

Dispersion matrix S

Contains the variances and covariances of the p descriptors. Therefore S is an association matrix. All eigenvalues of S are positive or null. Ideally, the matrix S should be estimated from a number of observations n larger than the number of descriptors p. When $n \leq p$ then the matrix has null eigenvalues but usually the first few are not affected.

Correlation matrix R

The covariance measures the joint dispersion of two random variables aorund their means. The correlation is defined as a measure of the dependence between two random variables y_j and y_k. Sometimes, descriptors don't have a common scale. For example when the descriptor or a site are two environmental variables and each have their own different units. In these cases, calculating covariances doesnt make sense, unless the descriptors are reduced to a common scale. This common scale standandizes the values to their standard normal distribution using: $z_i = \frac{y_i - \overline{y}}{s_y}$ where $\overline{y}$ is the mean and s_y is the standard deviation of that descriptor.

The covariance (S) matrix of two standardized descriptors is the linear correlation. Therefore the correlation matrix is the dispersion matrix of the standardized variables.

Matrices S and R are related to each other by the diagonal matrix of the standard deviations of Y, simbolized by $D(\sigma)$.

$$\Sigma = D(\sigma)RD(\sigma) \tag{3}$$

Multinormal Distribution

Central limit theorem: When a random variable results from several independent and additive effects, of which none has a dominat variance, then this variable ttends towards a normal distribution even if it's effects are not themselves normally distributed.