15. INTRODUCTION TO MULTIVARIATE ANALYSES

15.1. Multivariate data

A multivariate data set includes more than one variable recorded from a number of replicate sampling or experimental units, sometimes referred to as objects. If these objects are organisms, the variables might be morphological or physiological measurements; if the objects are ecological sampling units, the variables might be physico-chemical measurements or species abundances. We have already considered multivariate data in linear models with two or more predictor variables, e.g. multiple regression (Chapter 6) and multifactor analysis of variance (Chapters 9-11). For these analyses, we have multiple predictor (independent) variables. The multivariate analyses we will discuss in the remaining chapters either deal with multiple response variables (e.g. MANOVA - Chapter 16) or multiple variables that could be response variables, predictor variables or a combination of both. This chapter will introduce some aspects of multivariate data and analysis that apply generally to many of the methods we will describe in the subsequent three chapters. We will illustrate these aspects with four data sets from the recent biological literature. For each data set, there are $i = 1$ to n objects with $j = 1$ to p variables measured for each object.

Chemistry of forested watersheds

In Chapter 2, we first described the study of Lovett *et al*. (2000) who examined the chemistry of forested watersheds in the Catskill Mountains in New York. They chose 39 first and second order streams (objects) and measured the concentrations of ten chemical variables (NO_3^-, total organic N, total N, NH_4^-, dissolved organic C, SO_2^{2-}, Cl^-, Ca^{2+}, Mg^{2+}, H^+), averaged over three years, and four watershed variables (maximum elevation, sample elevation, length of stream, watershed area).

Plant functional groups and leaf characters

In Chapter 9, we described the study of Reich *et al*. (1999) who examined the generality of leaf traits from different species across a range of ecosystems and geographic regions. We will use a subset of their data, Wisconsin forbs, with ten species as the objects. There were five variables measured for each species: specific leaf area, leaf nitrogen concentration, mass-based net photosynthetic capacity, area-based net photosynthetic capacity and leaf diffusive conductance at photosynthetic capacity.

Wildlife underpasses in Canada

Clevenger & Waltho (2000) reported on the effectiveness of road underpasses for wildlife in Banff National Park in Alberta, Canada. For part of their study, they quantified the human activity at the underpasses as numbers of people on bikes, on horses and on foot. The objects were the eleven underpasses and the variables were the three human activities and the data were counts.

Bats and African woodlands

Fenton *et al*. (1998) studied the effects of woodland disturbance on species richness and abundance of bats in northern Zimbabwe. They has four groups of sites: nine intact and nine impacted sites in Mana, six intact sites in Kanyati and six impacted sites in Matusadona. The sites within each area and disturbance category are not true replicates for assessing effects of disturbance so, like Fenton *et al*. (1998), we will combine the sites within each group. There were four objects (area and disturbance combinations) and 15 variables,

species of bats. The data were numbers of each species of bat and there were numerous zero values, i.e. species absent.

15.2. Distributions and associations

In a univariate context, we can describe the distribution of each variable and many of the parametric univariate analyses for estimating linear models and testing hypotheses about their parameters assume that the distribution of the response variable being analyzed is of a particular form (Chapters 5, 6, 8 to 14). For example, classical linear models assume normality (although the analyses are robust to this assumption under many circumstances), while generalized linear models allow other distributions from the exponential family (e.g. binomial, Poisson etc.). Although the multivariate analyses we will introduce in the next three chapters are mainly descriptive, interval estimation and hypothesis tests of parameters can also be relevant and usually require the assumption of multivariate normality, where all variables and linear combinations of variables are normally distributed (Tabachnick & Fidell 1996). The simplest multivariate normal distribution is the bivariate normal distribution described in Chapter 5. Other multivariate distributions are obviously possible, although less commonly used in multivariate analyses.

One measure of the centre of a multivariate distribution is the centroid. In multivariate space where each dimension is a variable, the centroid is the point represented by the univariate means of the distributions of each of the variables (Figure 15-1). The centroid is not usually estimated by a single value but is used as a description of centre of a multivariate normal distribution and for detecting multivariate outliers (Section 15.9.1).

We can summarize variation in single variables by sums-of-squares (SS) and variances (Chapter 2). When we have more than one variable, we not only have variances for each variable but also covariances between variables. To represent variation in multivariate data sets, we must use some simple matrix algebra. A data matrix (**Y**) for *n* objects by *p* variables is represented in Table 15-1, and illustrated using the data from Reich *et al*. (1999) for Wisconsin shrubs.

With more than one variable, we calculate both sums-of-squares for each variable and sums-of-cross-products between variables to get a *p* by *p* sums-of-squares-and-cross-products (**SSCP** or **S**) matrix (Table 15-2). The rows and columns of this matrix represent the variables (j = 1 to p). The main diagonal of this matrix contains the sums-of-squares for each variable. The other entries are the sums-of-cross-products, the sum of the product of the deviations of the value for each variable from its sample mean. Note that this matrix is symmetrical, i.e. the sum-of-cross-products between Y_1 and Y_2 is the same as the sum-of-cross-products between Y_2 and Y_1.

We can convert this matrix to a *p* by *p* matrix of variances and covariances (**C**) by dividing the sums-of-squares and sums-of-cross-products by their degrees-of-freedom (n-1), where the main diagonal contains the variances for each variable and the other entries are the covariances between pairs of variables (Table 15-3). The covariance matrix can also be obtained directly from the raw data matrix **Y**, if each variable is centered (to a mean of zero), by multiplying $\mathbf{Y}.\mathbf{Y}^{-1}$, where $\mathbf{Y}^{-1}$ is the inverse ("reciprocal") of the centered raw data matrix.

There are two ways we can summarize the variability of a multivariate data set based on the variance-covariance matrix (Jackson 1991):

- The determinant of a square matrix is a single number summary of the matrix. The determinant of the variance-covariance matrix ($|\mathbf{C}|$) represents the generalized variance of the matrix.
- The trace of the variance-covariance [Tr(**C**)] matrix is the sum of the diagonal values, i.e. the sum of the variances of the centered individual variables.

Finally, we can also standardize these covariances by dividing by the standard deviations of the two variables involved to produce correlations and thus a correlation matrix (**R**), where r_{12} is the correlation coefficient between variables 1 and 2 etc. (Table 15-4). Note the main diagonal consists of ones because the correlation between each variable and itself is one. Covariances and correlations are measures of association between variables. Other measures of association include the c^2 statistic, discussed in Chapter 14 as a measure of association for contingency tables.

If our objects occur in groups (e.g. experimental treatments), then we can calculate these matrices for between and within groups, analogous to analyses of variance in Chapters 8 to 11. Analyses based on multiple variance-covariance matrices nearly always have the assumption that the within-groups matrices have equal variances and covariances.

15.3. Linear combinations, eigenvectors and eigenvalues

15.3.1. Linear combinations of variables

One of the fundamental techniques in multivariate analyses is to derive linear combinations of the variables that summarize the variation in the original data set. Basically, we are "consolidating" (*sensu* Tabachnick & Fidell 1996) the variance from a data matrix into a new set of derived variables, each of which is a linear combination of the original variables. For $i = 1$ to n objects and $j = 1$ to p original variables:

$$z_{ik} = c_1 y_{i1} + c_2 y_{i2} + ... c_j y_{ij} + ... + c_p y_{ip} \qquad (15.1)$$

In equation 15.1, z_{ik} is the value of the new variable k for object i, y_{i1} to y_{ip} are the values of the original variables for object i and c_1 to c_p are weights or coefficients that indicate how much each original variable contributes to the linear combination. Depending on the analysis, these new variables are termed, variously, discriminant functions, canonical functions or variates, principal components or factors. This linear combination is analogous to a regression equation. For some analyses, the linear combination may include a constant (an intercept in regression terminology):

$$z_{ik} = \text{constant} + c_1 y_{i1} + c_2 y_{i2} + ... c_j y_{ij} + ... + c_p y_{ip} \qquad (15.2)$$

The form in equation 15.2 is common when the variables are not standardized to zero mean and unit variance; if they are, then the constant becomes zero and equation 15.1 is appropriate.

The derived variables are extracted so the first explains most of the variance in the original variables, the second explains most of the remaining variance after the first has been extracted but is uncorrelated with the first, the third explains most of the remaining variance after the first and second have been extracted but is uncorrelated with either the first or second etc. The new derived variables are independent of, uncorrelated with, each other. The number of new derived variables is the same as the number of original variables (p), although the variance is usually consolidated in the first few derived variables.

15.3.2. Eigenvalues

Eigenvalues, also termed characteristic or latent roots (l_1, l_2, l_3, ... l_k ... l_p), represent the amount of the original variance explained by each of the $k = 1$ to p new derived variables. These eigenvalues are population parameters and we estimate them using maximum likelihood (ML) to produce (l_1, l_2, l_3, ... l_k ... l_p) and can also determine their approximate standard errors. Note from Box 15-1 that if we use a covariance matrix and centred variables, then the sum of the eigenvalues is equal to the trace of the original covariance matrix, i.e. the sum of the variances of the original centred variables. If we use a correlation matrix and centred and standardized variables, the sum of the eigenvalues would equal the trace of the correlation matrix, i.e. the sum of the variances of the original standardized variables. We have simply rearranged the variance in the association matrix so that the first

few derived variables explain most of the variation that was present (between objects) in the original variables. The eigenvalues can also be expressed as proportions or percentages of the original variance explained by each new derived variable (component).

15.3.3. Eigenvectors

Eigenvectors (characteristic vectors) are lists of the coefficients or weights showing how much each original variable contributes to each new derived variable. In general terms, the eigenvectors contain the c_j in equation 15.1 but these coefficients can be scaled in different ways so are often represented as u_j, v_j or w_j in matrix descriptions of multivariate analyses – see Box 15-1). The eigenvectors are commonly scaled so the sum of squared coefficients equals one; other forms of scaling are possible. We estimate the coefficients with maximum likelihood and can determine approximate standard errors. These linear combinations can be solved to provide a score (z_{ik}) for each object for each new derived variable. Note that there is the same number of derived variables as there are original variables (p). The new derived variables, each with an eigenvector of coefficients and an eigenvalue, are extracted sequentially so that they are uncorrelated with each other.

15.3.4. Derivation of components

We can derive the new variables (components) with matrix algebra in two ways. We can use a spectral decomposition of a p by p square matrix of association among variables (e.g. **SSCP**, **C** or **R** matrices) or we can use a singular value decomposition of the n by p original data matrix. The two approaches produce equivalent results if there is a match between the association matrix used and the standardization of variables in the data matrix. One of the biggest problems facing biologists trying to become familiar with multivariate statistical techniques is the bewildering range of terminology, with different textbooks using different terms for the same property and also different labels for the relevant matrices. We have tried to summarize these two approaches for extracting components from a multivariate data set in Box 15-1, following the terminology of Jackson (1991) where possible.

The usual derivation of components is from an association matrix of covariances or correlations between variables (Box 15-1). This is sometimes termed an *R*-mode analysis and we can calculate scores for the derived variables (components) for each object (Jackson 1991, Ludwig & Reynolds 1988). We could also derive components from matrices representing covariances or correlations between objects and the derived variables (components) are linear combinations of the objects. We can calculate component scores for each variable and this is termed a *Q*-mode analysis. These two sets of component scores are related via matrix algebra and we can obtain component scores for objects from the eigenvectors of the variables and vice-versa (Jackson 1991). In practice, *Q*-mode analyses comparing objects are more commonly based on dissimilarity measures (Box 15-2; Figure 15-2; Section 15.4).

The calculation of eigenvectors and their eigenvalues for new derived variables (components) from a multivariate data set is fundamental to canonical correlation analysis, principal components analysis and correspondence analysis (Chapter 17). If our data set contains groups, we can extract the components in a way that maximizes the between group differences and this is the basis of multivariate analysis of variance and discriminant function analysis (Chapter 16).

15.4. Multivariate distance and dissimilarity measures

The methods described in the previous section deal with multivariate data sets by rearranging the variance based on the association (covariances or correlations) between the variables (*R*-mode analyses). Another approach to multivariate data analyses (*Q*-mode analyses) is based on a measure of similarity or dissimilarity, sometimes termed a resemblance measure (Ludwig & Reynolds 1988), between objects.

Similarity indices measure how alike objects are, e.g. how similar sampling units are in terms of species composition or how alike specimens are in morphology. Dissimilarity indices measure how different objects are and should represent multivariate distance - if each variable is represented by an axis (or dimension) then multivariate distance is how far apart the objects are in multidimensional space. These dissimilarity indices are also called distances and are calculated for every possible pair of objects. There are numerous dissimilarity indices and the preferred ones are those that most closely represent biologically meaningful differences between objects. Particular difficulties arise when variables are measured on very different scales or when some of the variables include zero values, e.g. the variables are abundances of species of organisms and many objects have zero abundance for one or more species.

We usually represent the dissimilarities between objects as a dissimilarity matrix, converting an *n* rows by *p* columns data matrix to an *n* rows by *n* columns dissimilarity matrix. Like the covariance and correlation matrices described in Section 15.2, dissimilarity matrices are identical above and below the diagonal, which will be zeros indicating zero dissimilarity between an object and itself.

15.4.1. Dissimilarity measures for continuous variables

There is a broad range of measures of dissimilarity between objects based on continuous variables (see Digby & Kempton 1987, Faith *et al*. 1987, Legendre & Legendre 1998, Ludwig & Reynolds 1988). Their proliferation is partly due to the requirement by ecologists for measures of dissimilarity between sampling units in species composition that best represent underlying environmental gradients. We illustrate some of the commonly used measures in Box 15-2 and describe them briefly below. Legendre & Legendre (1998) provide a very thorough coverage.

Euclidean:

This is based on simple geometry as a measure of the distance between two objects in multidimensional space. It is the square root of the sum, over all the variables, of the square of the difference between the values of each variable for the two objects. It is only bounded by zero for two objects with exactly the same values for all variables and has no upper limit, even when two objects have no variables in common with positive values.

City block or Manhattan:

This is the sum (across variables) of the absolute differences in the value of each variable between two objects. It has properties similar to Euclidean distance and will be dominated by variables with large values.

Minkowski:

Euclidean and City block are both versions of the more general Minkowski metric. Some software will, by default, "normalize" both measures by dividing by the sample size, i.e. the number of variables that contribute to the distance measure. This is only relevant if you wish to compare dissimilarities between data sets with different numbers of variables.

Canberra:

This is the City block measure above, except that the difference between objects for each variable is divided by the sum of the variable values in the two objects before summing across variables. To ensure it has an upper limit of one, we standardize it by the number of variables that are greater than zero in both objects, e.g. the number of species present in at least one of the objects. This standardization is not always provided in texts (e.g. see Digby & Kempton 1988). The Canberra measure is less influenced by variables with very large values (Krebs 1989) than the City block measure.

Bray-Curtis:

Developed by botanists in Wisconsin, this is also a modification of the Manhattan measure where the sum of differences between objects across variables is standardized by the sum of the variable values across objects, also summed across variables. Equivalently, it can be calculated as one minus twice the sum of the lesser value of each variable when it is greater than zero in both objects, standardized by the sum of the values of all variables in both objects. It ranges between zero (same variables and values in both objects - completely similar) and one (no variables in common with positive values - completely dissimilar) and is sometimes called percent dissimilarity (when expressed as a %; Ludwig & Reynolds 1988) or Czekanowski's coefficient. It is well suited to species abundance data because it ignores variables that have zeros for both objects (joint absences). Its value is determined mainly by variables with high values (e.g. species with high abundances; see Krebs 1989) because these variables are likely to be more different between the objects.

Kulczynski:

This complicated measure, also termed the quantitative symmetric measure, was introduced to biologists by Faith *et al*. (1987). Like Bray-Curtis, it ranges between zero and one and has similar properties.

Chi-square:

This dissimilarity measure, implicit in some multivariate analyses (e.g. correspondence analysis – Chapter 17), is only applicable when the variables are counts, such as species abundances. It is based on differences between objects in the proportional representation of each species, also adjusted for species totals.

15.4.2. Dissimilarity measures for dichotomous (binary) variables

Another group of dissimilarity coefficients has been developed for variables measured on a binary scale (e.g. presence and absence). Let a be the number of variables with non-zero values in both objects, b is the number of variables with non-zero values in object 1 and c is the number of variables with non-zero values in object 2. A simple measure of dissimilarity between two objects is Jaccard's coefficient:

$$1 - \frac{a}{(a+b+c)} \tag{15.3}$$

A slight modification is Sorenson's coefficient which replaces a by $2a$. Sorenson's coefficient is identical to the Bray-Curtis measure for dichotomous variables.

15.4.3. General dissimilarity measures for mixed variables

Gower (1971) introduced a general dissimilarity measure that is useful for situations that include a mixture of continuous and categorical variables:

$$\frac{\sum_{j=1}^{p} w_{12j} s_{12j}}{\sum_{j=1}^{p} w_{12j}} \tag{15.4}$$

In equation 15.4, s_{12j} is the similarity between objects 1 and 2 based on variable j and w_{12j} equals one if the two objects can be compared for variable j and zero if they can't. So Gower's coefficient is "an average over all possible similarities" (Cox & Cox 1994) for objects 1 and 2. Gower's coefficient handles a mixture of variable types by calculating similarity for each variable separately (using appropriate coefficients for binary and

continuous variables), then averaging those similarities. With all continuous variables, Gower's coefficient becomes (Cox & Cox 1994, Faith et al. 1987):

$$\sum_{j=1}^{p} \frac{\left|y_{1j} - y_{2j}\right|}{(\max_j - \min_j)} \qquad (15.5)$$

15.4.4. Comparison of dissimilarity measures

One characteristic of dissimilarities is whether they meet the criterion of being metric. A dissimilarity coefficient is metric if the dissimilarity between objects 1 and 2 is less than the sum of the dissimilarities between objects 1 and 3 and 2 and 3. This means that it is possible to construct a triangle whose sides match the three dissimilarities between three objects. Dissimilarity measures that meet the condition of being metric are commonly termed dissimilarity metrics. Not all dissimilarity measures are metric, e.g. Minkowski and chi-square are but Bray-Curtis is not. If the dissimilarity is to be used in linear models (see Chapter 18), then being metric is important but otherwise the choice of dissimilarity measure for the analyses we describe in Chapter 18 is not usually based on whether it is metric or not.

Which of the many dissimilarity measures to use depends on the purpose of the analysis, the nature of the data and is closely linked to standardizations discussed in Section 15.6. When variables are measured on similar scales and have no zero values, Euclidean, City block or Canberra are good measures of dissimilarity between objects. If the scales of measurement are not consistent for different variables (e.g. the leaf characteristics from Reich *et al*. 1999), then the data need to be standardized before calculating these dissimilarities. Where the variables are species abundances (i.e. counts), an ideal dissimilarity coefficient should reach a constant maximum value when two sampling units have no species in common (i.e. it doesn't classify sampling units as similar because they have no species in common). Bray-Curtis, Kulczynski and Canberra meet this criterion whereas Euclidean and Chi-square do not. For this and other reasons, Faith *et al*. (1987) recommended the Bray-Curtis or Kulczynski coefficients for comparing objects when the variables are abundances of different species, as simulations showed these measures best matched ecological gradients. The suitability of some multivariate analyses for certain types of data is closely linked to the chosen or implicit dissimilarity measure that is used; we will discuss this further in the next two chapters.

For binary data, Kent & Coker (1992) argued that Sorenson's coefficient is preferred because it weights species (variables) in common higher than species absences (see also Krebs 1989). Remember that Sorenson's coefficient is the same as the Bray-Curtis measure with binary variables.

The general Gower dissimilarity measure is particularly useful when the data are a mixture of binary and continuous variables or when there are missing observations (but see Section 15.9.2), although Faith *et al*. (1997) showed that the version for continuous variables did not represent underlying ecological distances very well.

15.5. Comparing distance and/or dissimilarity matrices

Biologists often wish to test whether two or more matrices, or at least their corresponding elements, are correlated with each other. Such questions are particularly relevant when we are dealing with distance and/or dissimilarity matrices. For example, Sokal & Rohlf (1995) compared the matrix of genetic distances between ten villages of the Yanomama Amerindians in South America to the matrix of geographic distances between the villages. Fortin & Gurevitch (1993) emphasized the importance of examining spatial structure in field experiments, where one matrix might be differences in response of experimental units and the other might be the actual physical distances between the units.

Mantel's test is used for testing null hypotheses about correlations between matrices. It uses a randomization procedure (Chapter 3) to test whether the relationship between two matrices is more different than we would expect by chance (Manly 1997, Sokal & Rohlf 1996). We simply calculate the correlation coefficient between the corresponding elements of the two matrices, using only the lower (or upper) half of each matrix because they are symmetrical. However, the dissimilarities or distances within each matrix are not independent of each other (the dissimilarity between object 1 and 2 uses some of the same information as the dissimilarity between object 1 and 3 etc.). This is why we use a randomization test (Chapter 3) for the H_0 that the correlation between the two matrices is no different than we would expect by chance. Other statistics equivalent to the correlation coefficient for testing the H_0 in Mantel's test include Z (the sum of the products of the corresponding elements in the two matrices) and the regression coefficient (slope) for elements in one matrix regressed against elements in the other matrix. If the distances in the two matrices are standardized to zero mean and unit variance (Chapter 4), the values of the correlation coefficient, the regression slope and Z/m, where m is the number of elements in each matrix, will be the same (Manly 1997).

McCue *et al*. (1996) described genetic structure of a rare annual plant (*Clarkia springvillensis*) in California. They identified eight subpopulations and calculated Cavalli-Svorza genetic distances between subpopulations from isozyme analysis of tissue samples. They had two distance matrices – one for genetic distances between subpopulations and one for geographic distance (in metres) between subpopulations. The correlation coefficient between the two matrices was 0.632 with a randomization P value of 0.032 and we would conclude that there is a statistically significant positive relationship between genetic and geographic distance for populations of *C. springvillensis*). Note that in this example, the subpopulations were either really close (<500 m) or around 8000 m apart so our interpretation of the relationship between genetic and geographic distance is constrained by the absence of data for separations between 500 and 8000 m.

The correlations can be extended to more than two matrices, using an analogue of the coefficient of multiple correlation (r^2) and partial correlations, called partial Mantel's test (Manly 1997). For example, Sklenar & Jorgensen (1999) measured floristic similarity between six mountains in Ecuador using Sorenson's index for presence-absence data. They used Mantel's test to show that there was a significant correlation between floristic similarity and differences in sampling intensity and they used a partial Mantel's test to test for a correlation between floristics and distance, holding sampling intensity constant.

15.6. Data standardization

Transformations, which change the scale of measurement of the data, were discussed in Chapter 4 in relation to meeting the normality assumption of parametric analyses and the homogeneity of variance assumption of most of these analyses. Transformations are particularly important for multivariate procedures based on eigenanalysis (e.g. principal components analysis - see Chapter 17) because covariances and correlations measure linear relationships between variables. Transformations that improve linearity will increase the efficiency with which the eigenanalysis extracts the eigenvectors.

Transformations such as log or $\sqrt{}$ will normalize positively skewed data and also reduce the influence of variables with high values (e.g. very abundant species) in multivariate procedures based on dissimilarity indices (Digby & Kempton 1987). Clarke & Warwick (1994) argued that 4^{th} root transformations should always be used for species abundance data before calculating dissimilarities to reduce the influence of very abundant species. One difficulty with this approach is that the effect of the transformation will depend on the underlying distributions of the variables (e.g. species) and therefore the degree of reduction of influence of very abundant species will be inconsistent. Cao et al. (1999) also had concerns about log transformation of water quality variables, pointing out that this

transformation “indiscriminately increases the importance of a low range across all variables”.

Standardizations work slightly differently from transformations by adjusting the data so that means and/or variances or totals for each variable are the same. For example (Table 15-5):

- Centering the data subtracts the variable mean from each observation for each variable, resulting in all variables having a mean of zero. Spectral decomposition of a covariance matrix extracts components from centered data.
- Standardizing the data divides the centered observations by the standard deviation for each variable, resulting in all variables having a mean of zero and a standard deviation (and variance) of one. Spectral decomposition of a correlation matrix extracts components from standardized data.
- Data can also be standardized so that each observation is expressed relative to the maximum value of that variable across all objects. This standardization results in observations being expressed as a proportion of the largest value for a variable, and is basically standardization based on the range within a variable.
- Cao et al. (1999) proposed a novel standardization for water quality data, whereby each variable is standardized in relation to the water quality standard of that variable and its range. Although acknowledging problems with their new standardization, they argued that it does allow natural variability in each variable to contribute to the results of a multivariate analysis.

These standardizations of variables are important if variables are measured in very different units or scales, because otherwise those variables with larger values or larger variances will often be more influential on the results of an analysis than variables with smaller values or smaller variances. Standardization of variables is essential if the variables are measured in very different units. For species abundances, such standardizations make all species have similar "importance" and thus "avoids a strong weighting by a few highly abundant species" (Ludwig & Reynolds 1988, p. 215). Without this standardization, rare species are often making little contribution to dissimilarities - of course, this may be the most biologically sensible interpretation.

In the same way that variables could be standardized, objects (e.g. sampling units) can also be standardized so the value for any variable for each object is expressed relative to the maximum value for that object in the whole data matrix. For species abundance data, this standardization is very important if the size of the sampling unit, and hence the total number of individuals, varies because it removes any effect of different total abundances in different sampling units, i.e. all sampling units are considered to have the same total abundance across all species.

Finally, converting abundance data to presence and absence might be considered an extreme combination of transformation and standardization. There are specific dissimilarity measures for such binary data (see Section 15.4.2).

It is often useful to analyze the same data with different standardizations, particularly in ecological research. For example, comparing the results of an analysis using raw data with one using sample-standardized data will indicate what influence different total abundances in samples have. Raw data versus species-standardized data will illustrate what influence the most abundant species have (simply leaving out different combinations of rarer species will provide similar information). Finally, to remove all effects of abundance, we can analyze just presence-absence data.

15.7. Standardization, association and dissimilarity

Measures of association between variables described in Section 15.2 have implicit standardizations (see also Chapter 5). Covariances measure the linear relationships between centered variables whereas correlations measure the linear relationships between standardized (zero mean and unit variance) variables. The choice of association matrix on which to base subsequent multivariate analyses (Chapter 17) depends on whether differences in variances between variables represent important biological information that you don't wish to lose. Standardizations are also important for dissimilarity measures. Some dissimilarity measures are implicitly standardized and are unaffected by data standardizations (Faith *et al*. 1987). Some become identical after data standardization, e.g. Bray-Curtis, Kulczynski and City block are identical for count data if objects are standardized to the same total abundance. Others, e.g. Bray-Curtis and Kulczynski, produce nonsensical values when standardization is to zero mean (centering) or zero mean and unit variance (because of negative values). Standardizing by the range is a better option for these measures if you wish to reduce the influence of very abundant variables (e.g. species).

15.8. Multivariate graphics

Many of the exploratory data analysis techniques described in Chapter 4 are very applicable to multivariate data sets. In particular, describing distributions and checking for outliers for each variable separately with boxplots and examining bivariate relationships between variables with scatterplot matrices (SPLOMS) are always useful.

We may also wish to represent each observation or object in symbolic form, so that each symbol describes the relative value of all of the variables. A number of approaches have been developed to represent the different variables in a single "icon". The best known method is using Chernoff faces, where different features of the face represent different variables (Chernoff 1973; see also Everitt & Dunn 1991, Flury & Riedwyl 1988). These plots have been criticized, primarily because of the difficulty of rationally assigning variables to face features (Cox 1978), but they also have their supporters (Everitt & Dunn 1991, Flury & Riedwyl 1988). We illustrate these face plots with the Wisconsin forb data from Reich *et al*. (1999) in Figure 15-3, for both raw and standardized data. The differences between species are more noticeable for standardized variables, especially nose features representing mass-based and area-based photosynthetic capacity. Nonetheless, practice on known data sets is required to become familiar with recognizing similar and dissimilar faces.

An alternative, less "cartoonish", icon plot is to represent each object with a star, where each variable is represented by a point on the star, and the value of the variable is indicated by how far the point is from the centre. There are no limits to the number of points, and therefore variables, for each star although the stars become difficult to interpret when there are too many variables. The difference between raw and standardized variables is often very obvious on star plots. In Figure 15-4, we again illustrate the Wisconsin forb data from Reich *et al*. (1999). It is clear that *S. purpurea* is very different from the remaining species and *S. terebinthinaceum*, *P. peltatum*, *B. leucophaea* and *T. grandiflora* have larger values for leaf diffusive conductance at photosynthetic capacity, indicated by the extension of their stars to the left.

Finally, a very common method of graphing relationships between objects is to use a scatterplot where the axes represent the new derived variables from an eigenanalysis. These plots are common in the analyses described in Chapters 16 and 17, especially discriminant function analysis, principal components analysis and correspondence analysis. Alternatively, we can graphically represent a dissimilarity matrix between objects in a scatterplot, the basis of multidimensional scaling described in Chapter 18. Both types of plots are used especially by ecologists to represent the relationships between sampling or experimental units based on species composition, where they are termed "ordination" plots, the term ordination being derived from attempts to order units along some environmental

gradient (Digby & Kempton 1988). Ordination is not a term familiar to most statisticians, or even non-ecological biologists, so we will call such plots of objects "scaling plots".

15.9. Screening multivariate data sets

In Chapter 4, we emphasized the importance of exploratory data analyses before proceeding with univariate statistical procedures, especially those with distributional assumptions. We also pointed out that unusual values (outliers) can have very influential effects on the conclusions from a statistical analysis, both in terms of estimation and hypothesis testing, and checking for outliers is an important precursor to any formal analysis. The need for exploratory screening of data is even more important for multivariate data sets because their complexity means that visual inspection of the raw data is likely to miss unusual patterns or observations. Additionally, the issue of missing observations is much more critical for the analyses we will describe in the next three chapters.

All of the univariate procedures we described in Chapter 4, especially graphical explorations (see previous section), can and should be used for multivariate data sets. In this section, we will focus on two particular issues: detecting multivariate outliers and dealing with missing observations.

15.9.1. Multivariate outliers

We discussed in Chapter 4 how unusually extreme values can influence the outcome of a statistical analysis. Multivariate outliers are more difficult to detect because they may not be univariate outliers for any of the individual variables (Jobson 1992). Additionally, outliers are often defined as large departures from a fitted statistical, usually linear, model to our data. For example, an observation may be an outlier from a fitted regression model (Chapters 5 & 6) and may have undue influence on the estimates of model parameters and tests of hypotheses about these parameters. In contrast, many of the multivariate techniques we will introduce in the next three chapters are more descriptive in nature, although new summary variables are often derived and can be used as response or predictor variables in subsequent linear models.

A multivariate outlier is an object with an unusual pattern of values for the variables (Tabachnick & Fidell 1996) and can be detected by measuring its distance, in multivariate space, from the centroid (Figure 15-1). The square of this distance (d_i^2 for object *i*) is called Mahalanobis distance (see Flury & Riedwyl 1988, Jackson 1991, Jobson 1992 for computational details) and is provided by most software in one or more of the multivariate analysis routines. If multivariate normality holds, the d_i^2 follow a c^2 distribution with *p* (the number of variables) df (Manly 1994) so we can test for outliers, possibly using a strict significance level like 0.001 (Tabachnick & Fidell 1996).

Dealing with univariate outliers has been described in Chapter 4. The options for multivariate outliers are similar. If we decide that an object has such an unusual pattern of values for one or more variables that it is unlikely to be part of the population of objects we wish to describe or make inferences about, then we might delete that object from the analysis. Transformations of the variable(s) can also reduce the influence of outliers if they are extreme values in a positively skewed distribution.

15.9.2. Missing observations

Occasionally, we will have missing observations in our data set, i.e. no value was recorded for one or more variables for one or more objects. The approaches for dealing with missing observations depend on the missing data mechanism, as introduced in Chapter 4 (see also Heitjan 1997, Little & Rubin 1987, Roth 1994). If the probability that an observation is missing is independent of the observed and missing values, the missing observations are termed missing completely at random (MCAR). This implies that the missing observations

are a random subset of the data. The probability that an observation is missing might not depend on the unobserved missing value but be dependent on the values of the other variables for that object. For example, the pattern of missing data may depend on the group in which the object occurs, where another variable classifies objects into groups. This termed missing at random (MAR). Finally, the missing values might be non-ignorable because whether an observation is missing depends on its value.

Consider the data set from Lovett *et al*. (2000) and imagine that one stream was missing a value for concentration of H^+. If the value is missing because of a random malfunction of a meter or a mistake by a researcher who forgot to write the value down then this observation might be MCAR. Our experience is MCAR is a common missing data mechanism in ecological sampling programs. If the value is missing because the stream was at a high altitude and weather conditions precluded access, then the observation might be MAR because the value of another variable (elevation), but not the unobserved H^+ value, determines the probability of it being missing. Finally, if the value is missing because the original H^+ reading was so high (e.g. Winnisook Creek) that the researcher assumed that the reading was a mistake and ignored it, the missing value is clearly non-ignorable. This situation is more common in situations when the observations depend on responses from subjects, such as in marketing surveys or clinical trials, although studies on animal behavior may suffer from this type of non-response. MCAR and MAR are much easier to deal with.

Basically, there are three approaches to dealing with missing observations (Little & Rubin 1987, Roth 1994). Our objective in this section is simply to make biologists aware that there are alternatives to simply "omitting whole rows of data", although some of the methods are sophisticated and usually require advice from statisticians experienced with their use. It is important to remember that avoiding missing data is the best solution because all of the alternatives are imperfect. We illustrate the results from some of the methods for dealing with missing observations in using a subset of the data from Reich *et al*. (1999). Our emphasis is not on the calculations, as these require appropriate software, but on the interpretation of the different methods.

Deletion

The simplest approach is to delete the entire object that has the missing value. This may be an appropriate strategy when the proportion of objects with missing values is low and the pattern is MCAR. It does result in loss of information because the non-missing values of variables for the object with the missing value are also excluded from the analysis. This is sometimes termed listwise deletion and is often the default for multivariate analyses in statistical software. If the analysis is based on pairwise associations between variables (e.g. correlations), an alternative is to use pairwise deletion. Here an object is only excluded for the calculation of the association between the two variables for which one value is missing but not excluded for the calculation of associations between other variables. This is the preferred deletion strategy when pairwise associations are the basis for the analysis.

Imputation

Imputation involves replacing (substituting) the missing values with some estimate of what the values might have been. There have been three common methods for imputing missing observations. The first is to replace the observation with the mean value of the variable calculated from the non-missing observations. Unfortunately, this tends to result in an underestimate of the true variance for that variable because these means do not contribute to the sum of squared deviations (Roth 1997). The second is to use a regression model to predict the imputed observation from other variables in the data. For example, we could determine which variable has the highest correlation with the variable with missing values from the complete objects and develop a regression model where the variable with missing values is the response variable and the other variable is the predictor. For the object with the missing value, the observed value of the predictor could then be used to predict the missing

value from this regression model. Alternatively, we could use two or more predictors in a multiple regression model. Generalized linear models could be used if the assumption of normal error terms for the regressions was untenable or even generalized additive models if the shape of the relationship between the variables is not linear, although we have not seen either of these used in practice. Finally, hot-deck imputation simply replaces the missing value with the actual value from an object with similar characteristics (Roth 1997).

There are two main difficulties with these imputation methods. The first is that the imputed values are not independent of the observed data for a given variable and the precision (variances and standard errors) of the estimates of parameters based on these imputed values is generally underestimated. The second problem is that imputing a single value provides no indication of the effect that different imputed values have on the estimation of the relevant parameter (e.g. correlation), i.e. no measure of imputation uncertainy (Little 1999). Rubin (1987) developed a method termed multiple imputation as a solution to the second problem (see also Schafer 1999). Multiple imputation basically imputes a range of values for each missing observation, these values being simulated from a specific distribution for the missing values. The complete data sets (observed and imputed values) are then analyzed in the usual manner. The estimate of any parameter is simply the mean of estimates from the analyses of the imputed data sets. The standard error of this average estimate includes both the variance between imputations and the variance within each data set. Multiple imputation is clearly a sensible approach and a considerable improvement over single imputation, giving us some indication of how different imputed values affect the outcome of our analysis. The really tricky bit is developing the distribution of values from which the multiple imputations are derived. Rubin (1987) recommended a Bayesian strategy whereby the posterior distribution of missing values is conditional on the prior distribution of observed values, although the computations are complex (Schafer 1999). Multiple imputation routines are not readily available in commonly used statistical software but specialist products do exist and macros for some programs are available (Rubin 1996 and references therein).

Maximum likelihood and EM

A different approach is to use maximum likelihood (ML) techniques to estimate the parameters of interest (e.g. means, correlation coefficients) from the observed, incomplete data (Little & Rubin 1987). Basically we use the distribution of the observed data and the conditional distribution of the pattern of missing data given the observed data. The likelihood function for any parameter can be complex with missing data so Little & Rubin (1987) also proposed methods based on factoring the likelihoods. The likelihood for a given parameter is decomposed into the sum of the likelihoods of distinct parameters given complete subsets of the data. These ML methods can estimate the missing observations once the parameters are estimated but do not use imputed values to estimate the parameters.

A combination of imputation and ML estimation is the Expectation-Maximization (EM) algorithm. This is an iterative procedure whereby the missing values are imputed, the parameters are estimated by ML, the missing values are re-estimated and imputed, the parameters re-estimated by Ml etc. until convergence of the likelihood of the parameter given the observed data is achieved. Technically, the missing values are not directly imputed using the EM method, but some function of the missing data like a predictive distribution is incorporated into the likelihood function (Little & Rubin 1987, Schafer 1999). The EM algorithm is now available in some commonly used statistical software. Multiple imputation may be more robust than EM methods for small data sets (Schafer 1999). Both straight ML and the EM method require the missing data to be at least MAR.

15.10. General issues and hints for analysis

General Issues

- Variation within, and linear relationships between, two or more variables can be summarized with a sums of squares and cross products matrix (raw data), covariance matrix (centered data) or a correlation matrix (standardized data).
- Spectral decomposition of one of these matrices produces new derived variables (components), extracted so the first explains most of the original variation, the second most of what is left etc. and so that the new variables are uncorrelated with each other. Equivalent results are obtained from a singular value decomposition of the original data matrix, appropriately standardized.
- These new variables are linear combinations of the original variables and the coefficients (summarized as an eigenvector) indicate the contribution of each original variable to the new variable.
- Differences between pairs of objects are measured with dissimilarities that are based on the sum of the differences for each variable between objects, often standardized so they range between zero and one.
- For measurement variables, either Euclidean or one of its modifications (City block or Canberra) are reliable dissimilarity measures, usually based on standardized data. For species abundances (counts with possible zero values), Bray-Curtis or Kulczynski are recommended.
- Graphical representations of multivariate data are available. SPLOMs display pairwise bivariate relationships and icon plots (Chernoff faces or stars) visually represent objects in terms of the relative values for the variables.
- The default for handling missing data with most software is to omit whole objects. Other approaches are generally preferred unless the sample size is large and the observations are missing completely at random.

Hints for analysis

- Before extracting components or determination of dissimilarities between objects when variables are measured in different scales or units, some type of standardization (based on standard deviation or range) is recommended.
- For species abundance, i.e. count, variables, different standardizations can provide useful comparative information. Standardizing objects to equal totals corrects for different sized sampling units, standardizing species to equal totals means that the most abundant species do not dominate the dissimilarity measure.
- Some standardizations can result in Bray-Curtis and Kulczynski dissimilarities not being bounded by one; standardize by range rather than by standard deviations when using these measures.
- We prefer standardizations to transformations for reducing the influence of variables with large values, although transforming variables may be relevant to improve linearity or if univariate analyses on the same variables also require transformation.

Box 15-1 Deriving components (modified from Jackson 1991):

There are two different strategies for extracting eigenvectors (components) and their eigenvalues from multivariate data set of *n* objects by *p* variables. First, we can use a spectral decomposition of a *p* by *p* association matrix between variables. Second, we can use a singular value decomposition (SVD) of a *n* by *p* data matrix, with variables standardized as necessary. The SVD is more generally applicable (see Chapter 17) although most biologists are more familiar with obtaining eignevectors and eigenvalues from a covariance or correlation matrix.

Consider the matrix (**Y**) of raw data from Clevenger & Waltho (2000) who recorded the numbers of people on bicycles, horses and on foot for eleven underpasses also used by wildlife in Alberta, Canada.

	Raw			Centered		
Underpass	Bicycle	Horse	Foot	Bike	Horse	Foot
1	0	6	7	-118.727	-37.273	-55.364
2	5	3	45	-113.727	-40.273	-17.364
3	6	6	14	-112.727	-37.273	-48.364
4	21	5	20	-97.727	-38.273	-42.364
5	189	42	34	70.273	-1.273	-28.364
6	8	138	77	-110.727	94.727	14.636
7	462	186	129	343.273	142.727	66.636
8	19	12	80	-99.727	-31.273	17.636
9	595	58	241	476.273	14.727	178.636
10	1	10	10	-117.727	-33.273	-52.364
11	0	10	29	-118.727	-33.273	-33.364

Spectral decomposition

We will illustrate spectral decomposition of a matrix of associations between variables ($\mathbf{Y}'\mathbf{Y}$). This might be a matrix of variances and covariances, **C**, among *p* variables based on *n* objects (Table 15-3):

	Bike	Horse	Foot
Bike	44906.018		
Horse	7336.382	3862.018	
Foot	13084.709	2205.191	4903.655

Note that we could also use a correlation matrix. Basically, we derive two matrices, **L** and **U**, so that:

$$\mathbf{L} = \mathbf{U}'\mathbf{C}\mathbf{U}$$

U is a *n* by *p* matrix whose columns contain the eigenvectors (characteristic vectors), the coefficients of the linear combinations of the original variables. The elements of each eigenvector *k* are u_{jk}, the coefficient for the *j*th variable in the *k*th eigenvector. Note that we clearly need to have to some constraints imposed on the coefficients within each eigenvector, otherwise simply increasing the absolute sizes of the coefficients could increase the variance

explained by each new variable. The simplest and most commonly used constraint is to restrict the sum of squared coefficients to zero, i.e. $\sum_{j=1}^{p} u_{jk}^2 = 1$. Eigenvectors that are independent and scaled to unity are termed orthonormal. Additional scaling options for the eigenvectors are available to make the variances of the eigenvectors similar (Jackson 1991), e.g. $v_{jk} = \sqrt{l_k} u_{jk}$ so the eigenvectors are in a **V** matrix and $w_{jk} = u_{jk}/\sqrt{l_k}$ so the eigenvectors are in a **W** matrix.

L is a *p* by *p* matrix whose diagonal of contains the eigenvalues $l_1, l_2, \ldots l_k \ldots l_p$ (estimates of $l_1, l_2, \ldots l_k \ldots l_p$, the latent or characteristic roots) of **C**. The eigenvalues measure the variance explained by each of the eigenvectors. The number of eigenvalues is the same as the number of rows and columns in the covariance matrix and therefore the same as the number of original variables (*p*).

The matrix **L** for our example data set with the eigenvalues on the diagonal is:

50075.681	0	0
0	2592.350	0
0	0	1003.660

The trace of this matrix, the sum of its diagonal elements, is the sum of the variances of the original centred variables. The sum of the eigenvalues from an eigenanalysis of a sums-of-squares and cross-products matrix or a correlation matrix would equal the sum of the variances of the original variables or the centered and standardized variables respectively. The matrix **L** represents, therefore, a reorganization of the variances of the variables from the original data matrix. Each eigenvalue is associated with each eigenvector and it is clear that the eigenvectors are extracted in order of decreasing proportions of the total variance. We often convert these eigenvalues to %:

	1	2	3
Eigenvalue	50075.681	2592.350	1003.660
% of total variance	93.300	4.830	1.870

More formally, determination of the eigenvalues involves solving the characteristic equation:

$$\left|\mathbf{C} - l\mathbf{I}\right| = 0$$

where **I** is an identity matrix of equivalent dimensions to **C**. The resulting polynomial (*p*th degree) in *l* is used to obtain $l_1, l_2 \ldots l_p$.

Based on the three human activity variables (bicycle, horse, foot) for eleven underpasses in Alberta from Clevenger & Waltbo (2000), the matrix **U** is:

	1	2	3
Bicycle	0.945	0.160	0.284
Horse	0.164	-0.986	0.011
Foot	0.282	0.036	-0.959

Each column is an eigenvector (u_k where k = 1 to p), the values in the eigenvector representing the coefficients or weights for that linear combination of the original variables. For example, the linear combination comprising eigenvector 1 is:

(0.945)Bicycle + (0.164)Horse + (0.282)Foot

where the values of each variable are centered because we used the covariance matrix to extract the eigenvectors. These linear equations are often termed components or factors (Chapter 17) and represent new variables derived from the original variables. Note that each variable contributes differently to each component (different coefficients or weights) and that these coefficients will depend on the units of each variable and whether standardizations are used. These linear equations can be solved to produce a component score (z_{ik}) for each object or observation for each component. For example, the score for component 1 for underpass 1:

$$(0.945)(-118.727) + (0.164)(-37.273) + (0.282)(-55.364) = -133.946$$

Singular value decomposition (SVD)

The SVD of a *n* by *p* data matrix is based on the product of the characteristic vectors of a matrix of associations between variables, the characteristic vectors of a matrix of associations between objects and their characteristic roots (eigenvalues, which are the same for both association matrices). If **Y** is a matrix of centered data (comparable to the covariance matrix used above), then $\mathbf{Y'Y}$ is the covariance matrix between variables (matrix **C** above) and $\mathbf{YY'}$ is the covariance matrix between objects (note these would be SSCP matrices for raw data and correlation matrices for centered and standardized data). The characteristic roots (eigenvalues) of these two matrices are the same.

The SVD of **Y** is:

$$\mathbf{Y = ZL^{1/2}U'}$$

where **L** contains the eigenvalues, **U** is a *p* by *p* containing the eigenvectors of $\mathbf{Y'Y}$ as defined above and **Z** is a *n* x *p* matrix of eigenvectors of $\mathbf{YY'}$ and are also the principal component scores for objects scaled by the square root of the eigenvalues. Note that we now have the square root of the eigenvalues because we are dealing with the original variables rather than covariances or correlations (Jackson 1991). If **Y** contains raw data, then **L** and **U** will be the equivalent to that from the spectral decomposition of the SSCP matrix. If **Y** contains centered data, then **L** and **U** will be the equivalent to that from the spectral decomposition of the covariance matrix. If **Y** contains centered and standardized data, then **L** and **U** will be the equivalent to that from the spectral decomposition of the correlation matrix. Note that we can determine the original variables (centered and standardized if appropriate) from the matrix of component scores and vice-versa when all components are extracted.

The advantage of using SVD is that extraction of eigenvectors and their eigenvalues is a one step process and SVD can also be applied to association matrices that are not square, e.g. chi-square matrices from contingency tables as used in correspondence analysis (Chapter 17). The advantage of spectral decomposition is that the choice of matrix (e.g. covariance vs correlation) will automatically center or standardize the data. As most multivariate analyses require statistical software, we rarely have to make this choice in practice.

Box 15-2 Measures of dissimilarity between objects for continuous variables.

Consider two objects (i = 1 and 2), e.g. two sampling units, and a number of variables (j = 1 to p) recorded from each object, e.g. abundances of p species from each sampling unit. The same variables are recorded from each object (even if some variables have zero values for an object). First, we need a few definitions:

- y_{1j} and y_{2j} are the values of variable j in object 1 and object 2,
- $\min(y_{1j}, y_{2j})$ is the sum of the lesser value of each variable when it is greater than zero in <u>both</u> objects,
- p is the number of variables, and
- q is the number of variables that are zero for objects 1 and 2.

For example, y_{1j} and y_{2j} might be the abundances of species j in sampling units 1 and 2, $\Sigma\min(y_{1j}, y_{2j})$ is the sum of the lesser abundance of species j when it is present in both sampling units, p is the number of species and q is the number of species that are missing (zero values) from both samples. The formulae presented below are from Faith *et al*. (1987), except we present a more common version of the Canberra measure (see Digby & Kempton 1987) and correct their typographical error for chi-square.

Dissimilarity	Equation
Minkowski	$\left[\sum_{j=1}^{p}\left\|y_{1j}-y_{2j}\right\|^{l}\right]^{1/l}$
Euclidean (l = 2)	$\sqrt{\sum_{j=1}^{p}(y_{1j}-y_{2j})^{2}}$
City block (Manhattan: l = 1)	$\sum_{j=1}^{p}\left\|(y_{1j}-y_{2j})\right\|$

Canberra

$$\frac{1}{p-q}\sum_{j=1}^{p}\frac{\left|y_{1j}-y_{2j}\right|}{(y_{1j}+y_{2j})}$$

Bray-Curtis (Czekanowski)

$$1-\frac{2\sum_{j=1}^{p}\min(y_{1j},y_{2j})}{\sum_{j=1}^{p}(y_{1j}+y_{2j})}=\frac{\sum_{j=1}^{p}\left|y_{1j}-y_{2j}\right|}{\sum_{j=1}^{p}(y_{1j}+y_{2j})}$$

Kulczynski

$$1-\frac{\left[\frac{\sum_{j=1}^{p}\min(y_{1j},y_{2j})}{\sum_{j=1}^{p}(y_{ij})}+\frac{\sum_{j=1}^{p}\min(y_{1j},y_{2j})}{\sum_{j=1}^{p}(y_{ik})}\right]}{2}$$

Chi-square

$$\sqrt{\sum_{j=1}^{p}\frac{\left(y_{1j}\Big/\sum_{j=1}^{p}y_{1j}-y_{2j}\Big/\sum_{j=1}^{p}y_{2j}\right)^{2}}{\sum_{i=1}^{n}y_{j}}}$$

To illustrate these dissimilarity measures, we have calculated the dissimilarity between three species of Wisconsin forbs based on five leaf character variables from Reich *et al*. (1999). We have used the original variables and also variables centered and standardized to zero mean and unit variance:

Dissimilarity between	Euclidean	City block	Canberra	Bray-Curtis	Kulczynski
C. thalictroides vs *D. laciniata:*					
Raw data	238.355	82.500	0.231	0.217	0.212
Standardized data	2.992	1.143	NA	NA	NA
C. thalictroides vs *P. peltatum:*					
Raw data	121.005	34.300	0.111	0.104	0.100
Standardized data	1.337	0.498	NA	NA	NA
D. laciniata vs *P. peltatum:*					
Raw data	198.911	55.520	0.166	0.155	0.138
Standardized data	2.482	0.865	NA	NA	NA

Note that all measures show the same basic pattern, with the dissimilarity between *C. thalictroides* and *D. laciniata* the greatest and the between *C. thalictroides* and *P. peltatum* the least. Standardizing the variables to zero mean and unit variance doesn't change the relative dissimilarities although such a standardization cannot be applied to Canberra, Bray-Curtis and Kulczynski because they already include standardization as part of the calculation.

We also compared intact and impacted forest locations, based on the abundance of 15 species of bats, from Fenton *et al*. (1998). This data set allows us to include the chi-square measure, which requires integer values:

Dissimilarity between	Euclidean	City block	Canberra	Bray-Curtis	Kulczynski	Chi-square
Mana intact vs *Mana impacted:*						
Raw data	35.875	77.000	0.754	0.336	0.252	0.036
Standardized data	5.679	17.323	NA	NA	NA	NA
Range standardized data	2.720	8.255	0.835	0.770	0.435	0.428
Kanyati intact vs *Matusadona impacted:*						
Raw data	21.119	48.000	0.715	0.444	0.416	0.087
Standardized data	4.831	13.706	NA	NA	NA	NA
Range standardized data	2.390	6.663	0.792	0.719	0.703	0.491

Here the different dissimilarities produce different patterns. The intact vs impacted difference is greater for Mana than for Kanyati/Matusadona when measured with Euclidean, City block and Canberra but the reverse is true for Bray-Curtis, Kulczynski and Chi-square. None of the standardizations changed the relative sizes for any of the measures except for Bray-Curtis.

Box 15-3 Dealing with missing data

The data set on physiological variables for a range of plant species from different locations and functional groups from Reich *et al*. (1999) will be used to illustrate some of the methods for handling missing observations. We will use a subset of their data, trees from Venezuela, where there were 22 species (objects). There were five variables: specific leaf area (SLA), leaf nitrogen concentration (Leaf N), mass-based net photosynthetic capacity (A_{mass}), area-based net photosynthetic capacity (A_{area}) and leaf diffusive conductance at photosynthetic capacity (G_s). Five of the possible 110 observations were missing: SLA and A_{area} for *Eperua purpurea* and A_{mass}, A_{area} and G_s for *Micropholis maguirei*. We will assume these values are at least MAR and use listwise and pairwise deletion, regression imputation (using all other variables with complete data as predictor variables) and the EM algorithm to estimate means, standard deviations and pairwise correlations between variables. The EM algorithm converged in four iterations with –2(log-likelihood) of 650.85.

Means (standard deviations)

	SLA ($cm^2.g^{-1}$)	Leaf N ($mg.g^{-1}$)	A_{mass} ($nmol.g^{-1}.s^{-1}$)	A_{area} ($mmol.m^{-2}.s^{-1}$)	G_s ($mmol.m^{-2}.s^{-1}$)
Listwise	89.85 (24.04)	14.29 (4.71)	78.96 (55.23)	8.28 (3.68)	622.60 (535.76)
All Values	88.20 (24.62)	14.04 (4.68)	77.82 (54.09)	8.28 (3.68)	602.90 (529.94)
EM	88.15 (24.18)	14.04 (4.68)	74.49 (55.39)	8.01 (3.67)	580.68 (535.92)
Regression	89.85 (24.04)	14.29 (4.71)	78.96 (55.23)	8.28 (3.68)	622.60 (535.76)

Correlations based on deletions

	SLA		Leaf N		A_{mass}		A_{area}		G_s	
	List	Pair	List	Pair	List	Pair	List	Pair	List	Pair
SLA	1.000	1.000								
Leaf N	0.569	0.607	1.000	1.000						
A_{mass}	0.789	0.789	0.708	0.699	1.000	1.000				
A_{area}	0.550	0.550	0.684	0.684	0.931	0.931	1.000	1.000		
G_s	0.498	0.498	0.546	0.530	0.851	0.851	0.894	0.894	1.000	1.000

Note that only the correlation between SLA and Leaf N differs much between the two methods of deletion.

Correlations based on regression imputation and EM

	SLA		Leaf N		A_{mass}		A_{area}		G_s	
	Regress	EM	Regress	EM	Regress	EM	Regress	EM	Regress	EM
SLA	1.000	1.000								
Leaf N	0.601	0.602	1.000	1.000						
A_{mass}	0.789	0.795	0.714	0.719	1.000	1.000				
A_{area}	0.555	0.563	0.681	0.685	0.931	0.932	1.000	1.000		
G_s	0.503	0.511	0.541	0.546	0.853	0.854	0.893	0.895	1.000	1.000

There are differences between the estimated correlations based on the two methods but for these data, the differences are small.

Observed data with regression and EM imputed values (in bold):

SLA	Leaf N	A_{mass}	A_{area}	G_s
144.60	24.70	252.20	17.70	2272.00
114.30	17.90	159.30	13.80	889.00
126.40	16.50	115.50	9.10	597.00
105.40	16.40	140.40	12.80	975.00
78.10	16.90	111.50	14.00	1707.00
129.90	15.10	99.00	7.80	300.00
103.10	18.40	65.00	6.40	479.00
90.30	15.90	91.80	10.30	1009.00

82.80	6.80	46.50	5.60	490.00
75.20	7.80	47.20	6.20	693.00
86.60	8.60	34.70	4.00	321.00
82.60	10.70	52.20	6.50	411.00
82.00	17.70	67.20	8.20	381.00
67.80	9.30	38.80	5.70	241.00
76.80	15.00	44.90	5.90	329.00
67.30	13.00	53.80	8.00	378.00
86.20 (Regress)	15.20	55.10	**6.40 (Regress)**	209.00
87.10 (EM)			**6.32 (EM)**	
95.10	12.50	35.10	3.70	173.00
72.10	21.40	47.70	6.70	235.00
58.40	10.80	43.30	7.40	298.00
55.30	8.00	**4.76 (Regress)**	**4.26 (Regress)**	**114.03 (Regress)**
		20.94 (EM)	**5.01 (EM)**	**247.29 (EM)**
58.10	10.30	33.00	5.70	274.00

Note that the regression and EM imputed values are similar for *Eperua purpurea* (row 17) but very different for A_{mass} and G_s for *Micropholis maguirei* (row 21). The latter differences probably reflect the fact that only two predictor variables are available for this species for predicting the missing observations using a regression and the observed values for both of those variables are at the low end of the range for those variables. The EM imputed values are probably more reliable for this species.

$$\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \cdots & \cdots & y_{ij} & \cdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

	SLA ($cm^2.g^{-1}$)	Leaf N ($mg.g^{-1}$)	A_{mass} ($nmol.g^{-1}.s^{-1}$)	A_{area} ($mmol.m^{-2}.s^{-1}$)	G_s ($mmol.m^{-2}.s^{-1}$)
Caulophyllum thalictroides	425.0	58.2	254.0	5.9	134
Dentaria laciniate	297.0	53.0	432.0	14.2	227
Erythronium americanum	222.0	42.0	263.0	11.9	359
Silphium terebinthinaceum	133.0	14.4	175.0	13.4	615
Podophyllum peltatum	309.0	44.7	244.0	7.9	164
Baptisia leucophaea	106.3	35.9	159.0	15.0	481
Trillium grandiflora	357.0	51.6	209.0	5.8	499
Echinacea purpurea	128.5	15.0	122.9	9.8	480
Silphium integrifolium	116.3	16.6	116.0	10.0	478
Sanguinaria canadensis	321.0	53.6	255.0	7.9	208
Sarrachenia purpurea	78.1	11.4	22.8	2.9	144

Table 15-1 Raw data matrix of p variables (j = 1 to p) for n objects (i = 1 to n), illustrated with data from Reich et al. *(1999) for eleven species of Wisconsin forbs (objects) and five variables: SLA is specific leaf area, leaf N is leaf nitrogen concentration, A_{mass} is mass-based net photosynthetic capacity, A_{area} is area-based net photosynthetic capacity and G_s is leaf diffusive conductance at photosynthetic capacity.*

$$\begin{vmatrix} \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2 & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)(y_{i1}-\bar{y}_1) & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)(y_{i1}-\bar{y}_1) \\ \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2) & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2 & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)(y_{i2}-\bar{y}_2) \\ \cdots & \cdots & \sum_{i=1}^{n}(y_{ij}-\bar{y}_j)^2 & \cdots \\ \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{ip}-\bar{y}_p) & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)(y_{ip}-\bar{y}_p) & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)^2 \end{vmatrix}$$

	SLA	Leaf N	A_{mass}	A_{area}	G_s
SLA	144120.13				
Leaf N	19873.03	3335.73			
A_{mass}	87160.14	15162.00	112204.77		
A_{area}	-1290.94	-23.86	1635.93	148.98	
G_s	-97696.97	-14505.68	-50261.55	3412.31	301594.73

Table 15-2 Sums-of-squares-and-cross-products matrix between p variables (j = 1 to p) for n objects (i = 1 to n), illustrated with data from Reich et al. *(1999). Main diagonal entries are sums-of-squares, off diagonal entries are sums-of-cross-products. Variables defined in Table 15-1.*

$$\begin{bmatrix} s_1^2 & s_{12}^2 & \cdots & s_{p1}^2 \\ s_{12}^2 & s_2^2 & \cdots & s_{p2}^2 \\ \cdots & \cdots & s_j^2 & \cdots \\ s_{1p}^2 & s_{2p}^2 & \cdots & s_p^2 \end{bmatrix}$$

	SLA	Leaf N	A_{mass}	A_{area}	G_s
SLA	14412.01				
Leaf N	1987.30	333.57			
A_{mass}	8716.01	1516.20	11220.48		
A_{area}	-129.09	-2.39	163.59	14.89	
G_s	-9769.69	-1450.57	-5026.16	341.23	30159.47

Table 15-3 Variance-covariance matrix between p variables (j = 1 to p), illustrated with data from Reich et al. (1999). Main diagonal entries are variances, off diagonal entries are covariances. Variables defined in Table 15-1.

$$\begin{Bmatrix} 1 & r_{21} & \dots & r_{p1} \\ r_{12} & 1 & \dots & r_{p2} \\ \dots & \dots & 1 & \dots \\ r_{1p} & r_{2p} & \dots & 1 \end{Bmatrix}$$

	SLA	Leaf N	A_{mass}	A_{area}	G_s
SLA	1.00				
Leaf N	0.91	1.00			
A_{mass}	0.69	0.78	1.00		
A_{area}	-0.28	-0.03	0.40	1.00	
G_s	-0.47	-0.46	-0.27	0.51	1.00

Table 15-4 Correlation matrix between p variables (j = 1 to p), illustrated with data from Reich et al. (1999). All entries are Pearson correlations. Variables defined in Table 15-1.

	Unstandardized	Centered	Standardized
Caulophyllum thalictroides	58.20	22.16	1.21
Dentaria laciniate	53.00	16.96	0.93
Erythronium americanum	42.00	5.96	0.33
Silphium terebinthinaceum	14.40	-21.64	-1.18
Podophyllum peltatum	44.70	8.66	0.47
Baptisia leucophaea	35.90	-0.14	-0.01
Trillium grandiflora	51.60	15.56	0.85
Echinacea purpurea	15.00	-21.04	-1.15
Silphium integrifolium	16.60	-19.44	-1.06
Sanguinaria canadensis	53.60	17.56	0.96
Sarrachenia purpurea	11.40	-24.64	-1.35
Mean	36.04	0.00	0.00
Standard deviation	18.26	18.26	1.00

Table 15-5 Comparison of unstandardized, centered (zero mean) and standardized (zero mean and unit variance) observations for leaf N concentration for the eleven species of Wisconsin forbs from the study by Reich et al. *(1999).*

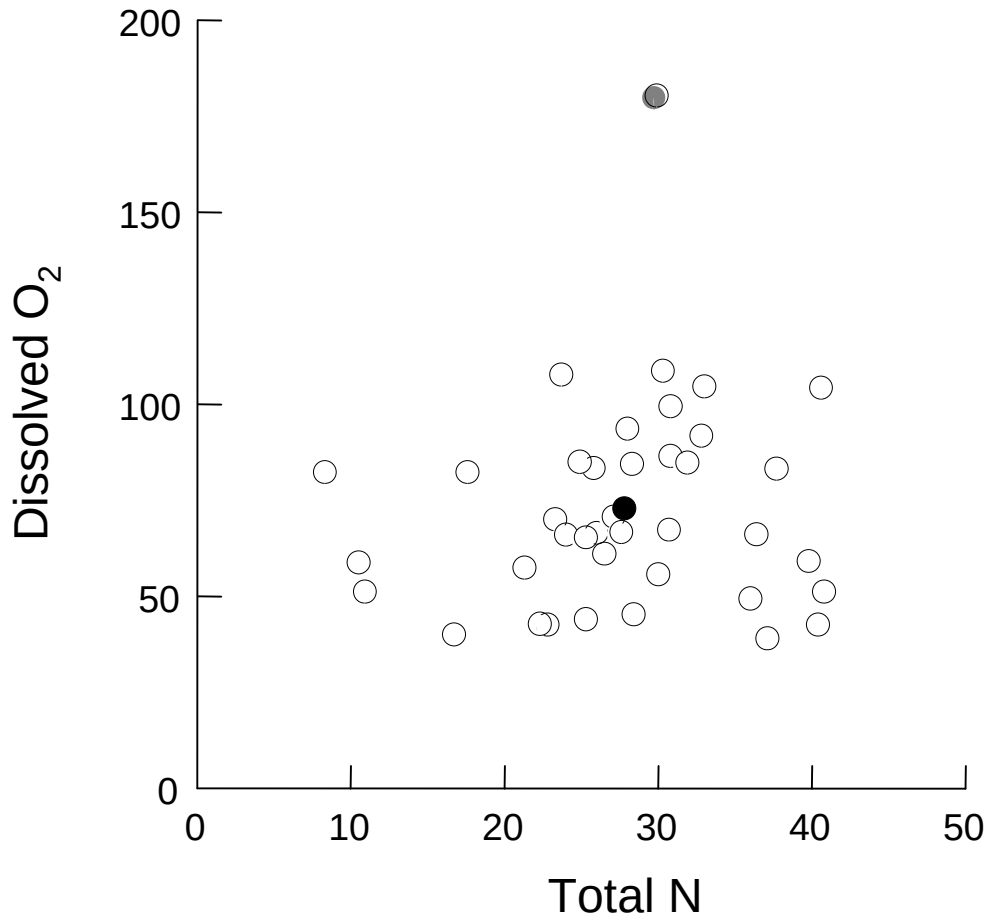
Captions to Figures

Figure 15-1 Scatterplot of dissolved oxygen against total nitrogen for 39 streams from Lovett et al. (2000). The centroid, the point represented by the mean of dissolved oxygen and total nitrogen is filled. In this example, one object (grey fill) is an outlier for dissolved oxygen and also a multivariate outlier.

Figure 15-2 Distinction in initial steps between R- and Q-mode analyses. A n rows by p columns data matrix is converted to a p by p matrix of associations between variables (e.g. correlations) or a n by n matrix of dissimilarities between objects.

Figure 15-3 Chernoff face representation of the eleven species of Wisconsin forbs for five leaf characteristics based on raw data (a) and standardized data (b) from Reich et al. *(1999). The features of the Chernoff faces are curvature of mouth for specific leaf area, angle of brow for leaf nitrogen concentration, width of nose for mass-based net photosynthetic capacity, length of nose for area-based net photosynthetic capacity, and length of mouth for leaf diffusive conductance at photosynthetic capacity. The species are, from left to right and row by row: Caulophyllum thalictroides, Dentaria laciniate, Erythronium americanum, Silphium terebinthinaceum, Podophyllum peltatum, Baptisia leucophaea, Trillium grandiflora, Echinacea purpurea, Silphium integrifolium, Sanguinaria canadensis, Sarrachenia purpurea.*

Figure 15-4 Star plot face representation of the eleven species of Wisconsin forbs for five leaf characteristics based on raw data (a) and standardized data (b) from Reich et al. *(1999). The features of the stars are, clockwise from the top, specific leaf area, leaf nitrogen concentration, mass-based net photosynthetic capacity, area-based net photosynthetic capacity, leaf diffusive conductance at photosynthetic capacity. The species are, from left to right and row by row: Caulophyllum thalictroides, Dentaria laciniate, Erythronium americanum, Silphium terebinthinaceum, Podophyllum peltatum, Baptisia leucophaea, Trillium grandiflora, Echinacea purpurea, Silphium integrifolium, Sanguinaria canadensis, Sarrachenia purpurea.*

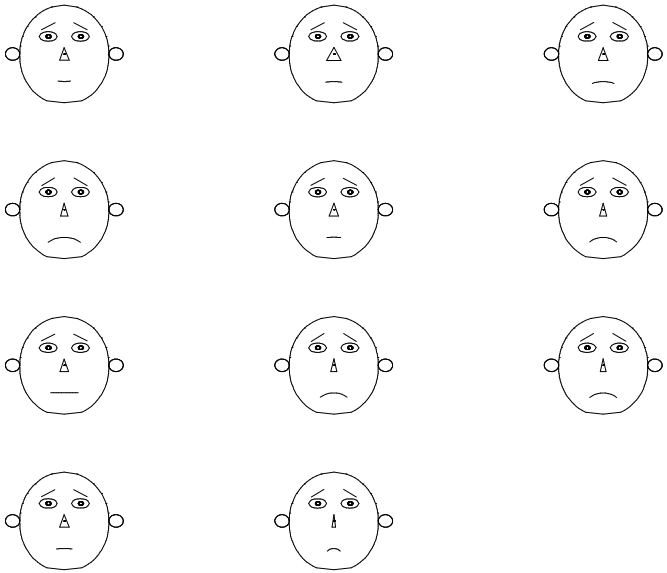

200
150
100
50
0
Dissolved O_2
0
10
20
30
40
50
Total N

Objects	Variables 1	2	3	etc.
1	y_{11}	y_{12}	y_{13}	
2	y_{21}	y_{22}	y_{23}	
3	y_{31}	y_{32}	y_{33}	
etc.				

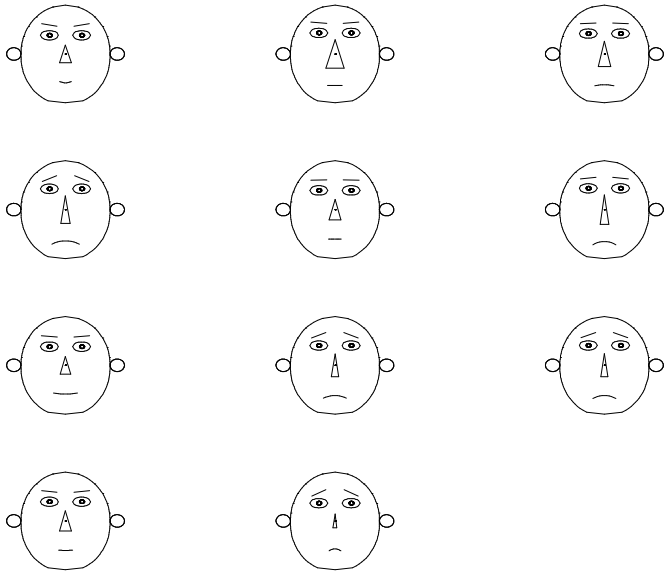
Variables	Variables 1	2	3	etc.
1	1	r_{12}	r_{13}	
2	r_{21}	1	r_{23}	
3	r_{31}	r_{32}	1	
etc.				

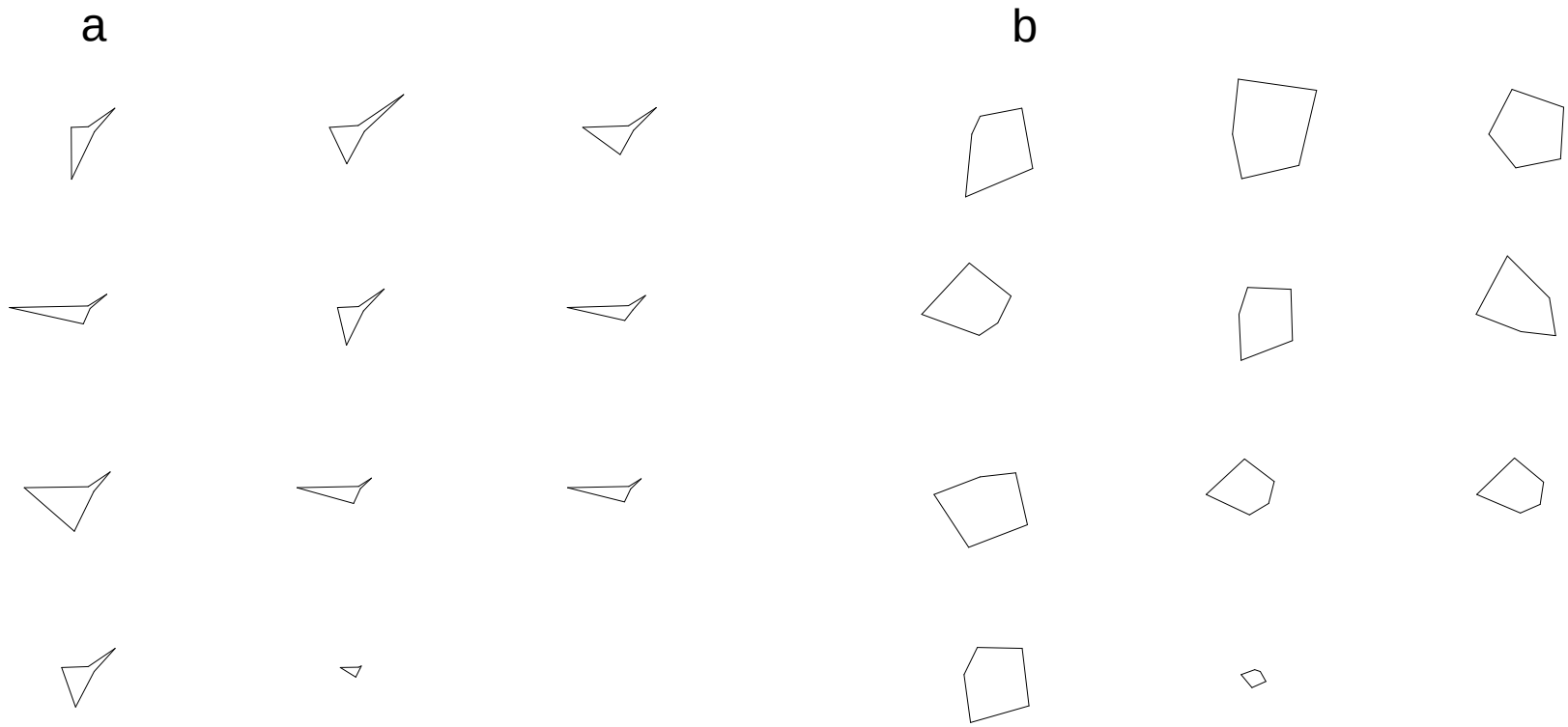
Objects	Objects 1	2	3	etc.
1	0	d_{12}	d_{13}	
2	d_{21}	0	d_{23}	
3	d_{31}	d_{32}	0	
etc.				

a



b




a
b