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.

Three sets of parameters are therefore necessary to specify a multidimensional normal distribution: the vector of means μ, the diagonal matrix of the standard deviations $D(\sigma)$ and the correlation matrix P.

$$\mu = \begin{bmatrix}\mu_1 & \mu_2\end{bmatrix} D(\sigma) = \begin{bmatrix}\sigma_1 & 0 \\ 0 & \sigma_2\end{bmatrix} P = \begin{bmatrix}1 & \rho \\ \rho & 1\end{bmatrix}$$

Plotting this distribution leads to a series of ellipsoids. When the variables are standardized, the family of elipsoids is centered on $\mu = (0, 0)$. As ρ approaches zero, the shapes of the ellipses tend to become circular. As ρ approaches +1 or -1 the ellipses tend to elongate. The sign of the correlation determines the orientation of the ellipses relative to the axis.

Principal Axis

The first principal axis is the line that passes through the greatest dimension of the ellipsoid. The next principal axes go through the next greatest dimensions, smaller and smaller of the p-dimensional ellipsoid. These principal axis are the basis of PCA.

These principal axis are perpendicular to each other in the hyperspace and the vector of coordinates which specifies the first principal axis is one of the eigenvectors of the matrix Σ.

In order to find the vectors of coordinates specifying the p succesive principal axes:

1. Rank the eigenvalues of the matrix Σ.
2. Associate the p eigenvectors to their corresponding eigenvalues.
3. Calculate a new p-dimensional variable along which the dispersion ellipeses are positioned with respect to the principal axes instead of the original cartesian system.

Multiple and partial correlations

Multiple linear correlation

It applies in cases where there is one response variable and several explanatory variables. The coefficient of multiple determination measures the fraction of the variance of y_k which is explained by a linear combination of $y_1, y_2, y_3 ... and y_p$ where p is the number of explanatory variables. Then the multiple correlation coefficient R is the square root of the coefficient of multiple determination.

Main properties:

- The multiple correlation coefficient measures the intensity of the relationship between a response variable and a linear combination of several explanatory variables
- The square of the multiple correlation coefficient, *coefficient of multiple determination*, measures the fraction of the variance of the reponnse variable which is explained by a linear combination of the explanatory variables.

Partial correlation

The second approach to correlation in the multidimensional context, applies to situations where the relationship between teh two variables is influences by their relationships with other variables. It measures what the correlation between y_j and y_k would be if other variables $y_1, y_2, y_3 ... and y_p$, suposed to influence both y_j and y_k were held constant at their means. In order to calculate partial correlation coefficients, the set of variables is divided into two subsets, the first one contains the variables between which the partial correlation is computed while contorlling for the influence of the variables in the second subset.

The partial correlation coefficient is a parameter of the multinormal conditional distribution.

Main properties:

- The partial correlation coefficient measures the intensity of the linear relationship between two rancom variables while taking into account their relationships with other variables.

Interpretation of correlation coefficients

Correlation coefficients are often interpreted in terms of casual relationships among descriptors. However, the choise of causal model always requires hypotheses, or input of external information.

Causal modelling can be done using three variables, looking at partial and simple correlation coefficients. For this type of modelling, at least two of the simple correlation coefficients must be significantly different from zero. Using these variables, four different linear models can be formulated (See fig 4.11 for the example).

Partial correlations do not provide the same information as path analysis. In path analysis, one is mainly interested in partitioning the relationship between predictor and criterion variables into direct and indirect components.

For these type of analysis, linearity must be checked since it is based on linear correlation coefficients. If the data is not normally distributed, then it must be transformed or non parametric methods must be used.

Tests for normality and multinormality

The first step in testing for normality is plotting the frequency distributions of the data.
Next, for unimodal distributions, you can measure skewness and kurtosis. Skewness is a measure of asymetry.

Skewness is 0 for the normal distribution. Kurtosis is a measure of flatness or peakedness. Kurtosis is 3 for a normal distribution. Distributions that are flatter will have a lower kurtosis and a higher kurtosis means the distributions have more observations around the mean than the normal distribution.

For tests of goodness of fit, you can use the Kolmogorov-Smirnov test or the Shapiro-Wilk test.