17. PRINCIPAL COMPONENTS AND CORRESPONDENCE ANALYSIS

This chapter and the next deal with analyses of multiple variables recorded from multiple objects where there are two primary aims. First is to reduce many variables to a smaller number of new derived variables that adequately summarize the information and can be used for further analysis, i.e. variable reduction. Multivariate analysis of variance and discriminant function analysis described in the previous chapter also have these aims. The discriminant functions represented the new derived variables that are extracted while explicitly accounting for group structure in the data set. Comparison of groups in the methods covered in this chapter and the next require subsequent analyses because the extraction of the summary variables does not consider group structure.

The second aim is to reveal patterns in the data, especially among objects, that could not be found by analyzing each variable separately. One way of detecting these patterns is to plot the objects in multidimensional space, the dimensions being the new derived variables. This is termed scaling, or multidimensional scaling, and the objects are ordered along each axis and the distance between objects in multidmensional space represents their biological dissimilarity (Chapter 15). Ecologists often use the term "ordination" instead of scaling, particularly for analyses that arrange sampling or experimental units in terms of species composition or environmental characteristics. Ordination is sometimes considered as a subset of gradient analysis (Kent & Coker 1992). Direct gradient analysis displays sampling units directly in relation to one or more underlying environmental characteristics. Indirect gradient analysis displays sampling units in relation to a reduced set of variables, usually based on species composition, and then relates the pattern in sampling units to the underlying environmental characteristics.

There are many different approaches to achieving the aims of variable reduction and scaling (ordination). In this chapter, we will describe methods based on extracting eigenvectors and eigenvalues from matrices of associations between variables or objects (Chapter 15). Methods based on measures of dissimilarity between objects will be the subject of Chapter 18.

17.1. Principal Components Analysis

Principal components analysis (PCA) is one of the most commonly used multivariate statistical techniques and it is also the basis for some others. For $i = 1$ to n objects, PCA transforms $j = 1$ to p variables ($Y_1, Y_2, Y_3,\ldots,Y_p$) into $k = 1$ to p new uncorrelated variables ($Z_1, Z_2, Z_3,\ldots, Z_p$) called principal components (or factors). The scores for each object on each component are called z-scores (Jackson 1991). For example, Naiman *et al.* (1994) examined the influence of beavers on aquatic biogeochemistry. Four habitats were sampled for soil and pore water constituents. Variables were N, Nitrate-N, ammonium-N, P, K, Ca, Mg, Fe, Sulfate, pH, Eh, % organic, bulk density, N fixation, moisture, redox. Three components explained 75% of the variation, with component I representing N & P, component II representing moisture and organic matter and component III representing ammonium-N and redox.

We will use two data sets from previous chapters, plus a new one, to illustrate principal components analysis.

Chemistry of forested watersheds

In Chapters 2 and 15, we described the work of Lovett *et al.* (2000) who studied 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). We will use PCA to reduce these variables to

a smaller number of components and use these components to examine the relationships between the 39 streams (Box 17-1).

Habitat fragmentation and rodents

In Chapter 13, we introduced the study of Bolger *et al.* (1997) who surveyed the abundance of seven native and two exotic species of rodents in 25 urban habitat fragments and three mainland control sites in coastal southern California. Besides the variables representing the species, other variables recorded for each fragment and mainland site included area (ha), % shrub cover, age (yr), distance to nearest large source canyon and distance to nearest fragment of equal or greater size. We will use PCA to reduce the species variables to a smaller number of components and use these components to examine the relationships between the habitat fragments and mainland sites (Box 17-2).

Geographic variation and forest bird assemblages

Mac Nally (1989) described the patterns of bird diversity and abundance across 37 sites in southeastern Australia. We will analyze the maximum abundance for each species for each site from the four seasons surveyed. There were 102 species of birds and we will use a PCA to try and reduce those 102 variables to a smaller number of components and use these components to examine the relationship between the 37 sites (Box 17-3).

17.1.1. Deriving components

Axis rotation

The simplest way to understand PCA is in terms of axis rotation (see Kent & Coker 1992, Legendre & Legendre 1998). Consider the study of Green (1996), who studied the ecology of red land crabs on Christmas Island (see Chapter 5). Part of that study measured two variables (total biomass of crabs and number of burrows) in ten quadrats in a forested site on the island. A scatterplot of these data is in Figure 17-1, with biomass on the vertical axis and burrow number on the horizontal axis. PCA can be viewed as a rotation of these principal axes, after centering to the mean of biomass and the mean of burrow number, so that the first "new" axis explains most of the variation and the second axis is orthogonal (right angles) to the first (see Figure 17-1). The first new axis is called principal component I and the second is called principal component II. The first component is actually a "line-of-best-fit" that is halfway between the least squares estimate of the linear regression model of biomass on burrow number and the regression line of burrow number on biomass. This is the estimate of the Model II regression (Chapter 5) and is the line represented by the correlation between burrow number and biomass (either raw or centered) and is also called the major axis. If the variables are standardized (to zero mean and unit standard deviation), then the first principal component represents the reduced major axis (Chapter 5). The second component is completely independent of, or uncorrelated with, the first component.

Decomposing an association matrix

When there are more than two variables, it is difficult (or impossible) to represent the rotation procedure graphically. In practice, the components are extracted either by a spectral decomposition of a sums-of-squares and cross products matrix, a covariance matrix or a correlation matrix among variables or by a singular value decomposition of the raw data matrix with variables standardized as necessary (see Chapter 15 and Box 15.1). Which matrix to use will be discussed in Section 17.1.2. There will be $k = 1$ to p principal components, each of which is a linear combination of the original variables:

$$z_{ik} = c_1 y_{i1} + c_2 y_{i2} + \ldots c_j y_{ij} + \ldots + c_p y_{ip} \qquad (17.1)$$

From Lovett et al. (2000):

$$z_{ik} = c_1(NO_3)_i + c_2(\text{total organic N})_i + c_3(\text{total N})_i + c_4(NH_4)_i + c_5(\log_{10} \text{dissolved organic C})_i + c_6(SO_4)_i + c_7(\log_{10} \text{Cl})_i + c_8(\text{Ca})_i + c_9(\text{Mg})_i + c_{10}(\log_{10} \text{H})_i \qquad (17.2)$$

In equations 17.1 and 17.2, z_{ik} is the value or score for component 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 forming this component. Although the number of components that can be derived is equal to the number of original variables, p, we hope that the first few components summarize most of the variation in the original variables.

The matrix approach to deriving components produces two important pieces of information - see Box 15.1. The eigenvectors contain the estimates of the coefficients for each principal component (the c_js in equation 17.1). Eigenvector I contains the coefficients for principal component I, eigenvector II for principal component II, etc.. As described in Box 15.1, the eigenvectors are usually scaled so that the sum of squared coefficients for each eigenvector equals one, although additional scaling is also sometimes used.

Estimates of the eigenvalues (or latent roots, λ_k) provide relative measures of how much of the variation between the objects, summed over the variables in the data set, is explained by each principal component. The components are extracted so that the first explains the maximum amount of variation, the second explains the maximum amount of that unexplained by the first, etc. If there are some associations between the variables, the first two or three components will usually explain most of the variation present in the original variables, so we can summarize the patterns in the original data based on a smaller number of components (variable reduction). In the analysis of the data from Lovett et al. (2000), the first three components comprised over 70% of the original variation (Box 17-1). If the original variables are uncorrelated, then PCA will not extract components that explain more of the variation than the same number of original variables – see analysis of data from Mac Nally (1989) in Box 17-3. Note that the sum of all the eigenvalues equals the total variation in the original data set, the sum of the variances of the original variables. PCA rearranges the variance in the original variables so it is concentrated in the first few new components.

17.1.2. Which association matrix to use?

The choice of association matrix between variables is an important one. The choice basically comes down to choosing between the covariance and the correlation matrix, because using the sums-of-squares and cross products matrix makes the resulting PCA sensitive to differences in mean values of the variables, even when they are measured in the same units and on the same scale. The covariance matrix is based on mean-centered variables and is appropriate when the variables are measured in comparable units and differences in variance between variables make an important contribution to interpretation. The correlation matrix is based on variables standardized to zero mean and unit variance and is necessary when the variables are measured in very different units and we wish to ignore differences between variances.

Most statistical software uses a correlation matrix by default in their PCA routines, although all should offer the covariance matrix as an alternative. Our experience is that most biologists use the correlation matrix but rarely consider the implications of analyzing variables standardized to zero mean and unit variance. For example, a PCA using the chemical data from Lovett *et al.* (2000) might be best based on a correlation matrix. Although the units of the variables are the same (μmol L^{-1}), the absolute values and variances are very different and we cannot attach an obvious biological interpretation to these very different variances (Box 17-1). In contrast, we might compare the results from using a covariance matrix with those from using a correlation matrix on the species abundance data from Bolger *et al.* (1998) to see if the different patterns of variance in abundance of species across fragments is important (Box 17-2). We argued in Chapter 15 that analyzing data with different forms of standardization can assist in interpretation. The message for using PCA is that using covariances will not produce the same components as using correlations (Jackson 1991, James & McCulloch 1990), and the choice depends on how much we want different variances among variables to influence our results.

17.1.3. Interpreting the components

The value of the components, and any subsequent use of them in further analyses, depends on their interpretation in terms of the original variables. The eigenvectors provide the coefficients (c_js) for each variable in the linear combination for each component. The further each coefficient is from zero, the greater the contribution that variable makes to that component. Approximate standard errors can be calculated for the coefficients (Flury & Riedwyl 1988, Jackson 1991), although the calculations are tedious for more than a few variables. Fortunately, these standard errors are default output from good statistical software and should be used when comparing the relative sizes of these coefficients. These standard errors are asymptotic only (i.e. approximate) and assume multivariate normality (Flury & Riedwyl 1988). The size of the standard errors can be relatively large compared to the size of the coefficients (Box 17-1).

Component loadings are simple correlations (using Pearson's r) between the components (i.e. component scores for each object) and the original variables. If we use centered and standardized data (i.e. a correlation matrix), the loadings are provided directly by scaled eigenvectors in the **V** matrix (see Box 15.1). If we use just centered data (i.e. a covariance matrix), the **V** matrix will contain covariances rather than correlations, although true correlations can be determined (Jackson 1991). High loadings indicate that a variable is strongly correlated with (strongly loads on) a particular component. The loadings and the coefficients will show a similar pattern (although their absolute values will obviously differ) and either can be used to examine which of the original variables contribute strongly to each component. Tabachnick & Fidell (1996) warn against placing much emphasis on components that are determined by only one or two variables.

Ideally what we would like is a situation where each variable loads strongly on only one component and the loadings (correlations) are close to plus/minus one (strong correlation) or zero (no correlation). It is also easier to interpret the components if all the strongly correlated variables have the same sign (+ve or -ve) on each component (which ones are +ve compared to -ve is actually arbitrary). What we usually get is much messier than this, with some variables loading strongly on a couple of components and many variables with loadings of about 0.5.

17.1.4. Rotation of components

The common situation where numerous variables load moderately on each component can sometimes be alleviated by a second rotation of the components after the initial PCA. The aim of this additional rotation is to obtain simple structure, where the coefficients within a component are as close to one or zero as possible (Jackson 1991). Rotation can be of two types. Orthogonal rotation keeps the rotated components orthogonal to, or uncorrelated with, each other after rotation. This includes varimax, quartimax, equimax methods, the first being the most commonly used. Oblique rotation produces new components that are no longer orthogonal to each other. Orthogonal rotation is simplest and maintains the independence of the components, although some (e.g. Richman 1986) have recommended oblique methods based on the results of simulation studies. Tabachnick & Fidell (1996) also argue that oblique rotation methods may be more realistic since the underlying processes represented by the components are unlikely to be independent.

The PCA on the chemical data for streams from Lovett *et al.* (2000) illustrates the advantages of secondary rotation, with more variables strongly correlated with just one of the retained components than with the unrotated solution (Box 17-1). This will not always be the case, but in our experience with biological variables, rotation often improves the interpretability of the components extracted by a PCA.

If the aim of the PCA is to produce components that will be used as predictor or response variables in subsequent analyses, and those analyses require that the variables are independent of each other (e.g. predictor variables in multiple linear regression models; Chapter 6), then oblique rotation methods should be avoided. Harris (1985), Jackson

(1991) and Richman (1986) provide the equations and statistical detail underlying rotations.

17.1.5. How many components to retain?

Although there are a number of approaches to determining how many components to keep (Jackson 1991, Jackson 1993), there is no single best method. Iit is important to examine the interpretability of the components and make sure that those providing a biologically interpretable result are retained. For example, there is little point retaining components with which no variables are strongly correlated, because these components will be difficult to interpret.

Eigenvalue equals one rule

We can use the eigenvalue equals one rule, which simply says to keep any component that has an eigenvalue greater than one when the PCA is based on a correlation matrix (Norman & Streiner 1994). The logic here is that the total amount of variance to be explained equals the number of variables (because using a correlation matrix standardizes the variables to a mean of zero and standard deviation of one), so by chance each component would have an eigenvalue of one. In the analysis of the water chemistry data from Lovett et al. (2000), three out of the ten possible components had eigenvalues greater than one (Box 17-1). In contrast, the analysis of the bird abundance data from Mac Nally (1989) resulted in 25 out of the 102 possible components with eigenvalues greater than one (Box 17-3).

Scree diagram

We can also examine the scree diagram, which simply plots the eigenvalues for each component against the component number. We are looking for an obvious break (or elbow) where the first couple of components explain most of the variation and the remaining group of components don't explain much more of the variation (Figure 17-2). The rule of thumb is to keep all components up to and including the first in that remaining group. Our experience is that scree diagrams don't offer more in interpretability than just simply examining the successive numerical eigenvalues for each component.

Tests of eigenvalue equality

There are tests for equality of a set of successive eigenvalues derived from a covariance matrix, such as Bartlett's and Lawley's tests (Jackson 1991, Jobson 1992), and we might use one of these to test the null hypothesis that the eigenvalues of the components not retained are equal. Bartlett's test is most common (and available in most statistical software as part of correlation or PCA routines) and the test statistic is compared to a χ^2 distribution. We usually test in a sequential manner, first testing that the eigenvalues of all components are equal (Bartlett's test is then a test of sphericity of a covariance matrix – see Chapters 10 and 11). If this is rejected, we then test equality of eigenvalues of all components except the first, and so on. Once we do not reject the null hypothesis, we retain all components above those being tested. This is a multiple testing situation so some adjustment of significance levels may be warranted (Chapter 3). Bartlett's and Lawley's tests are not applicable when using a correlation matrix because the test statistics do not follow a χ^2 distribution; approximate methods when using correlations are suggested by Jackson (1991).

Analysis of residuals

Residual analysis is also useful for PCA, just like for linear models. Remember that we can extract p components from the original (appropriately standardized) data and we can also reconstruct the original data from the p components. If we extract less than p components, then we can only estimate the original data and there will be some of the information in the original data not explained by the components – this is the residual. When we retain fewer than all p components, we are fitting a model analogous to a linear

model (Jackson 1991) with the original data (with variables usually standardized to unit variance) represented as a multivariate mean (centroid) plus a contribution due to the retained components plus a residual. This residual measures the difference between the observed value of a variable for an object and the value of a variable for that object predicted by our model with less than p components. Alternatively, we can measure the difference between the observed correlations or covariances and the predicted (reconstructed) correlations or covariances based on the less than p components – this is termed the residual correlation or covariance matrix (Tabachnick & Fidell 1996; see also Chapter 16).

We have a residual term for each variable for each object and the sum (across variables) of squares of the residuals, often termed Q (Jackson 1991), can be derived for each object. If the variances differ between the variables and some objects have much larger values for some variables, then the residuals, and Q-values, for those objects will probably be larger for a PCA based on a covariance matrix than one based on a correlation matrix. Whichever matrix is used, unusually large values of Q for any observation are an indication that the less than p components we have retained do not adequately represent the original data set for that object. Q-values can be compared to an approximate sampling distribution for Q to determine P values (the probability that a particular Q-value or one more extreme came from the sampling distribution of Q). When we retained three components from a PCA on the correlation matrix of the water chemistry data from Lovett et al. (2000), none of the residual values were statistically significant (Box 17-1).

However, formal statistical testing seems not very useful when exploring a multivariate data set for unusual values – just check unusual values relative to the rest. This is the same process for checking for outliers using residuals from linear models. Objects with large Q-values may be particularly influential in the interpretation of the PCA and a number of such objects would suggest that too few components have been retained to adequately describe the original data. These objects can be further examined to see which variable(s) contribute most to the large Q-value, i.e. which variables have the large difference between observed and predicted values.

17.1.6. Assumptions

Because it uses covariances or correlations as a measure of variable association, PCA is more effective as a variable reduction procedure when there are linear relationships between variables. Non-linear relationships are common between biological variables and under these circumstances, PCA will be less efficient at extracting components. Transformations can often improve the linearity of relationships between variables (see Chapter 4, Tabchnick & Fidell 1989).

There are no distributional assumptions associated with the ML estimation of eigenvalues and eigenvectors and the determination of component scores (the descriptive use of PCA). However, calculation of confidence intervals and tests of hypotheses about these parameters, such as a test that some of the eigenvalues are equal (see Section 17.1.5; also Jackson 1991, Jobson 1992), do assume multivariate normality. Outliers can also influence the descriptive results from a PCA, especially when based on a covariance matrix where the variances of variables contribute to the component structure. Multivariate outliers can be identified using Mahalanobis distances (Chapter 15).

When normality is questionable, because we have skewed univariate distributions of variables for example, then bootstrap standard errors and confidence intervals might be used. Alternatively, transformations of variables to achieve univariate normality might also improve multivariate normality, reduce the influence of outliers and also improve the linearity of the associations between variables.

Like all multivariate analyses, missing data are a real problem. The default setting for PCA routines in most statistical software is to omit whole objects that contain one or more missing observations. Unless the sample size (number of objects) is large and the objects with missing values are a random sample from the complete data set, then pairwise

deletion, multiple imputation or estimation based on the EM algorithm are more appropriate for dealing with missing observations (see Chapter 15).

17.1.7. Robust PCA

Robust PCA techniques allow us to derive components that are less sensitive to outliers. Two approaches have been suggested in the literature. The first is to use robust estimates of covariances or correlations (Jackson 1991). For example, we could use correlations based on ranked variable values, such as Spearman's rank correlation, for the PCA (Jobson 1992). Alternatively, we could calculate each correlation (or covariance) independently of the others, using trimmed observations or *M*-estimators, such as Huber's, that downweight extreme observations (Chapter 2). Calculating each pairwise covariance or correlation independently of the others, using all the available data for each pair of variables, is also an effective means of handling missing data (Chapter 15). The second approach is to use robust methods to derive components directly from the original data (Jackson 1991), although these are more complex to compute and there are no obvious criteria for choosing between the methods.

17.1.8. Graphical representations

Scaling (ordination)

The eigenvectors can be used to calculate a new score (*z*-score) on each component for each object. This is achieved by solving the linear combination for each object for each component (equation 17.1), using mean centered or standardized variables if the eigenvectors came from covariance or correlation matrices respectively (see Box 15.1). These scores can also be further standardized by dividing by the square root of the eigenvalue for the relevant component so that the variance of the scores for each component is one:

$$z_{ik}^{*} = \frac{z_{ik}}{\sqrt{l_k}} \qquad (17.3)$$

Some software may produce these standardized scores, rather than the original z-scores.

The objects can then be positioned on a scatterplot based on their scores with the first two or three principal components as axes (Figure 17-3). It doesn't matter whether *z*- or *z**-scores are used for the basic plot of objects, although some authors recommend that standardized scores should be used if the PCA is based on a correlation matrix (Jobson 1992). The interpretation of these plots is straightforward but subjective. Objects close together on the plot are more similar in terms of their variable values based on the components being a summary of the original variables; conversely for objects further apart. For a PCA on the data from Bolger et al. (1997), the sites Sandmark and Alta La Jolla are most similar in terms of native rodent species composition (Figure 17-3).

This type of graphical representation of objects from a multivariate analysis is termed scaling. When the objects are sampling units and the variables are species abundances, then ecologists describe analyses that produce such plots as ordinations and the plot an ordination plot. Clearly, we could plot each object using the original variables as axes, but such a plot is impractical beyond three variables. The plot of the component scores allows us to show the relationship between the objects based on the new derived components, given that the first two or three components can usually be interpreted in terms of the original variables and explain most of the original variance.

It is well known by ecologists that when we are dealing with data for species abundances for different sampling units (e.g. plots, sites etc.), then the scaling plot of the sampling units (objects) for the first two components of a PCA often shows an arching pattern (the "arch" and "horseshoe" effects). This arching is most apparent when the sampling units cover a long ecological gradient and those at each end of the gradient have few species in common (Minchin 1987, Wartenberg *et al.* 1987)). For example, the scaling of the bird

abundance data from Mac Nally (1989) shows a strong arch when sites are plotted for the first two principal component axes (Box 17-3; Figure 17-4). Although this arching may indicate the true ecological dissimilarities between the extreme sampling units, there is evidence that it distorts the true underlying pattern. One explanation for the arching is that the implicit measure of dissimilarity between objects that PCA uses, Euclidean distance, does not reach a constant maximum value when two sampling units have no species in common and thus can imply that two objects are similar due to joint absences. Sampling units with few or no species in common are most likely to occur at the extremes of an environmental or geographical gradient so the underlying relationship between dissimilarity and the environmental gradient is nonlinear. The inability to represent nonlinear relationships between dissimilarity and some gradient without distortion is not unique to PCA; correspondence analysis (Section 17.3) also has this problem. We will compare different approaches to scaling/ordination in Chapter 18.

We have described an *R*-mode analysis, where associations between variables are used to extract components. The PCA could be done as a *Q*-mode analysis where a matrix of associations between the objects is calculated (Legendre & Legendre 1998). Components can be extracted from either matrix and object scores derived from variable eigenvectors and eigenvalues and *vice versa*. Any differences relate to how variables or objects are standardized, since an *R*-mode PCA based on a correlation matrix standardizes variables to zero mean and unit variance. More commonly, *Q*-mode analyses are based on measured dissimilarities between objects (Chapter 18). It turns out that using the techniques in Chapter 18 to examine the relationship between objects based on a matrix of dissimilarities will produce almost identical scaling (ordination) plots to those produced by an *R*-mode PCA if we use Euclidean distance as the dissimilarity measure.

Biplots

One particular form of a scaling/ordination plot is called a biplot (Gower & Hand 1996), where both objects and variables (hence the "bi") are included on a single scaling plot. Biplots can use more than two axes although they are commonly plotted in two dimensions. The usual form of a biplot is a point-vector plot where the objects are points and the variables are represented by vectors (lines) drawn from the origin of the scaling plot. Biplots are possible because the singular value decomposition of a data matrix allows us to relate eigenvectors from a matrix of associations between variables to the eigenvectors from a matrix of associations between objects through the eigenvalues for the components (Box 15.1). The most common form of the biplot will use the component scores for objects as points and the variables are represented by the eigenvectors relating each variable to each component. If the PCA is based on a correlation matrix (i.e. centered and standardized variables), then the biplot will often use z^*-scores for the objects and component loadings to represent the variables on the biplot. In any case, some scaling of the eigenvectors or loadings for variables will usually be required so that the vectors are commensurate with the range of object scores.

Biplots are commonly used by ecologists in situations where the objects represent sampling units or sites and the variables are species abundances (e.g. Digby & Kempton 1987, Legendre & Legendre 1998). We have illustrated a PCA biplot for the 28 sites from the study of the effects of habitat fragmentation on rodents by Bolger *et al.* 1997 (Figure 17-3a; see also Box 17-2). We have included loading vectors for six of the species (vectors for all species resulted in a plot that was very crowded and difficult to read). The ends of the vectors represent the correlations of each species with each component, although the correlations have been scaled by three so they are roughly commensurate with site scores. For these point-vector biplots, it is not how close the head of the variable vector is to the object points on a biplot that is relevant because we usually have to scale the vectors in some way. It is the direction and relative length of these vectors that are important. The direction indicates that the values of the variable increase in that direction and the length indicates the rate of increase – long vectors are more gradual increases, short vectors are faster increases. So, the vector for *R. rattus* in Figure 17-3a indicates that this species increases rapidly in abundance in the opposite direction

from Balboa Terrace. The vector for *P. eremicus* indicates that this species increases more gradually in abundance in the direction of Sandmark and Alta La Jolla.

17.1.9. Other uses of components

One problem we face with many statistical analyses, particularly linear models, is dealing with numerous correlated response or predictor variables. We usually analyze each response variable separately with univariate regression or ANOVA techniques, which causes Type I error rate problems due to multiple testing, and we have difficulties using correlated predictor variables in these models because of the effects of collinearity on our parameter estimates and hypothesis tests. PCA may help in both situations because we can often reduce a large number of correlated variables down to a smaller number of components without losing much information and our linear model analyses can use these components as response or predictor variables.

Relationship to MANOVA

When we have multiple response variables in a design that we would usually analyze with an ANOVA model to estimate and test for differences between groups, there are two approaches we can use. The first is multivariate analysis of variance (MANOVA) that we described in Chapter 16. Basically, we analyze a component (discriminant function) that is extracted so it maximizes the explained variance between groups and the hypothesis being tested is about group differences on a linear combination of variables or differences between group centroids. The second approach is to initially ignore group differences and do a PCA on the whole data set, i.e. all objects, and then use as many of the derived components as deemed interpretable as response variables in univariate ANOVA models to test for group differences. The components are obviously independent of each other, although the F tests from univariate ANOVAs on these components technically are not (Jackson 1991).

The two approaches (MANOVA and ANOVA on components) will produce different results, although the broad patterns of group differences are likely to be similar. Analyzing components using ANOVA has some advantages. MANOVA is commonly described in terms of the first discriminant function and deriving output from software for other functions, especially for complex designs, is difficult. In contrast, ANOVA on components can analyze the second, third etc. components if they offer useful interpretations of the original variables. Also, post-hoc comparisons of groups are more straightforward under a univariate ANOVA framework.

Principal components regression

In Chapter 6, we discussed the problems caused by collinearity among predictor variables when fitting multiple regression models, especially the inflated standard errors of regression coefficients and the sensitivity of estimates of regression coefficients to which predictors are included in the model. One strategy sometimes suggested as a solution to this problem is principal components regression (Chaterjee & Price 1991, Lafi & Kaneene 1992, Rawlings *et al.* 1998). If there are serious correlations among the predictor variables, we can do a PCA on the predictors, usually centered (and maybe standardized), to extract the p components. We could then fit a regression model that uses all the components as the predictors, but such a model will predict the response variable with the same precision as a model based on the original variables. Usually, we fit a simpler model based on fewer than p components, although the choice of which components to retain is problematical (see below). If the components are easily interpretable, then principal components regression might be better than the original multiple regression because the components are orthogonal so there is no collinearity and no instability in the estimates of the regression coefficients.

We can also recalculate regression coefficients in terms of the original variables based on the relationship (Jackson 1991, Lafi & Kaneene 1992):

$$\mathbf{b} = \mathbf{U}\mathbf{b}_z \tag{17.4}$$

In equation 17.4, **b** is a matrix of regression coefficients on the original standardized variables, $\mathbf{b}_z$ is a matrix of regression coefficients on the principal components (derived using a correlation matrix) and **U** is the matrix of eigenvectors from the PCA on the predictor variables (see Box 15.1). When the PCA is based on a matrix of correlations between the predictors, then regression coefficients in **b** are standardized coefficients and relate to standardized predictor variables. Covariances could be used with just centred predictor variables.

Equation 17.4 simply states that we can obtain regression coefficients in terms of the original variables from the product of the regression coefficients for the principal components and the eigenvectors from the PCA. Using eigenvectors from the **U** matrix scales the coefficients so that the sum of squared coefficients equals one (Box 15.1).

The standard error of the regression coefficient for the *k*th principal component is (Chaterjee & Price 1991, Jackson 1991):

$$s_{b_k} = \sqrt{\frac{\mathrm{MS}_{\mathrm{Residual}}}{l_k}} \tag{17.5}$$

In equation 17.5, $\mathrm{MS}_{\mathrm{Residual}}$ is from the linear regression on the p principal components. So the standard errors are inversely proportional the eigenvalues and the first principal components will have smaller standard errors than later components.

If all p components are used, then the regression coefficients in **b** will be the same as those from the regression on the original (standardized) variables. If less than p components are used, then the regression coefficients in **b** will be different from the regression coefficients on the original (standardized) variables. These new coefficients will be biased, the bias increasing the fewer components we retain. In both cases (p or less than p components retained), the standard errors of the recalculated regression coefficients will also be smaller than those from the original multiple regression (Jackson 1991).

Chaterjee & Price (1991) provide a clear example of the calculations involved in principal components regression. Despite its attractiveness as a way of overcoming collinearity in multiple linear regression models, there are limitations to principal components regression. Hadi & Ling (1998) pointed out that the components that explain most of the variance in the predictor variables, i.e. the first few components derived using PCA, might not be the most important in explaining the variance in the response variable in a multiple regression model. The choice of which components to use in principal components regression should be based on their contributions to the $\mathrm{SS}_{\mathrm{Regression}}$, not just their eigenvalues from the original PCA.

17.2. Factor analysis

In Section 17.1.5, we pointed out that we can reconstruct the original data from the principal components but if we retain less than p components, we can only approximate the original data. The residual represents information in the original data not included within the less than p retained components. Factor analysis (FA) formalizes this into a structured model and we now use the term factors instead of components. FA is based on a correlation matrix, or less commonly a covariance matrix. The correlation matrix for the original variables is separated into two parts (Jackson 1991, Jobson 1992). The first is that generated by the common factors, those factors that explain all the correlations among the original variables. The second is that due to the unique factors, those factors representing information in the correlation matrix that is not explained by the common factors. So we have a model that basically includes explained and unexplained (residual) variability, although FA is "explaining" the correlation structure in the data rather than just the variance. The term communality is used for the variance of a variable explained by the common factors.

The mechanics of FA are pretty much the same as for PCA, although the procedure is more complex because we need to estimate both common factors and the residual variability associated with the unique factors. Jackson (1991) describes different approaches to estimation, the most commonly used called principal factor analysis where the matrix of correlations between the variables is modified so that the diagonal contains estimates of the communalities. A spectral decomposition is then applied to this new matrix to extract eigenvectors and eigenvalues.

The common factors are estimates of latent variables, the true variables causing the correlation structure in the data. Structural equation modeling (also termed latent variable analysis or causal modeling) combines FA with multiple regression so that the response and predictor variables may be measured variables or common factors (Tabachnick & Fidell 1996). When only measured variables are used, we have multiple regression modeling and the possible causal relationships between response and predictor variables can be displayed as a path diagram (Chapter 6). When we have factors on either side of our regression model, we have structural equation modeling and the path diagrams are more sophisticated. We strongly recommend Tabachnick & Fidell (1996) for a readable introduction to structural equation modeling.

Jackson (1991) summarized the differences between PCA and FA. The most fundamental is that PCA is trying to extract components that explain the variability in the original variables whereas FA is trying to explain correlations among the original variables. FA is not commonly used in biological research, probably because biologists are trying to extract a small number of new variables that explain most of the variability in the original variables and use these new variables in scaling or ordination plots. PCA is clearly more appropriate than FA for these purposes. Jackson (1991) and Manly (1997) include good introductions to FA and Tabachnick & Fidell (1996) compare some of the common statistical software routines for FA and PCA.

17.3. Correspondence analysis

Correspondence analysis (CA) was developed as a method for decomposing contingency tables of counts (see Chapter 14) into a small number of summary variables and representing the lack of independence between rows and columns of the contingency table as a low dimensional plot. CA is based on a raw data matrix of counts, classified by n rows (objects) and p columns (variables). In Chapter 14, we described tests for independence of rows and columns in a two way contingency table of counts. A simple test was based on the χ^2 statistic calculated as:

$$\chi^2_{(n-1)(p-1)} = \sum_{i=1}^{n}\sum_{j=1}^{p}\frac{(o_{ij}-e_{ij})^2}{e_{ij}} \qquad (17.6)$$

where o_{ij} are the observed cell counts and e_{ij} are the expected cell counts under independence. Large values of this statistic indicate lack of independence between rows and columns, i.e. the proportion of counts in different columns depends on the row and vice-versa. The main purpose of CA is to summarize the lack of independence between rows (objects) and columns (variables) of a contingency table as a small number of derived variables, sometimes called principal axes. The maximum number of derived variables is the minimum of $(n\text{-}1)$ and $(p\text{-}1)$, although usually only two axes are derived. The scores for each object and each variable on these axes are used in the scaling (ordination) plot, often with objects and variables plotted jointly.

We will illustrate the use of CA to scale jointly the 28 sites and nine species of rodents from the habitat fragmentation study of Bolger et al. (1997). This CA is presented in Box 17-4.

17.3.1. Mechanics

CA proceeds by a double transformation of the observed minus expected counts, dividing by the product of the square roots of the row totals (r_i) and column totals (c_j). This is equivalent to using standardized residuals from the model of independence for a two-way contingency table, adjusted by the total frequency:

$$\frac{1}{\sqrt{N}}\frac{(o_{ij}-e_{ij})}{\sqrt{e_{ij}}}=\frac{(o_{ij}-e_{ij})}{\sqrt{r_i}\sqrt{c_j}} \tag{17.7}$$

We could just use the observed counts in the numerator of equation 17.7 (Jackson 1991, Ludwig & Reynolds 1988) and the basic results of the CA are the same except that the first principal axis becomes trivial and is ignored in interpretation. The matrix approach to CA can be of two forms, like PCA. First, we can use a SVD on the matrix of transformed counts (**H**):

$$\mathbf{H}=\mathbf{U}^{*}\mathbf{L}^{1/2}\mathbf{U}' \tag{17.8}$$

In equation 17.8, **U*** represents the eigenvectors for each component with coefficients for variables, **U** represents the eigenvectors for each component with coefficients for objects and **L** represents a diagonal matrix containing the eigenvalues for each component (Box 15.1). Therefore, we have two sets of eigenvectors, one for objects and one for variables. Second, we can convert **H** into two association matrices, one between variables ($\mathbf{H}'\mathbf{H}$) and the other between objects ($\mathbf{H}\mathbf{H}'$) and use spectral decomposition of both these association matrices to extract the same eigenvectors and eigenvalues.

Because the eigenvectors for objects and variables are extracted jointly, after a double transformation of counts to contributions to the χ^2 statistic for lack of independence, the eigenvalues associated with the principal axes for rows and columns are the same. The sum of these eigenvalues is equal to the overall χ^2 statistic divided by the total frequency and is called total inertia, a measure of lack of independence. The eigenvalues are interpreted similarly to those from a PCA, with % of the total inertia explained by the successive axes usually presented. The first axis should explain a high proportion of the lack of independence between objects and variables. The axes are extracted in CA so that the correlation between variable and object scores is as high as possible. The axes are also orthogonal (independent) of each other.

17.3.2. Scaling and joint plots

The eigenvectors are used to determine a score for each principal axis for each object and for each variable. These scores are used for the scaling (ordination) plots. Commonly, objects and variables are plotted together as a joint plot (a "point-point" plot). The biggest difficulty in interpreting these joint plots from a CA is the numerous options for scaling (or standardizing) the object and variable eigenvectors and subsequent scores. As with PCA, the scores are scaled by a measure based on $\sqrt{l}$, where l is the estimated eigenvalue for that axis. An alternative scaling (Hill's method) uses $\sqrt{(l(1-l))}$. The different scaling options result in "minor, but irritating, variants in presenting CA results" (Gower 1996, p. 162), a problem exacerbated by the different terminology used by statisticians and biologists, especially ecologists. These different forms of scaling don't change the order of objects or species along the axes but do change their relative positions because the underlying dissimilarity measure differs. Not all the types of scaling allow sensible joint plots (see below).

Jackson (1991) described scaling options for objects and variables that result in the implicit dissimilarity between points being Euclidean distance (Chapter 15; see also Legendre & Legendre 1998). More commonly, especially for biological applications, we scale objects and/or variables so that the implicit dissimilarity between points is the chi-square metric (Chapter 15), and this is the usual output from CA routines in software. The distances between objects and/or variables in the scaling plot are proportional to their chi-square distances. Three common scalings available in specialist software used for

ecological applications (sampling units by species abundances) produce scores that can be used in biplots (see Legendre & Legendre 1998):

- Scores for sampling units are scaled so that they are positioned at the centroids of the species scores. The distances between sampling units are proportional to their chi-square distances and this scaling is appropriate when the main focus is on relative positions of sampling units (objects).
- Scores for species are scaled so they are positioned at the centroids of the sampling unit scores. The distances between species are proportional to their chi-square distances and this scaling is appropriate when the main focus is on relative positions of species (variables).
- Compromise scaling tries to scale sampling unit and species scores comparatively with a method "half-way" between the first two.

It often doesn't matter which scaling is chosen because the pattern of objects to variables in the joint plots will be similar – just the absolute scores are different and the values of the axis scores are not of much practical use. Note that some software plots either objects or variables as points and the other as vectors, as in a biplot, although CA actually produces a point-point plot of objects and variables jointly, not a true biplot. You also occasionally see the point-point joint plots called biplots. Finally, some programs do not scale scores in a manner that allows sensible joint plots, especially CA routines in general statistical software (Legendre & Legendre 1998).

The interpretation of the joint plot of object and variable scores is different from a biplot. In CA, objects and variables that occur together on the plot indicate that the variables have values greater than predicted under independence for those objects, or conversely, objects have greater values than predicted for those variables. Examining the joint plot in conjunction with a matrix of residuals from the independence model for the contingency table will be helpful since we can see which cells have large deviations from expected values. We would expect combinations of objects and variables with large positive deviations to be near each other on the plot, whereas combinations with large negative deviations to be in opposite quadrants of the plot. With the scaling options described above, those variables (e.g. species) contributing most to the position of the objects (e.g. sampling units) will be the ones closest to the particular object on the plot.

The scores produced by a CA can be used, like principal components scores, as response variables in subsequent analyses. For example, we could correlate the sampling unit scores from a CA with other environmental variables recorded for each unit or use the sampling unit scores to examine difference between groups of units.

17.3.3. Reciprocal averaging

Scaling the eigenvectors so that dissimilarities between points are chi-square distances also relates to an alternative approach to CA, termed reciprocal averaging (Hill 1973, 1974; see descriptions in Digby & Kempton 1987, Ludwig & Reynolds 1988). This is an iterative procedure that calculates object scores for the first axis as a weighted average of variable scores and vice versa. At each step, the object and variable scores are rescaled so they are comparable. Final scores are obtained when there is little change in scores between iterations and convergence is usually quick. The process is then repeated for the second axis. The reciprocal averaging procedure is tedious and produces the similar scores (given rounding error) as the much more efficient matrix approach to CA when the two methods are used with the equivalent scaling. However, the default settings will often be different between programs that use the reciprocal averaging algorithm and programs that use the matrix approach – don't be surprised by variations in output from competing software. The reciprocal averaging algorithm is particularly useful when we wish to constrain the axis scores by additional variables, as in canonical correspondence analysis (Section 17.6).

17.3.4. Use of CA with ecological data

The most common users of CA in biology are community ecologists, who often deal with data sets consisting of n objects (sampling units, sites etc.) and p variables (species abundances) – see Section 17.1. By treating these data sets as two-way contingency tables, CA can be used to scale objects and variables simultaneously by plotting the scores for sampling units and species. These data sets are often based on sampling units along ecological gradients so that units at each end of the gradient (i.e. units furthest apart spatially or temporally or most different along some underlying environmental gradient) have few or no species in common. Ecologists describe this as high beta diversity, i.e. large changes in species diversity along environmental gradients (Ludwig & Reynolds 1988). We have already pointed out that under these conditions, PCA can produce a distorted scaling/ordination plot of sampling units (objects) so that units at the ends of the gradient are closer together than they should be ("arch" effect) and may even curve back in ("horseshoe" effect) – see Legendre & Legendre (1998) for an excellent summary. This effect is partly because the PCA scaling plot is trying to display a potentially complex and nonlinear relationship between dissimilarity and true ecological distance in a simple form (two or three dimensions), using a dissimilarity measure (Euclidean) that does not represent these distances very well.

CA also suffers from this problem (Legendre & Legendre 1998), because the implied dissimilarity measure is chi-square distance and, like Euclidean, this does not reach a constant maximum value when two sampling units have no species in common (Chapter 15). Also, because chi-square distance is measuring differences in proportional representation of species between sampling units, it tends to weight rarer species higher in the calculation of dissimilarity than their overall abundance warrants (Minchin 1987). Therefore, sampling units with few or no species in common may appear more similar relative to other sampling units in the CA plot than we would expect from their species composition and abundance (Wartenberg *et al.* 1987). If we are using the CA scaling plot to look for underlying ecological gradients, then this distortion can make interpretation difficult, especially for the second axis, because patterns of sampling units related to a second gradient (assuming the first is displayed along the first axis) may be obscured. The second axis is a quadratic distortion of the first axis, rather than reflecting a second ecological gradient (Kent & Coker 1992). Van Groenewald (1992) simulated ecological data with clear gradients and showed that CA does not recover underlying gradients beyond the primary one very well if they are nearly as strong as the primary gradient. Therefore, we cannot recommend CA as an appropriate method for scaling sampling units across long ecological gradients.

17.3.5. Detrending

Hill & Gauch (1980) proposed detrended correspondence analysis (DCA) as a solution to the arching problem. Detrending breaks the first axis up into a number of segments, the number determined by the user, and rescales the second axis so its average is the same for all segments. Detrending is applied to the reciprocal averaging algorithm, with rescaling occurring at each iteration. While this method is effective at removing the arch effect, different numbers of segments used in the detrending process can affect the results (Jackson & Somers 1991). Also, the method assumes that the arch effect is an artifact of the CA, and not a real pattern in the data (Minchin 1987). Simulations by Minchin (1987) showed that DCA performed poorly relative to other methods (e.g. nonmetric multidimensional scaling; see Chapter 18) in trying to recover known ecological gradients, although this was due to both the instability of the results to detrending and the implicit chi-square dissimilarity measure. Therefore, we cannot recommend DCA as a scaling/ordination technique because of the arbitrary nature of detrending, its sensitivity to the number of segments chosen and even problems with order of data entry for some versions of the algorithm (Okansen & Minchin 1997).

17.4. Canonical correlation analysis

Biologists may have a data set where they wish to examine the correlation between one set of variables and another set of variables for the same objects. For example, consider the data from Lovett *et al.* (2000) described in Section 17.1. The variables recorded from each of the 39 stream sites were of two types: 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) – see Box 17-5. We might wish to examine the correlation between the set of chemical variables and the set of watershed variables. We could do this by examining all the pairwise correlations between the variables (30 pairwise correlations). Alternatively, we could use canonical correlation analysis where we extract linear combinations of variables (components) from the two sets of variables so that first component for one set has the maximum correlation with the first component from the second set. The components are termed canonical variates and the first component from each set forms one pair of canonical variates, the second component from each set forms a second pair, etc.. The number of canonical variates, and therefore pairs, is the number of variables in the smallest set.

The basic equation for canonical correlation analysis is:

$$\mathbf{R} = \mathbf{R}_{11}^{-1}\mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21} \qquad (17.9)$$

In equation 17.9, $\mathbf{R}$ is the matrix of canonical correlations, $\mathbf{R}_{12}$ and $\mathbf{R}_{21}$ are the correlation matrices between sets 1 and 2 and between sets 2 and 1 respectively, and $\mathbf{R}_{11}$ and $\mathbf{R}_{22}$ are the correlation matrices within sets 1 and 2 respectively. Basically, this is an eigenvalue-eigenvector problem similar to that outlined for PCA (Box 15.1), with the constraint that the canonical variates are paired so they have the maximum correlation among all possible pairs of canonical variates. The matrix calculations are tedious but described in detail by Jackson (1991), Jobson (1992), Manly (1994) and Tabachnick & Fidell (1996). In some software, canonical correlation analysis can be set up as a regression problem with one set of variables being the response set and the other set being the predictor set.

The output from running a canonical correlation analysis in most software will be familiar once you are used to eigenvalues, eigenvectors and component scores from PCA (Box 17-5). The descriptive output usually includes matrices of correlations within and between the two sets and regression statistics for each response variable regressed against each predictor variable (these are based on standardized variables because we are using correlations).

Output related specifically to the canonical correlation analysis includes eigenvectors and loadings for the canonical variates from each set, interpreted in the same way as eigenvectors and loadings from PCA. Remember that we are using correlation matrices here so the comparable PCA interpretation is for centered and standardized variables. The relative signs associated with eigenvector coefficients and loadings are arbitrary within a variate but the interpretation of the canonical correlations between variates depends on the signs associated with the variables within each variate. For example, the analysis of the correlation between the set of chemical variables and the set watershed variables from Lovett et al. (2000) showed negative loadings for NO_3 and negative loadings for maximum and site elevation for canonical variate 1. The interpretation here is that large values of NO_3 are associated with large values of maximum and site elevation. Positive loadings for variables in one set and negative loadings for variables in the other for a canonical variate indicate that large values of the variables is one set are associated with small values of the variables in the other. Always check your interpretation by examining the univariate correlations to make sure your interpretation of the direction of the multivariate relationship makes sense.

We also get a test of the H_0 that there is no correlation between any of the pairs of the canonical variates, usually provided as Bartlett's χ^2 statistic. If this H_0 is rejected, then we know that at least the first pair of canonical variates is significantly correlated. Most

software then provides tests for the subsequent pairs, usually sequentially by testing the remaining pairs after the first has been removed, then those still remaining after the first two have been removed etc.

Like PCA (Section 17.1.3), the interpretation of canonical correlation analysis really depends on how easily the canonical variates can be interpreted in terms of the original variables. Also like PCA (Section 17.1.4), rotation of the canonical variates is possible and may improve the simple structure for each pair of variates.

The nature of the matrix calculations in canonical correlation analysis means that it is very sensitive to collinearity among the variables in either set, especially when one or both sets have many variables (see Tabachnick & Fidell 1996). In these circumstances, omitting one or two variables can cause marked differences (instability) in the magnitude and signs of the variable loadings on the canonical variates. This is a similar problem that affects multiple regression (Chapter 6) and multivariate analysis of variance (Chapter 16) and other procedures that require matrix inversion. Removing redundant variables (those highly correlated with others) is about the only option. A method for assessing correlations between two sets of variables that is sensitive to correlations between pairs of variables within or between the sets must have limited applicability to real world data.

We have not found many examples of canonical correlation analyses in the biological literature, nor have we have we had much cause to consider using it ourselves. This may be because biologists are most interested in hypotheses about correlations between specific pairs of variables, rather than sets of variables or exploratory descriptions of relationships between objects based on some form of scaling (ordination).

17.5. Redundancy analysis

An obvious extension of canonical correlation analysis would be to distinguish response and predictor variables and develop a predictive model whereby we predict a linear combination of response variables from a linear combination of predictor variables. The proportion of the total variance in the response variables that can be explained by (predicted from or extracted by) a linear combination of the predictor variables is termed redundancy (Tabachnick & Fidell 1996). The statistical procedure for estimating this variance and developing the predictive model is termed redundancy analysis (RDA: Van den Wollenberg 1977). Legendre & Legendre (1998) and Legendre & Anderson (1999a) provide excellent descriptions of RDA. A multiple linear regression model relating each response variable to the set of predictor variables is estimated and a matrix of predicted Y-values for the response variables determined. This matrix is just like the raw data matrix comprising n objects by p variables, except that the values for each variable are those predicted by the regression model. This matrix of predicted Y-values is then subjected to a PCA via spectral decomposition of the covariance matrix of the predicted values (see Box 15.1) to extract eigenvectors and their "canonical" eigenvalues. The redundancy, the variance in the response variables explained by the predictor variables, is the sum of these eigenvalues. The eigenvectors can be used to calculate scores for each object and can be used as axes for scaling/ordination of the objects.

The contrast with PCA is important (Legendre & Legendre 1998). In a PCA, a covariance (or correlation) matrix of the response variables would be decomposed into eigenvectors and their eigenvalues, principal component scores determined for each object based on these eigenvectors and a scaling/ordination plot derived from these scores. In RDA, the response variables are first constrained to be a linear combination of some set of predictor variables, using multiple regression, and then the eigenvectors and their eigenvalues are extracted, object scores calculated and a scaling/ordination plot derived. The RDA eigenvectors are constrained to be a linear combination of the predictor variables, whereas the PCA eigenvectors are not related to predictor variables in any way (Jongman *et al.* 1995, Legendre & Anderson 1999a).

RDA can therefore be viewed as an extension of canonical correlation analysis that explicitly models multiple response variables against multiple predictor variables.

However, ecologists commonly use RDA as a modification of PCA to produce eiegnvectors and component scores for sampling units that are constrained to a linear combination of environmental variables recorded for each sampling unit (Legendre & Legendre 1998). For example, Verschuren *et al.* (2000) examined the composition of the fossil invertebrate community in different levels of a core taken from a lake bed in Kenya and used RDA to incorporate three environmental variables: salinity, lake level and papyrus-swamp development. The significance of the overall model relating the species abundance data set and the predictor variables, and also of individual predictor variables, can be tested using randomization procedures (Legendre & Anderson 1999a; Manly 1997). The predictor variables do not have to be continuous and an important application of RDA is when the predictors are dummy variables representing categories of categorical factors and their interactions (Legendre & Anderson 1999a; Chapter 18).

In the context of scaling/ordination, the logic of RDA can be illustrated with the data from Bolger *et al.* (1997). The response variables would be the abundance of the different rodent species for 28 fragments (objects) and the predictor variables would be the other fragment characteristics, such as area, % shrubs, age etc.. The scaling of the fragments in terms of species abundances would be constrained so that the components were linear combinations of the predictor variables. An alternative way of constraining axes of a scaling/ordination plot is within the context of correspondence analysis and will be described in the next section.

17.6. Canonical correspondence analysis

As indicated in the previous section, ecologists who work with data sets of species abundances for a number of sampling units sometimes also have additional variables (covariates) recorded for each site. For example, in the study of rodents in habitat fragments, Bolger *et al* (1997) also recorded the area of the fragment, the % of the area covered with shrubs, the age of the fragment, the distance to the nearest large "source" canyon and the distance to the nearest canyon fragment of equal or greater size. We might be interested in not only scaling the sites and species, such as with CA, but also examining how the relative positions of sites and species are related to the values of the additional covariates for each site. Canonical correspondence analysis (CCA) is a modification of CA where the principal axes are extracted not only so they explain most of the total inertia (lack of independence between objects and variables) but also so that their correlation with additional variables is maximized (Jongman et al. 1995, Kent & Coker 1992, Legendre & Legendre 1998, ter Braak & Verdonschot 1995).

CCA uses the reciprocal averaging algorithm for CA. At each step when sampling unit scores are determined, they are constrained to be a linear combination of environmental variables (usually standardized) using OLS multiple regression techniques (Chapter 6). The predicted values of the site scores from this multiple regression are then used to calculate species scores and the iterative process continues (Jongman *et al.* 1995). Incorporating the environmental variables in this way also ensures that the extracted axes maximize the dispersion of the species scores based on the linear combination of environmental variables. The axes in CA also maximize the dispersion of species scores but independently of any environmental variables.

The main decisions for users of software for CCA are about standardizations or transformations of species and/or environmental variables and standardization and scaling of site and species scores. Linear relationships between environmental variables and scores may be improved by transforming environmental variables so they have closer to a symmetrical distribution. The options for scaling the scores for CCA are similar to those for CA (Section 17.3.2) and the choice of scaling needs to be made carefully if the objects and variables are to be included in a joint plot.

The CCA algorithm produces axes that represent maximum correlations with linear combinations of the environmental variables, with the second axis being uncorrelated with the first. CCA produces two sets of site scores. The first are those produced without

being constrained by the environmental variables, although for some reason these are different when produced by CCA than when the same data are analyzed by CA. The second are those produced by the multiple regression of the above scores on the linear combination of environmental variables. Palmer (1993) termed these WA and LC scores respectively, and described them as the observed sites scores, as weighted averages of species scores, and those site scores predicted from the multiple regression on the environmental variables. He recommended plotting the LC scores, arguing that the meaning of the WA scores is unclear and they differ from the scores from a straight CA anyway. The relative positions of sites based on the three types of scores (CCA WA scores, CCA LC scores, CA scores) is usually different, although broad patterns are comparable.

Output from CCA algorithms includes axis scores for sites and species and vectors representing the correlations between the environmental variables and principal axes can also be included on these plots, creating a biplot. Canonical weights for the final multiple regression model are provided as well as correlations between the environmental variables and species and site scores. CCA can be run with the detrending option although as discussed in Section 17.3.5, detrending is difficult to justify. The big advantage of CCA is the simultaneous scaling of sites and species (like CA) while at the same time maximizing the correlations between the principal axes and linear combinations of environmental variables. Its disadvantages are those of CA described in Section 17.3.4, especially the chi-squared distance measure, and the limited availability of software; CCA is not available in any of the common commercial programs and specialist ecological software like CANOCO is required.

Blanche *et al.* (2001) illustrate the use of CCA in their experimental study of the effects of fire on the community of ground-active beetles in tropical savannahs of Kakadu National Park in northern Australia. There were three fire treatments (unburnt, early-season burn each dry season, late-season burn each dry season) and six years of sampling (pre-burn in 1988-89 and post-burn from 1990-94). Abundances of ground-dwelling beetles, sorted to family and species, in each of three replicate 15-20km^2 experimental compartments (small catchments) for each treatment in each year were measured with pitfall traps. The replicate compartments were combined for the analysis and two environmental covariates were also recorded for each year-treatment combination: fire intensity and rainfall just prior to sampling. The CCA showed that the effects of treatment were contingent on both sampling rainfall and fir intensity (Figure 17-6). Treatment-year combinations favoured by high rainfall tended to be pre-burn years and unburnt treatments and late-burn treatments were correlated with less rainfall and more intense fires.

We illustrate a worked example of CCA based on the rodent data from Bolger *et al.* (1997) in Box 17-6. The 25 habitat fragments were scaled based on the abundances of nine rodent species, with three variables used to constrain the ordination: area of the fragment (ha), the age of the fragment (yr), and the distance to the nearest large "source" canyon (m). All three variables were important in determining the associations of fragments with species (Figure 17-7) and the biplot was quite different to that produced by a CA on the same data, ignoring the environmental variables (compare Figure 17-7 with Figure 17-8).

The logic of CCA is to include the environmental variables as part of the site and species scaling (ordination). An alternative approach is to scale the sites separately and then examine which species contribute most to the pattern and also relationships with environmental variables. We will discuss these approaches in Chapter 18.

17.7. Constrained and partial "ordination"

Both RDA and CCA are known as constrained scaling procedures because the relative positioning of the objects in the scaling (ordination) plot is constrained by a set of covariates. In an ecological setting, we usually have sampling units being scaled based on the abundances of multiple species, with the covariates being environmental variables

recorded for each sampling unit or even spatial coordinates of each sampling unit. These constrained methods are very informative because they allow the relationship between the environmental variables and scaling of sampling units or species to be explored simultaneously. RDA, like PCA, is most appropriate when the relationship between species abundances and underlying environmental gradients is linear. This is unlikely in practice, especially for long environmental gradients, so CCA, like CA, is more suited when the relationship between species abundances and underlying environmental gradients is unimodal (Jongman *et al.* 1995). Forms of scaling/ordination that have fewer assumptions about the relationship between species abundances and underlying gradients, such as nonmetric multidimensional scaling, will be described in Chapter 18.

An interesting development that can be applied to any of the constrained scaling/ordination methods is partial ordination (Legendre & Legendre 1998). Imagine a situation where we have two sets of environmental variables, and we wish to use one set to constrain a scaling of sampling units based on species abundance after eliminating the effects of the second set. An example given by Jongman *et al.* (1995) is where there is one or more "impact" variables representing effects of some human activity and one or more covariates representing other sources of variation we are less interested in, such as seasonal factors (ter Braak & Versonschot 1995). A partial scaling of sampling units would use the impact variables after removing the effects of the other covariates. This would be achieved by fitting multiple regression models with each of the covariates of prime interest (e.g. impact variables) as the response variable and the secondary covariates we are partialling out as the predictor variables. The residuals from each of these models represent the variation in each of the primary covariates that is not explained by the linear relationship with the secondary covariates. These residuals are then used instead of the original primary covariates in a CCA or RDA.

These partial ordination techniques allow us to examine the relationships between a scaling based on species abundances and some environmental variables after partialling out the effects of other covariates. For example, Verschuren *et al.* (2000) examined the fossil invertebrate communities in a core of sediment from a lake in Kenya. They used RDA to examine the relationships between the scaling of sampling units (sections of the core) and three environmental variables (salinity, lake level, papyrus-swamp development) and used partial RDA to look at the effects of each of these covariates after removing one or both of the remaining ones. We might also be interested in how much of the variation between sampling units in abundances of multiple species can be attributed to a set of environmental variables, a set of spatial coordinates, the variation shared by the environmental and spatial components and the undetermined (residual) variation. Bocard *et al.* (1992) described a method based on either partial RDA or CCA to determine the variation in the original sampling units by species data matrix into these four components. The residuals from multiple regression models of either environmental variables on spatial coordinates or vice-versa are used to examine the contribution of the environmental variables and spatial coordinates independently of each other. Note that for partial RDA, it is variance being partitioned; for partial CCA, it is inertia. In both cases, percentage contributions can be determined.

17.8. General issues and hints for analysis

General issues

- The implicit dissimilarity measures used in scaling/ordination techniques, such as Euclidean for PCA and RDA and chi-square for CA and CCA, may not be best suited to all types of data, especially species abundance data.
- The choice between covariance and correlation for the association matrix in a PCA is important. Use covariance if you wish differences in variance for each variable to contribute to the analysis. Use correlation if the variables are measured on different scales abd you do not wish differences in variance for each variable to have any influence on the analysis.