centroids for two or more variables. In the Peckarsky *et al.* (1993) example, we would test whether there is an effect of predator treatment on a combination of body mass, egg mass, percentage of eggs, total mass, and maturation time of individual mayflies.

We will illustrate MANOVA with two examples from the biological literature.

Trace metals in marine sediments

Haynes *et al.* (1995) carried out a pilot study to test for differences between sites in trace metal concentrations in marine sediments off the Victorian coast in southern Australia. They had three sites: Delray Beach, site of a proposed wastewater outfall, and two possible control sites, Seaspray and Woodside. At each site, they had four randomly chosen stations and at each station, two randomly chosen cores of sediment. They recorded the concentrations of copper, chromium, cadmium, lead, iron, nickel, manganese and mercury. We will test for the effects of site on a subset of these response variables taken together. Although this is strictly a nested design, site would be tested against the random station effect so we will average the replicate cores for each station and use a single factor MANOVA for comparing sites. The analysis of these data is presented in Box 16.1.

Plant functional groups and leaf characters

In Chapters 9 and 15, 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 analyze a subset of their data (to avoid missing cells), with two locations (Colorado and Wisconsin) and two functional groups (forbs and shrubs) in a crossed design. There were between three and eleven species in each cell and five response variables were measured: specific leaf area ($\log_{10}$ transformed), leaf nitrogen concentration, mass-based net photosynthetic capacity, area-based net photosynthetic capacity and leaf diffusive conductance at photosynthetic capacity. We will test for the effects of location and functional group, and their interaction, on these five response variables taken together. The analysis of these data is presented in Box 16.2.

16.1.1 Single factor MANOVA

Linear combination

The simplest design where a MANOVA is appropriate is when we have n replicate experimental or sampling units ("objects" from Chapter 15) allocated to two or more levels of a factor (groups) and we record p (where p is greater than two) response variables from each unit. The MANOVA is based on a linear combination (z) of the p response variables as defined in Chapter 15 (see Equations 15.1 and 15.2). In the example from Haynes *et al.* (1995), there were n equals four replicate stations in each of three groups (sites) with p equals four response variables (trace metals). The MANOVA uses the linear combination (z) of response variables, out of the infinite number of possible linear combinations, which maximizes the ratio of between-group and within-group variances of z. This linear combination is also called the discriminant function for the difference between groups and is used in discriminant function analysis (see Section 16.2):

$$z_{ik} = \text{constant} + c_1 y_{i1} + c_2 y_{i2} + \ldots c_j y_{ij} + \ldots + c_p y_{ip} \quad (16.1)$$

For example, from Haynes *et al.* (1995):

$$z_{ik} = \text{constant} + c_1(\log_{10} \text{Cu})_i + c_2(\log_{10} \text{Pb})_i + c_3(\log_{10} \text{Ni})_i + c_4(\log_{10} \text{Mn})_i \quad (16.2)$$

From Reich *et al.* (1999):

$$z_{ik} = \text{constant} + c_1(\log_{10} \text{specific leaf area})_i + c_2(\text{leaf N})_i + c_3(\text{mass-based photosynthetic capacity})_i + c_4(\text{area-based photosynthetic capacity})_i + c_5(\text{leaf diffusive capacity})_i \quad (16.3)$$

In Equations 16.1, 16.2 and 16.3, z_{ik} are the values for object i for linear combination k, the combination that maximizes the ratio of between-group and within-group variances of z_{ik}. From Haynes *et al.* (1995), this is the value for station i from solving Equation 16.2 for linear combination k. The coefficients (c_j) are the weights measuring the relative contribution of each variable to the linear combination. As described in Box 15.1, these coefficients will be scaled in some form and will be represented in matrix descriptions of MANOVA as elements of a matrix of eigenvectors (Box 15.1). Note that if the variables are

Box 16.1 Worked example of MANOVA: heavy metals in marine sediments

Haynes *et al.* (1995) carried out a pilot study to test for differences between sites in trace metal concentrations in marine sediments off the Victorian coast in southern Australia. They had three sites: Delray Beach, site of a proposed wastewater outfall, and two possible control sites, Seaspray and Woodside. At each site, they recorded the concentrations of copper, chromium, cadmium, lead, iron, nickel, manganese and mercury (means of two sediment cores) at four randomly chosen stations. We used only the 1991 data in our analyses. There were strong correlations among some of the metals (e.g. Cu and Cr, Fe and Ni) so only four variables (Cu, Ni, Pb, Mn) were included in the analysis. There was strong indication of skewness for the four variables, so all were $\log_{10}$-transformed. There were a few cases with significant ($P<0.001$) Mahalanobis distances ($D_{ij}^2>16.3$) but these were not extreme and remained in the analysis. All variables except Cu (Levene's test, $P=0.023$) had similar variances between groups.

The multivariate test statistics all result in rejection of the H_0 that there is no difference in site group centroids.

	Statistic	df	F	P
Wilk's λ	0.058	8, 12	4.728	0.008
Pillai trace	1.272	8, 14	3.058	0.033
Hotelling–Lawley trace	10.549	8, 10	6.593	0.004

Pairwise contrasts among the sites, with a sequential Bonferroni (Holm's method) adjustment of P values, indicated that only the difference between Delray Beach and Woodside was significant.

Contrast	Pillai trace	df	F	P	Adj P
Delray vs Seaspray	0.713	4, 6	3.719	0.074	0.078
Delray vs Woodside	0.909	4, 6	14.924	0.003	0.009
Seaspray vs Woodside	0.772	4, 6	5.092	0.039	0.078

The univariate F tests indicate significant differences between sites for all four metals.

Source	df	MS	F	P
Log Cu				
Site	2	0.098	5.208	0.031
Residual	9	0.019		
Log Pb				
Site	2	0.136	4.834	0.038
Residual	9	0.028		
Log Ni				
Site	2	0.083	8.655	0.008
Residual	9	0.009		
Log Mn				
Site	2	0.244	23.608	<0.001
Residual	9	0.010		

The raw and standardized coefficients for the discriminant function obviously differ but because the variables were $\log_{10}$-transformed, the difference in scales between the variables is not great and the basic pattern is the same. The standardized coefficients suggest that lead contributes least to the difference between sites and manganese the most. The loadings simply reflect the univariate *F*-ratio statistics from above, and the pattern is the same as for the coefficients. Mn and Ni are most important, and Pb least important, at separating the sites.

Variable	Raw coefficient	Standardized coefficient	Loading
Constant	−29.013		
Log Cu	1.253	0.172	0.334
Log Pb	−0.494	−0.083	0.258
Log Ni	6.690	0.653	0.428
Log Mn	9.308	0.945	0.724

We had no theoretical basis for ordering our variables so we entered them in a step-down analysis in order of their univariate *F*-ratios. Log Mn entered first, then we tested log Ni with log Mn as a covariate, then log Cu with log Mn and log Ni as covariates and finally log Pb with log Mn, log Ni and log Cu as covariates. We were not interested in testing hypotheses about the covariates and adjusted the significance levels for the site effects with a Holm correction (Chapter 3).

Source	df	MS	*F*	*P*(Adj *P*)
Log Mn				
Site	2	0.244	23.608	<0.001 (0.004)
Residual	9	0.010		
Log Ni				
Site	2	0.034	3.407	0.085 (0.255)
Log Mn	1	0.007		
Residual	8	0.010		
Log Cu				
Site	2	0.011	0.512	0.620 (0.910)
Log Mn	1	<0.001		
Log Ni	1	0.023		
Residual	7	0.021		
Log Pb				
Site	2	0.033	0.901	0.455 (0.910)
Log Mn	1	0.021		
Log Ni	1	0.022		
Log Cu	1	0.002		
Residual	6	0.037		

The step-down analysis suggests that none of the variables contributes significantly to the difference between groups when entered after log Mn, i.e. none of the site effects for any variable is significant once log Mn is included as a covariate.

Box 16.2 Worked example of MANOVA: plant functional groups and leaf characters

Reich *et al.* (1999) examined the generality of leaf traits from different species across a range of ecosystems and geographic regions. We will use two of their locations (Colorado and Wisconsin) and two of their functional groups (forbs and shrubs) in a crossed design. There were between three and eleven species in each cell and five response variables were measured: specific leaf area ($\log_{10}$-transformed), leaf nitrogen concentration, mass-based net photosynthetic capacity, area-based net photosynthetic capacity and leaf diffusive conductance at photosynthetic capacity.

There is some concern about the assumption of homogeneity of variances and covariances, especially as Levene's test for homogeneity of variances was statistically significant for three (log specific leaf area, leaf N and G_s) out of the five variables.

There were no significant multivariate test statistics for either main effect or the interaction. Note that since there were only two levels of each factor, the df, the approximate *F*-ratios and the *P* values were identical for each term for all three multivariate statistics.

	Wilk's λ	Pillai trace	Hotelling–Lawley trace	df	*F*	*P*
Location	0.573	0.427	0.745	5, 16	2.384	0.085
Functional group	0.549	0.450	0.819	5, 16	2.622	0.065
Interaction	0.836	0.164	0.196	5, 16	0.626	0.682

The univariate *F* tests indicate that the only significant effect was that of functional group for nitrogen concentration in leaves, although the effect of functional group for mass-based net photosynethetic capacity and of location for leaf diffusive conductance were marginal.

Source	df	*F*	*P*
Location	1, 19		
Log specific leaf area		0.880	0.359
Leaf N		0.005	0.947
A_{mass}		1.025	0.323
A_{area}		0.042	0.841
G_s		3.756	0.069
Functional group	1, 20		
Log specific leaf area		2.299	0.145
Leaf N		5.305	0.032
A_{mass}		3.254	0.086
A_{area}		1.148	0.297
G_s		2.645	0.119
Interaction	1, 20		
Log specific leaf area		1.979	0.175
Leaf N		0.774	0.389
A_{mass}		1.624	0.217
A_{area}		0.065	0.802
G_s		1.112	0.304

The standardized discriminant function coefficients for each main effect and interaction would not normally be of much interest given that there were no significant effects from the MANOVA. We present them simply to illustrate that there is a separate discriminant function for each effect in the model and we can interpret these coefficients just as we would for single factor MANOVAs.

Variable	Location	Functional group	Interaction
Log specific leaf area	−2.002	−1.721	1.309
Leaf N	1.798	1.499	−1.294
A_{mass}	0.612	1.409	0.479
A_{area}	−1.489	−1.615	0.338
G_s	1.436	1.472	−0.709

standardized to zero mean and unit variance, the constant equals zero.

The determination of the linear combination that maximizes the ratio of between-group and within-group variances is best done using simple matrix algebra, some of which we have already described in Chapter 15. The steps for a single factor MANOVA are as follows.

1. The between-groups, within-groups and total SS used in an ANOVA are replaced by sums-of-squares-and-cross-products matrices (SSCP or S; see Chapter 15), one matrix for between groups (the hypothesis or effect matrix, **H**), one for within groups (the error or residual matrix, **E**) and one for total (the total matrix, **T**). The values in the main diagonal of these matrices are the univariate sums-of-squares for each variable, either between group means (**H**) or pooled across replicates within groups (**E**). The other elements are the sums-of-cross-products between any two of the variables. For example, the cross product for the between-groups matrix for two variables is the sum of (i) the product of the differences between each group mean and the overall mean for one variable and (ii) the differences between each value and the mean for the other variable – see Table 16.1.

2. We multiply **H** by the inverse of **E** (i.e. $\mathbf{HE}^{-1}$). Matrix inversion is the multivariate analogue of division so what we are really doing here is "dividing" **H** by **E**, the between-groups SSCP matrix "divided by" the within-groups SSCP matrix.

3. We then decompose the resulting matrix product (Box 15.1) to calculate characteristic roots or eigenvalues of each linear combination (eigenvector). The eigenvalues measure how much of the total between-group variance in the variables (the sum of the between-group variances of each of the variables) is explained by each linear combination or eigenvector. The eigenvectors contain the coefficients for each linear combination.

4. The linear combination producing the largest eigenvalue is the linear combination that maximizes the ratio of between-group and within-group variance (i.e. maximizes the explained variance between groups) and the eigenvector is a vector of coefficients or weights for that linear combination.

Null hypothesis

The H_0 for a single factor MANOVA is that the population effect of the groups or treatments is zero with respect to all linear combinations of the response variables. This is equivalent to no difference between population centroids (multivariate means). This H_0 can be tested by using statistics based on one of the measures of variance of a matrix, such as the determinant or the trace (Chapter 15; see also Harris 1985, Johnson & Field 1993, Stevens 1992, Tabachnick & Fidell 1996).

- Wilk's lambda (λ), which is the ratio of the determinants of the within-groups SSCP and the total SSCP: $|\mathbf{E}|/|\mathbf{T}|$. Remember that the determinant of a matrix is a measure of generalized variance for that matrix (Chapter 15), so Wilk's λ is a measure of how much of the total variance is due to the residual, with smaller values indicating larger group differences.
- Hotelling–Lawley trace, which is the ratio of the determinants of the between-groups SSCP and the within-groups SSCP: $|\mathbf{H}|/|\mathbf{E}|$. This is also the sum of the eigenvalues (trace) of the matrix product $\mathbf{HE}^{-1}$. Larger values indicate greater differences between group centroids.
- Pillai trace, which is the sum of the eigenvalues (trace) of $\mathbf{HT}^{-1}$, i.e. the variance between groups.
- Roy's largest root, which is the largest eigenvalue of $\mathbf{HE}^{-1}$, i.e. the eigenvalue of the linear combination that explains most of the variance and covariance between groups. This statistic is less commonly provided by statistical software.

Table 16.1 | Groups (a) and residual (b) and total (c) sums-of-squares-and-cross-products matrices for data from Haynes *et al.* (1995)

(a)

	Log_{10} Cu	Log_{10} Pb	Log_{10} Ni	Log_{10} Mn
Log_{10} Cu	0.196			
Log_{10} Pb	0.152	0.273		
Log_{10} Ni	0.164	0.192	0.165	
Log_{10} Mn	0.306	0.275	0.273	0.487

(b)

	Log_{10} Cu	Log_{10} Pb	Log_{10} Ni	Log_{10} Mn
Log_{10} Cu	0.169			
Log_{10} Pb	0.001	0.254		
Log_{10} Ni	0.045	0.031	0.086	
Log_{10} Mn	−0.011	0.033	−0.026	0.093

(c)

	Log_{10} Cu	Log_{10} Pb	Log_{10} Ni	Log_{10} Mn
Log_{10} Cu	0.369			
Log_{10} Pb	0.153	0.523		
Log_{10} Ni	0.209	0.223	0.251	
Log_{10} Mn	0.295	0.308	0.247	0.579

Note:
The main diagonals are sums-of-squares between groups, within groups and total and the other elements are cross-products.

The sampling distributions of these statistics are not well understood and they are usually converted to approximate *F*-ratio statistics (Tabachnick & Fidell 1996). Wilk's, Hotelling's and Pillai's statistics produce identical *F* tests when there are only two groups and become Hotelling's T^2 statistic – see example based on plant functional group data in Box 16.2. This is the multivariate extension of the *t* test for comparing two groups (Harris 1985, Tabachnick & Fidell 1996). They will generally produce similar results with more than two groups, although Pillai's trace seems to be the most robust of the tests (Johnson & Field 1993), especially when the assumption of similar variance–covariance matrices might be violated (Section 16.1.4). In our two worked examples (Box 16.1 and Box 16.2), the conclusions from Wilk's, Hotelling's and Pillai's statistics were the same. Most statistical software will provide **H**, **E**,

maximum λ and all the multivariate test statistics with their approximate F tests.

16.1.2 Specific comparisons

Most statistical software allow contrasts among the factor levels in MANOVA, analogous to planned contrasts in the univariate ANOVA (Chapter 8). Unplanned multiple comparisons are a more difficult problem, although use of Bonferroni-adjusted (see Chapter 3) pairwise MANOVAs is one conservative solution. Harris (1993) and Johnson & Field (1993) have reviewed other approaches for comparing specific groups after a MANOVA.

16.1.3 Relative importance of each response variable

If the null hypothesis of no difference between group centroids is rejected, we usually are interested in which of the response variables contributes most to the group differences. There are several methods of assessing the relative contribution of each response variable to the difference between groups in a MANOVA.

Univariate ANOVAs

We can examine the univariate ANOVAs on each response variable separately. Indeed, univariate hypotheses about group differences for each response variable will often be relevant. These univariate results do not necessarily indicate the relative contribution of each variable to the MANOVA result because they ignore correlations between variables. Correlations between variables can have marked effects on the power of MANOVA tests (Cole *et al.* 1994). Some authors (e.g. Harris 1985, 1993) also emphasize the problem of increasing family-wise Type I error rates when doing multiple univariate ANOVAs, a problem inherent in any multiple testing situation (see Chapter 3).

Step-down analysis

Step-down analysis is an analogue of forward selection stepwise multiple regression (Chapter 6) but taking into account the group structure (Tabachnick & Fidell 1996). This procedure relies on ordering the response variables based on theoretical expectations of their importance or using univariate analyses to choose the variable that shows the greatest difference between groups. First, the response variable with the highest priority is decided; for example, this might be the variable with the largest F-ratio from univariate ANOVAs on all the response variables. Each response variable is then tested sequentially, in the order determined *a priori*, in an ANCOVA model (Chapter 12), with groups as the categorical predictor and the higher priority response variables as covariates. We are interested in how much each additional variable adds to the variance explained by the variables already included.

Automated step-down analysis is available in some statistical software; otherwise, it must done with a series of ANCOVAs. Step-down analysis suffers from the problems we described in Chapter 6 for stepwise multiple regression, although we are not trying to find the "best" model in this situation, just assess the relative importance of each of the response variables. Step-down analysis also results in numerous unplanned significance tests so you need to be aware of the high family-wise Type I error rate. Huberty (1994) describes similar approaches, such as deleting variables one at a time and running a MANOVA on each set of the $p - 1$ remaining variables. The variables can be ordered based on the size of the change in MANOVA test statistic for each set.

Coefficients of linear combination

A more subjective approach is based on examining the discriminant function, i.e. the linear combination of the response variables that maximizes the ratio of between-group to within-group variance. There is a coefficient for each variable in the discriminant function, plus one for the grand mean (i.e. intercept or constant). If the different variables are measured on comparable scales (or we have values of a single variable recorded repeatedly through time in a repeated measures design), then the relative size of these coefficients (also termed "weights") provides a comparable measure of the contribution of each variable to the variance explained by the discriminant function and thus the difference between groups. If the variables are measured on very different scales, then we need to standardize them so that the coefficients can be compared. The simplest method is to standardize the discriminant function by the within-group variances, although whether this produces coefficients that are directly comparable is debatable (see Harris 1985 and Huberty 1994 for differing opinions). Standard errors can be estimated for each discriminant function coefficient (Flury & Riedwyl 1988), although they are rarely provided by statistical software.

Loadings

Loadings are the correlations between each variable and the discriminant function (see Chapters 15 and 17). These simply represent the correlations between the value of a variable and the score for the discriminant function with the units as replicates. The loadings of each variable on each discriminant function can be found by multiplying the within-group correlation matrix between variables (pooled across groups) by the matrix of standardized discriminant function coefficients (Tabachnick & Fidell 1996). Correlations automatically standardize the variables and examining loadings is popular because correlation coefficients are familiar and easily interpretable. However, these loadings are directly proportional to the univariate F-ratio statistics for each variable tested between groups so they ignore any relationships between the variables (Harris 1985). Note that one of the effects of highly correlated response variables can be a contradictory pattern when coefficients are compared with loadings.

Comments

Most statistical software will provide all of these coefficients and the loadings, either in MANOVA output or as part of a discriminant function analysis. The terminology used in the output does, however, vary considerably between programs and Tabachnick & Fidell (1996) provide a detailed comparison of the major software. Our experience is that unless there are many variables with some high correlations, the different approaches will produce a similar pattern. In our worked example of trace metals in sediments (Haynes *et al.* 1995; Box 16.1), the variables were log-transformed but not standardized. The univariate F-ratios, loadings and function coefficients showed the same pattern, with the order of importance being log Mn, followed by log Ni, log Cu and log Pb. The step-down analysis showed that none of the variables contributed significantly to site differences besides log Mn.

16.1.4 Assumptions of MANOVA

It is important to check normality, homogeneity of variance, and outliers for each response variable using univariate exploratory data analysis procedures (boxplots, residual plots, pplots, etc.; see Chapter 4). Given that the multivariate tests (especially Pillai's trace) are relatively robust to deviations from multivariate normality, particularly if each response variable has approximate univariate normality and sample sizes are equal, two multivariate assumptions are of major concern (Johnson & Field 1993, Tabachnick & Fidell 1996).

First, MANOVA tests are sensitive to multivariate outliers, which are cases with an unusual pattern of values for all the response variables considered simultaneously. Mahalanobis distance, the distance of each observation from the centroid or multivariate mean, can be used to detect multivariate outliers and is provided by most statistical software (Chapter 15).

Second, homogeneity of variances and covariances (i.e. equality of the variance–covariance matrices for each group) is an important assumption – this is the multivariate extension of univariate homogeneity of within-group variances. If this assumption is not met, then the pooled within-group matrix (**E**) will be misleading. Box's M test can test the H_0 of equal variance–covariance matrices but it is very sensitive to deviations from multivariate normality and is not recommended. There is no easy check for this assumption (but see discussion in Johnson & Field 1993), although it is more likely to be met when univariate homogeneity holds for each response variable. Like univariate ANOVA tests, MANOVA tests are more reliable when sample sizes are equal. Reducing the dimensionality (reducing the number of variables) of the analysis improves the robustness of all the MANOVA tests statistics (Johnson & Field 1993).

Johnson & Field (1993) provided strong evidence from simulation studies that Pillai's trace statistic is the most robust to deviations from the assumption of homogeneity of the variance–covariance matrices across groups. Suitable transformations of individual variables should

Table 16.2 | MANOVA results from Juenger & Bergelson (2000) who tested the effects of clipping, emasculation and their interaction on four response variables (flower, fruit, and seed production, total seed mass) of the perennial wildflower, the scarlet gilia

Source	df	Wilk's λ	F	P
Clipping (C)	4, 56	0.467	23.950	<0.001
Emasculation (E)	4, 45	0.936	3.251	0.439
C × E	4, 56	0.826	2.768	0.029

also be used where appropriate and including quadratic terms in the discriminant function can also help (Section 16.2.3).

Collinearity between variables is also a problem in the same way as for multiple regression (Chapter 6). The discriminant function coefficients, just like multiple regression coefficients, will be sensitive to which variables are included or excluded when variables are highly correlated. Most statistical software provides collinearity diagnostics, such as tolerance or variance inflation factors, and examinations of pairwise correlations between variables will be informative. Not including highly correlated (redundant) variables will help lessen the impact of collinearity and, since it reduces the dimensionality of the data matrix, will also make the MANOVA more robust to heterogeneous variance–covariance matrices (see above).

16.1.5 Robust MANOVA

Approaches to MANOVA that are robust to the underlying assumptions of multivariate normality and homogeneity of the variance–covariance matrices have been based on randomization procedures (Johnson & Field 1993). Edgington (1995) and Manly (1997) describe numerous possible test statistics for randomization MANOVA tests. These include a test based on Wilk's lambda, one using the sum of the logs of the univariate t or F statistics for each variable and one that compares the sum of squared Euclidean distances between objects and their sample centroids between groups and within groups. Manly (1997) pointed out that with large data sets (many variables and/or observations), only a subsample of all possible randomizations of observations on all variables to groups will be possible.

Another type of test is to determine distances or dissimilarities between all pairs of objects and compare the between-groups and within-groups dissimilarities. These tests will be described in Chapter 18 when we consider in detail multivariate analyses based on dissimilarities.

16.1.6 More complex designs

MANOVAs can also be used to test null hypotheses about combinations of variables in more complex designs. The matrix calculations described in Section 16.1.1 are done for each effect and error term that would have been used if univariate ANOVA models were fitted (Harris 1985). Separate linear combinations of variables are thus constructed for each main effect and interaction (Box 16.2) and the contribution of each variable needs to be assessed separately for each effect and its appropriate discriminant function. An ecological example is from Juenger & Bergelson (2000), who studied interactions between herbivory and pollination on various aspects of reproduction in the perennial wildflower *Ipomopsis aggregata* ssp. candida (the scarlet gilia) in Colorado, USA. Their experimental design had two factors: artificial grazing or clipping (two levels: control vs experimentally clipped) and male function (two levels: control vs emasculation, i.e. anther removal). There were 20 replicate plants in each combination (cell) of the two factor crossed design and a number of response variables were measured for each plant: total number of flowers, fruits and undamaged seeds and total seed mass. They used Wilk's lambda to test the two main effects and the interaction effect on the combination of the four response variables (Table 16.2).

Table 16.3 | MANOVA results from Pennings & Calloway (1996) who set up an experiment in a saltmarsh with three factors: *Cuscuta* infection by the parasitic plant *Cuscuta salina*, zone within saltmarsh and size of patch. They recorded the biomass of three non-parasitic plant species and analyzed these three response variables with a three factor (infection, marsh zone, patch size) crossed MANOVA

Source	df	Pillai's trace	F	P
Infected or not	3, 54	0.58	24.43	<0.001
Marsh zone	3, 54	0.38	11.00	<0.001
Patch size	3, 54	0.51	18.56	<0.001
Infection × zone	3, 54	0.19	4.12	0.010
Infection × size	3, 54	0.41	12.60	<0.001
Zone × size	3, 54	0.13	2.60	0.062
Infection × zone × size	3, 54	0.09	1.84	0.150

A more complex factorial MANOVA was used by Pennings & Callaway (1996), who studied the effects of a parasitic plant (*Cuscuta salina*) on a saltmarsh community. They set up an experiment with three factors: *Cuscuta* infection, zone within saltmarsh and size of patch. They recorded the biomass of three non-parasitic plant species and analyzed these three response variables with a three factor (infection, marsh zone, patch size) crossed MANOVA and tested the hypotheses for each main effect and interaction using Pillai's trace statistic (Table 16.3).

Note that one of the commonest applications of MANOVA in biology is in the analysis of repeated measures designs (Chapters 10 and 11), where the differences between pairs of repeated measurements are analyzed as multiple response variables using MANOVA statistics.

16.2 | Discriminant function analysis

Discriminant function analysis (DFA) is a "classification" technique, introduced by Fisher (1936) and recently reviewed by Huberty (1994). DFA is used when we have observations from pre-determined groups with two or more response variables recorded for each observation. DFA generates a linear combination of variables that maximizes the probability of correctly assigning observations to their pre-determined groups and can also be used to classify new observations into one of the groups. We might also wish to have some measure of the likelihood of success of our classification. Examples of DFA are common in the biological literature. For example, Skelly (1995) used DFA to test how well three variables (survivorship, size and larval period) could be used to classify individuals of two species of frogs (chorus frogs and spring peepers). Petit & Petit (1996) used DFA to separate four habitats based on ten variables (canopy cover, canopy height, density of various stem sizes) measured around nest boxes occupied by warblers along the Tennessee River.

We will illustrate DFA with the same two data sets we used for MANOVA.

Trace metals in marine sediments

We will analyze the data from Haynes *et al.* (1995), previously used in Box 16.1 for a MANOVA, with a discriminant function analysis. Our aim is to classify stations to each of the three sites (Delray Beach, Seaspray, Woodside) based on trace metal concentrations in marine sediments off the Victorian coast in southern Australia. The DFA of these data is in Box 16.3.

Plant functional groups and leaf characters

We will also analyze the data from Reich *et al.* (1999), used in Box 16.2 for a MANOVA, with a discriminant function analysis to classify species into one of four location and plant functional group combinations (Colorado–forb, Colorado–shrub,

Box 16.3 Worked example of discriminant function analysis: trace metals in marine sediments

We will illustrate a discriminant function analysis using the data from Haynes *et al.* (1995) – see Box 16.1. The aim here is to try and predict site membership of stations based on the four variables recorded for each station. The variance explained by each discriminant function was as follows.

	Eigenvalue	Percentage of variance
Function 1	9.979	94.6
Function 2	0.570	5.4

The first discriminant function explains most of the between-group (between-site) variance. The MANOVA test showed a significant difference between sites in the first discriminant function (Pillai trace = 1.272, df = 8, 14, *F*-ratio = 3.058, $P = 0.033$; see Box 16.1).

The relative contributions of each of the four trace metals to each discriminant function were as follows.

	Raw coefficient		Standardized coefficient		Loading	
	1	2	1	2	1	2
Constant	−29.013	−0.822	0	0		
Log Cu	1.253	3.030	0.172	0.415	0.334	−0.271
Log Pb	−0.494	−5.042	−0.083	−0.847	0.258	0.845
Log Ni	6.690	−3.126	0.653	−0.305	0.428	0.409
Log Mn	9.308	2.864	0.945	0.291	0.724	−0.159

The general pattern is the same for raw and standardized coefficients and loadings. Manganese and nickel contribute the most to the first function (Box 16.1), whereas lead contributes most to the second function. Note that within a discriminant function, the direction of the sign for each variable is arbitrary, i.e. the positives and negatives could be reversed with no change in interpretation.

The classification functions for each site are tabulated below.

	Delray	Seaspray	Woodside
Constant	−339.675	−421.174	−534.398
Log Cu	14.723	14.191	22.851
Log Pb	−24.237	−18.519	−27.090
Log Ni	171.225	195.893	216.269
Log Mn	258.338	282.345	320.356

These classification functions were solved for each station and each station classified to the site with the highest value.

	Delray	Seaspray	Woodside	Percentage correct
Delray	4	0	0	100
Seaspray	0	4	0	100
Woodside	0	0	4	100
Total	4	4	4	100

Note that percentage successful prediction is perfect. The classification matrix produced using a jackknife technique was as follows.

	Delray	Seaspray	Woodside	Percentage correct
Delray	3	1	0	74
Seaspray	1	3	0	75
Woodside	1	1	2	50
Total	5	5	2	67

Note that the jackknifed model results in lower percentage successful prediction but these percentages may be a more reliable indicator of classification success because we have excluded each observation when calculating the classification coefficients.

A discriminant function plot using group mean scores showed that the three sites discriminate clearly along function 1 but there is little separation along function 2, not surprisingly since function 1 explained nearly all of the variation between sites (Figure 16.1).

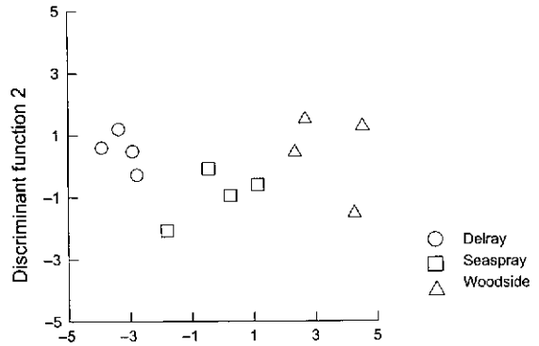

Figure 16.1 Plot of discriminant function scores for each replicate station for the first two functions from discriminant function analysis of data from Haynes *et al.* (1995). The four variables were concentrations of the metals Cu, Pb, Ni and Mn in the sediment, all $\log_{10}$-transformed.

Wisconsin–forb, Wisconsin–shrub) based on five response variables: specific leaf area ($\log_{10}$-transformed), leaf nitrogen concentration, mass-based net photosynthetic capacity, area-based net photosynthetic capacity and leaf diffusive conductance at photosynthetic capacity. The DFA of these data is in Box 16.4.

16.2.1 Description and hypothesis testing

Discriminant function analysis (DFA) is mathematically identical to a single factor MANOVA, although the former emphasizes classification and prediction rather than tests of hypotheses about group differences. However, the first step in any DFA is to derive discriminant functions (also called canonical discriminant functions) that are linear combinations of the original variables. The first discriminant function is the linear combination of variables that maximizes the ratio of between-groups to within-groups variance (i.e. maximizes the differences between groups) and is the linear combination used for the MANOVA test of no differences between group centroids derived in Section 16.1.1. The second discriminant function is independent of (uncorrelated with) the first and best separates groups using the variation remaining (the residual variation) after the first discriminant function has been determined, and so on for the third, fourth, etc., discriminant functions.

The number of discriminant functions that can be extracted depends on the number of groups and the number of variables – it is the lesser of the degrees of freedom for groups (number of groups minus one) and the number of variables (Tabachnick & Fidell 1996). In the example from Haynes *et al.* (1995), with three groups (sites) and four variables, there can be only two discriminant functions. Even in situations when there are more functions, the first one or two usually have the most discriminating power. Most statistical software also provides eigenvalues

Box 16.4 Worked example of discriminant function analysis: plant functional groups and leaf characters

We examined our ability to discriminate between the four location and functional group combinations for species of plants on which Reich *et al.* (1999) measured five variables – see Box 16.2. Like the two factor MANOVA earlier, the multivariate tests indicated no significant differences between the four groups (e.g. Pillai Trace = 0.902, df = 15, 54, *F*-ratio = 1.548, $P = 0.121$) for the first discriminant function.

The following classification functions were solved for each species (object).

	Colorado forb	Colorado shrub	Wisconsin forb	Wisconsin shrub
Constant	−535.136	−570.858	−576.439	−593.038
Log specific leaf area	557.743	580.616	582.442	592.914
Leaf N	−2.688	−2.981	−3.010	−3.129
A_{mass}	−1.126	−1.154	−1.134	−1.173
A_{area}	24.450	25.467	25.356	26.184
G_s	−0.227	−0.249	−0.248	−0.262

Each species was classified to the location and functional group combination with the highest value for the classification function. The classification matrices showed that we could more correctly classify species to some combinations than others.

	Colorado forb	Colorado shrub	Wisconsin forb	Wisconsin shrub	Percentage correct
Colorado forb	3	0	0	0	100
Colorado shrub	0	3	0	1	75
Wisconsin forb	1	3	6	1	55
Wisconsin shrub	0	1	0	5	83
Total	4	7	6	7	71

The jackknifed classification matrix was as follows.

	Colorado forb	Colorado shrub	Wisconsin forb	Wisconsin shrub	Percentage correct
Colorado forb	3	0	0	0	100
Colorado shrub	2	0	0	2	0
Wisconsin forb	2	2	5	2	45
Wisconsin shrub	0	2	0	4	67
Total	7	4	5	8	50

We were most successful at classifying species from the Colorado–forb combination and least from the Colorado–shrub combination.

The plot of the scores for the first two discriminant functions shows that there is considerable overlap between the different groups for both functions (Figure 16.2). Colorado forbs were the tightest group and we were most successful at classifying these species.

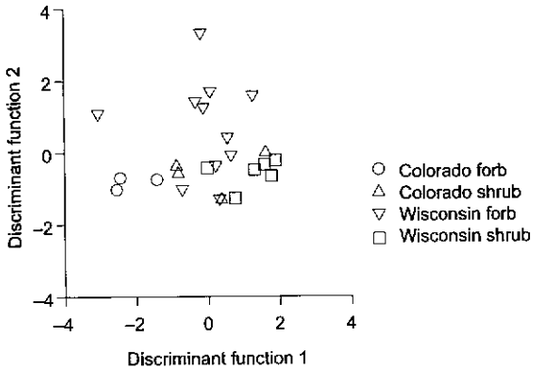

Figure 16.2 Plot of discriminant function scores for each replicate species for the first two functions from discriminant function analysis of data from Reich *et al.* (1999). The five variables were leaf characters: log specific leaf area, leaf N, A_{mass}, A_{area}, and G_s.

(how much of the between-group variance is explained by each function) and the proportion of total variance explained.

Determining which variables contribute most to discriminant functions, and therefore to group separation, is done in the same way as for MANOVA (Section 16.1.3). The relative sizes of the standardized coefficients for each discriminant function indicate which variables are more important to each discriminant function. Also useful are loadings, which measure the correlation between each variable and each discriminant function, although they ignore any correlation between variables. With a large number of variables, stepwise discriminant function analysis can be used, similar to stepwise multiple regression. The stepwise approach enters and removes variables in a model-building process to try and produce a discriminant function with only the "important" variables. Our criticisms of stepwise procedures (see Chapter 6) are just as applicable here and we do not recommend stepwise discriminant analysis.

The test of the H_0 of no difference between group centroids (MANOVA) is usually the first step in a discriminant function analysis because if it is not significant, the discriminant functions will not be very useful for separating groups and therefore classifying observations. Successive discriminant functions can be tested for significance (the second one is tested after the first has been extracted) using the MANOVA tests described in Section 16.1.1.

We can calculate discriminant function scores (z_{ik}) for each observation on each function (k) by simply solving each discriminant function as in Equation 16.1. These scores can be used in a linear discriminant function (LDF) plot (Huberty 1994), with the first discriminant function scores on one axis and the second discriminant function scores on the other axis. Either individual observations or centroids can be plotted. These plots indicate subjectively how similar or different groups are in terms of the discriminant functions. For example, there was a clear separation between sites from Haynes *et al.* (1995) when the first two discriminant functions were plotted (Figure 16.1), although most of the difference was for function one.

LDFs can also be presented as biplots where the loadings (correlations) of each variable on each function are plotted as vectors, scaled so that the vectors are commensurate with the scale of functions scores. The direction of the vectors indicates an increase in the values of the variable towards those objects in that direction on the plot, and the length of the vector indicates the rate of increase. Biplots will be explored in more detail in Chapter 17.

16.2.2 Classification and prediction

The second purpose of a DFA is to classify each observation into one of the groups and assess the success of the classification. A classification equation is derived for each group and is a linear combination of variables like a discriminant function, including a constant (Equation 16.1).

For example, for Delray using the variables from Haynes *et al.* (1995):

$$C_{Delray} = \text{constant} + c_1(\log_{10} \text{Cu}) + c_2(\log_{10} \text{Pb}) + c_3(\log_{10} \text{Ni}) + c_4(\log_{10} \text{Mn}) \quad (16.4)$$

There are four steps in determining and using the classification function for any group.

- The coefficients (c) of the classification equation are termed classification coefficients (Tabachnick & Fidell 1996) and are found by multiplying the within-group covariance

matrix (pooled across all groups) by the matrix of means for each variable for that group.
- The constant for a group is determined by multiplying the matrix of classification coefficients for that group (i.e. the coefficients for each variable) again by the matrix of means for each variable for that group.
- A classification score for each observation for each group is then calculated by using the actual values for each variable to solve the classification equation for that group.
- Each observation is formally classified into the group for which it has the highest score. This may or may not be the actual group from which the observation came.

Tabachnick & Fidell (1996) have provided a fully worked example of the calculations and Huberty (1994) has a more detailed theoretical background.

Discriminant analysis routines in most statistical software provide classification matrices that indicate to which group each observation was classified and whether that classification was correct. The success of classifications of observations will be greater if the groups were clearly distinguishable on the first discriminant function. For example, the stations from Haynes *et al.* (1995) were clearly separable into groups (highly significant MANOVA) and the classification success was also high (Box 16.3). In contrast, there was no significant separation of groups in the Reich *et al.* (1999) data and the classification success was lower (e.g. only six out eleven species correctly classified as being from forbs from Wisconsin – Box 16.4).

One difficulty with the classification methodology we just described is that the classification functions are calculated using all observations and these functions are then used to classify the same observations, i.e. we classify each observation with an equation that already used that observation. One way of avoiding the resulting inherent bias is to use a jackknife procedure (Chapter 2). The classification of each observation is based on group classification functions that are determined when the observation is omitted and only the remaining observations are used to calculate coefficients and constants. In our examples, the jackknife classifications were less successful, but probably more robust, than the usual classifications using all observations (Box 16.3 and Box 16.4). The biggest difference between the usual and jackknifed classifications will often be for groups with the smallest sample size, again illustrated for the classification of the Reich *et al.* (1999) data where our classification success for Colorado shrubs went from 75% to 0% when the jackknife approach was used (Box 16.4).

Table 16.4 Discriminant function analysis from Skelly (1995). The grouping variable was frog species (*Pseudoacris triseriata* and *P. crucifer*) and standardized coefficients and loadings for the three variables on the first discriminant function are provided. Larval period contributed most to the separation between species

Variable	Standardized coefficient	Loading (correlation)
Survivorship	0.208	0.015
Size	0.634	0.593
Larval period	1.015	0.757

Most uses of DFA we have found in the biological literature have focused on description and hypothesis testing, rather than classification. For example, Petit & Petit (1996) derived three discriminant functions to separate four habitats based on ten variables (canopy cover, canopy height, density of various stem sizes) measured around nest boxes occupied by warblers along the Tennessee River. They found that the first function explained 96.7% of the variance and canopy cover was the variable most highly correlated (loading = 0.84) with this first discriminant function. Skelly (1995) used a number of discriminant function analyses in his study of tadpole behaviour and performance. In one, he tested how well three variables (survivorship, size, and larval period) classified individuals into one of two species of frogs (chorus frogs and spring peepers). He presented a single discriminant function, which significantly separated the two species (MANOVA). Larval period had the highest coefficient (1.015) and loading (0.76) for this function, i.e. larval period separated the species more than size and survivorship (Table 16.4).

16.2.3 Assumptions of discriminant function analysis

DFA has the same assumptions as MANOVA (Section 16.1.4). The most important of these assumptions is homogeneity of the within-group variance–covariance matrices, especially for the classification part of discriminant analysis because this is quite sensitive to heterogeneous variance–covariance matrices between groups. This assumption is very difficult to test formally and Tabachnick & Fidell (1996) suggested plotting the scores for each observation for the first two discriminant functions (e.g. Figure 16.1) and checking if the spread of points is similar among the groups. Transformations of variables will often help.

If there is clear heterogeneity across the within-group variance–covariance matrices, you can try fitting quadratic functions instead of the usual linear ones. Quadratic functions include coefficients for squares of the variables and do not assume equal within-group covariances; statistical software usually offers quadratic functions as an option. Quadratic terms are usually highly correlated with the linear term for the same variable. This can result in collinearity problems (Section 16.1.4) and centered variables may need to be used (Chapter 6).

16.2.4 More complex designs

Because DFA is identical to a MANOVA, DFA can be extended to more complex designs, such as factorial designs, as described in Section 16.1.6. However, when focusing on classification, we usually treat each combination of factor levels (cell) as a separate group and use methods developed for single factor designs.

16.3 MANOVA vs discriminant function analysis

MANOVA and DFA are mathematically identical (Tabachnik & Fidell 1996), although the terminology used in the two procedures often differs. In MANOVA, we test whether population centroids, based on a number of response variables, are different between groups. In DFA, we use the response variables to try and predict group membership and also to classify new observations to one or other of the groups with some measure of success of that classification. The linear combination of variables that maximizes the ratio of between-group to within-group variation in MANOVA is the first discriminant function. Discriminant function analysis goes further than MANOVA, however, by calculating additional discriminant functions and using the functions to classify observations to groups.

16.4 General issues and hints for analysis

16.4.1 General issues

- MANOVA can be used to analyze any design where there is more than one response variable and one or more categorical predictor variables and the question of interest concerns the response variables considered simultaneously.
- MANOVA is also used when analyzing partly nested models for "repeated measures" designs where the differences between levels of the within-subjects factor are treated as multiple response variables.
- Although checking the assumptions is more difficult for multivariate analyses compared with univariate analyses, the former are also more sensitive to departures from the assumptions.
- MANOVA and DFA are functionally equivalent, the former emphasizing between-group differences on a single discriminant function, the latter using more than one discriminant function and focusing on classification.

16.4.2 Hints for analysis

- Homogeneity of between-group variances and covariances is important. Keep sample sizes similar and at least ensure homogeneity of variances for each variable separately. Check for outliers with Mahalanobis distance, tested against a χ^2 distribution with p df and a strict significance level (0.001).

- Pillai's trace is the most robust of the test statistics for MANOVA and is recommended.
- The contribution of each variable to a discriminant function is best measured by the standardized coefficients.
- Loadings for each variable on each discriminant function ignore correlations between variables and will have the same pattern between groups as the univariate *F* tests for each variable
- Jackknifed classifications of each observation to each group are probably more reliable than standard classifications because the former do not include the observation being classified when calculating the classification score.

Chapter 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. The first is to reduce many variables to a smaller number of new derived variables that adequately summarize the original 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 this aim. 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 multidimensional 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,...,Y_p) into $k = 1$ to p new uncorrelated variables (Z_1, Z_2, Z_3,..., Z_p) called principal components or factors (Chapter 15). 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, percentage of organic matter, bulk density, N fixation, moisture, redox. Three components explained 75% of the variation, with component 1

representing N and P, component 2 representing moisture and organic matter, and component 3 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+}, and H^+), averaged over three years, and four watershed variables (maximum elevation, sample elevation, length of stream, and 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).

Box 17.1 Worked example of principal components analysis (PCA): chemistry of forested watersheds

The variables in the study of 39 stream sites in New York state by Lovett *et al.* (2000) fell into two groups measured at different spatial scales – watershed variables (elevation, stream length and area) and chemical variables for a site averaged across sampling dates. We only used the chemical variables for the PCA, as a PCA using all variables together was very difficult to interpret. Preliminary checks of the data showed that one stream, Winnisook Brook, was severely acidified with a concentration of H far in excess of the other streams so this site was omitted from further analysis. Additionally, three variables (dissolved organic C, Cl and H) were very strongly skewed and were transformed to $\log_{10}$. Summary statistics for each variable were as follows.

Variable	Mean	Standard deviation
NO_3	22.85	8.61
Total organic N	4.97	1.28
Total N	27.89	8.10
NH_4	1.65	0.73
$\log_{10}$ dissolved organic C	1.83	0.15
SO_4	62.08	5.22
$\log_{10}$ Cl	1.33	0.16
Ca	65.13	13.96
Mg	22.86	5.12
$\log_{10}$ H	−0.67	0.29

First, the PCA was done on all ten chemical variables and 38 streams. We used a correlation matrix because the variables had very different variances, with the variance in Ca concentration much greater than for all other variables, and we did not wish these variances to influence the analysis. Three components had eigenvalues greater than one and explained over 70% of the total variance.

Component	Eigenvalue	Percentage variance
1	3.424	34.239
2	2.473	24.729
3	1.171	11.711

Analysis of the residuals from retaining three components indicated that there were no Q values very different from the rest and all P values were >0.100.

The coefficients of the first three eigenvectors, with their standard errors, are shown below. Note that many of the standard errors are relatively large, some exceeding the value of the coefficient. Considering these standard errors, it appears that SO_4, $\log_{10}$ Cl, Mg (all +ve) and $\log_{10}$ H (−ve) contribute consistently to eigenvector 1. NO_3, total N, and Ca contribute consistently (−ve) to eigenvector 2. Finally, eigenvector 3 contrasts $\log_{10}$ dissolved organic C (+ve) with NH_4 (−ve), although the latter has low precision (large standard error).

Variable	Eigenvector 1	Eigenvector 2	Eigenvector 3
NO_3	−0.261 ± 0.260	−0.519 ± 0.138	0.049 ± 0.212
Total organic N	0.147 ± 0.181	0.299 ± 0.164	0.515 ± 0.404
Total N	−0.228 ± 0.258	−0.510 ± 0.133	0.154 ± 0.274
NH_4	0.228 ± 0.116	0.075 ± 0.192	−0.487 ± 0.478
$\log_{10}$ dissolved organic C	−0.288 ± 0.123	0.147 ± 0.201	0.562 ± 0.198
SO_4	0.368 ± 0.133	−0.225 ± 0.207	0.242 ± 0.221
$\log_{10}$ Cl	0.358 ± 0.110	0.158 ± 0.204	−0.018 ± 0.269
Ca	0.281 ± 0.227	−0.446 ± 0.156	0.081 ± 0.225
Mg	0.472 ± 0.058	−0.015 ± 0.247	0.301 ± 0.145
$\log_{10}$ H	−0.397 ± 0.150	0.281 ± 0.210	0.006 ± 0.218

The loadings (correlations) of each variable on each component reveal a similar pattern to the coefficients of the eigenvectors, although measures of sampling error are not available. Mg (+ve), $\log_{10}$ H (−ve), SO_4 (+ve) and $\log_{10}$ Cl (+ve) correlate highest with component 1, NO_3 (−ve), total N (−ve), and Ca (−ve) correlate with component 2 and $\log_{10}$ dissolved organic C (+ve) correlates with component 3, as do total organic N (+ve) and NH_4 (−ve) slightly less. Note that there are many variables that have moderate correlations (0.4 to 0.6) with the three components.

Variable	Component 1	Component 2	Component 3
NO_3	−0.483	−0.816	0.053
Total organic N	0.272	0.471	0.557
Total N	−0.423	−0.802	0.166
NH_4	0.422	0.118	−0.527
$\log_{10}$ dissolved organic C	−0.533	0.231	0.608
SO_4	0.682	−0.354	0.262
$\log_{10}$ Cl	0.662	0.248	−0.019
Ca	0.520	−0.701	0.087
Mg	0.873	−0.024	0.326
$\log_{10}$ H	−0.735	0.443	0.006

To see if we could get better simple structure for the components, we also applied a varimax (orthogonal) rotation to these eigenvectors. The total variance explained by the first three eigenvectors is the same as before.

Component	Eigenvalue	Percentage variance
1	2.908	29.081
2	2.719	27.185
3	1.441	14.413

The loadings (correlations) of each variable on each rotated component reveal an improved simple structure. SO_4 (+ve), Mg (+ve), $\log_{10}$ H (−ve) and Ca (+ve) correlate strongly with rotated component 1, NO_3 (−ve) and total N (−ve) stand out for component 2, and $\log_{10}$ dissolved organic C (+ve) and NH_4 (−ve) for component 3. The number of variables that have moderate correlations (0.4–0.6) with components has decreased from nine in the unrotated solution to four in the rotated solution.

Variable	Component 1	Component 2	Component 3
NO_3	0.046	−0.943	0.104
Total organic N	0.175	0.578	0.491
Total N	0.126	−0.893	0.192
NH_4	0.090	0.284	−0.617
$\log_{10}$ dissolved organic C	−0.324	−0.038	0.775
SO_4	0.808	0.064	−0.028
$\log_{10}$ Cl	0.393	0.551	−0.206
Ca	0.794	−0.327	−0.182
Mg	0.817	0.448	0.011
$\log_{10}$ H	−0.801	0.002	0.307

We also calculated component scores for each stream for each component based on the rotated solution and correlated the first three components with the watershed variables, adjusting the *P*-values with Holm's sequential Bonferroni method (Chapter 3). Elevation was negatively correlated with component 2. NO_3 and total N load negatively on component 2, indicating that streams with lower elevations also have lower concentrations of nitrogen.

	Max. elevation		Sample elevation		Stream length		Watershed area	
	r	*P*	*r*	*P*	*r*	*P*	*r*	*P*
Component 1	−0.330	0.387	−0.414	0.100	−0.165	1.000	−0.170	1.000
Component 2	−0.528	0.012	−0.496	0.022	−0.066	1.000	−0.048	1.000
Component 3	−0.084	1.000	−0.064	1.000	−0.229	1.000	−0.284	0.664

We also extracted the components based on a covariance matrix, to illustrate the influence that differences in variances have when using a covariance matrix compared with a correlation matrix for a PCA. A much higher proportion of the total variance is explained by the first three components. The eigenvalues are considerably larger than for the correlation matrix because the variables are not standardized to unit variance.

Component	Eigenvalue	Percentage variance
1	223.510	57.262
2	128.595	32.945
3	29.681	7.604

The loadings are now covariances rather than correlations and their pattern among variables is quite different from that based on a correlation matrix. Note that Ca dominates component 1 and this is the variable with the largest variance, with contributions from NO_3 and total N, both with next largest variances. These

two variables also structure component 2, as with the correlation-based PCA, and SO_4 and Mg make up component 3, whereas $\log_{10}$ dissolved organic C did so for the correlation-based PCA.

Note that our preference with these data would be to use a correlation matrix because we did not want the large differences in variances to contribute to our interpretation of components.

Variable	Component 1	Component 2	Component 3
NO_3	4.386	−7.373	0.499
Total organic N	−0.286	0.371	0.342
Total N	4.275	−6.729	1.215
NH_4	0.001	0.164	0.045
$\log_{10}$ dissolved organic C	−0.033	−0.040	−0.026
SO_4	2.514	2.022	3.741
$\log_{10}$ Cl	0.006	0.084	0.025
Ca	13.212	3.853	−1.638
Mg	1.532	3.135	3.340
$\log_{10}$ H	−0.207	−0.143	0.009

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), percentage shrub cover, age (years), 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 (1997), 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 1 and the second is called principal component 2. 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 model 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

Box 17.2 Worked example of principal components analysis (PCA): habitat fragmentation and rodents

Bolger et *al.* (1997) 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. Our aim is to reduce the nine species variables to fewer principal components and examine the relationships between the sites in terms of these components. All species variables were strongly skewed and the variances were very different between variables with many zeros. A fourth root transformation improved normality but did not consistently improve the strength of the linear correlations between variables, so we analyzed raw data. We did separate analyses using covariance and correlation matrices, the latter to remove the effects of the very different variances. With the correlation matrix, the first three components explained over 79% of the variation. With a covariance matrix, more of the variance was contained within component 1 and the first three components explained more than 90% of the variation.

	Correlation		Covariance	
Component	Eigenvalue	Percentage variance	Eigenvalue	Percentage variance
1	4.387	48.746	697.522	78.101
2	1.565	17.393	136.580	15.293
3	1.173	13.029	31.149	3.488

Varimax rotation resulted in different structures, especially for components 1 and 2, but these were not necessarily easier to interpret so we will discuss the unrotated solutions. The loadings (correlations) from a PCA based on a correlation matrix showed that component 1 represented a contrast between the two exotic species and the seven native species and component 3 was a contrast between the two exotics. Component 2 was a little harder to interpret, mainly involving *P. eremicus* and *N. lepida* (+ve) and *N. fuscipes* and *P. fallax* (−ve). Not surprisingly, the loadings (covariances) from a PCA based on a covariance matrix emphasized the species with large variances in their abundances. Component 1 was mainly the two *Peromyscus* species and component 2 was dominated by *P. eremicus*. None of the covariances for component 3 were very strong.

	Correlation			Covariance		
Variable	1	2	3	1	2	3
R. rattus	−0.350	0.293	0.664	−0.252	0.017	−0.249
M. musculus	−0.307	−0.146	−0.800	−1.355	0.013	0.009
P. californicus	0.836	−0.101	0.144	23.697	−3.753	−1.534
P. eremicus	0.750	0.575	−0.139	9.096	10.704	0.499
R. megalotis	0.852	0.051	0.025	3.471	0.283	0.245
N. fuscipes	0.825	−0.509	0.091	5.924	−1.752	4.520
N. lepida	0.685	0.626	−0.163	1.377	1.811	−0.047
P. fallax	0.573	−0.678	0.096	1.427	−1.174	2.799
M. californicus	0.840	0.057	−0.082	0.532	0.333	0.396

The scaling (ordination) scatterplots of component scores for each site from a PCA based on the correlation matrix show that three sites, Sandmark, Alta La Jolla, and Balboa Terrace, stood out from the rest, particularly along components 1 and 2 (Figure 17.3). A biplot of components 1 and 2 including loading vectors for six of the species (Figure 17.3(a)) showed that Sandmark and Alta La Jolla were in the opposite direction of the vector for *M. musculus* and Balboa Terrace was in the opposite direction of the vector for *R. rattus*. So these were sites with high abundance native species and few of the two exotics. Sandmark and Alta La Jolla were at the opposite extreme from Balboa Terrace for component 2, indicating very different numbers of *P. eremicus*, *N. lepida*, *N. fuscipes* and *P. fallax*. Sites with similar patterns for the two exotic species group together on component 3. For all three components, the control mainland sites were not obviously different from the spread of the urban fragments.

Box 17.3 Worked example of principal components analysis (PCA): geographic variation and forest bird assemblages

The data set from Mac Nally (1989) contains the maximum seasonal abundance of 102 species of birds across 37 sites in southeastern Australia. Our main interest is whether we can summarize the relationships between sites based on a small number of components representing the 102 species. This example illustrates problems often faced by ecologists trying to explore multivariate data sets – a large number of variables relative to objects and most of the variables having numerous zero values. We don't present detailed output from the analysis but the PCA based on a matrix of correlations between species showed that 25 components had eigenvalues greater than one and the first five components only explained about 48% of the variation in the original 102 variables. The components themselves were difficult to interpret because of the number of variables, many of them loading moderately on many components (although rotation did help).

A plot of the standardized component scores for the first two components also illustrates the problem of a horseshoe or arch in the pattern of sites along component 1, whereby sites at the extremes are compressed in their relationship to other sites (Figure 17.4). The extremes of this axis represented sites in central Victoria at one end and sites in the Dandenong Ranges close to Melbourne at the other. We will discuss this arch effect further in Chapter 18. Clearly, PCA is not a particularly efficient or interpretable method for examining patterns among sites for these data.

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

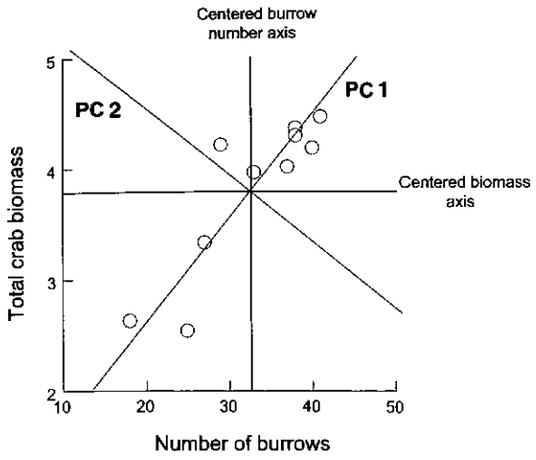

Figure 17.1 Geometric rotation of axes in PCA based on the correlation between crab numbers and burrow numbers of red land crabs on Christmas Island from Green (1997).

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:

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

From Lovett *et al.* (2000):

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

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 1 contains the coefficients for principal component 1, eigenvector 2 for principal component 2, 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.

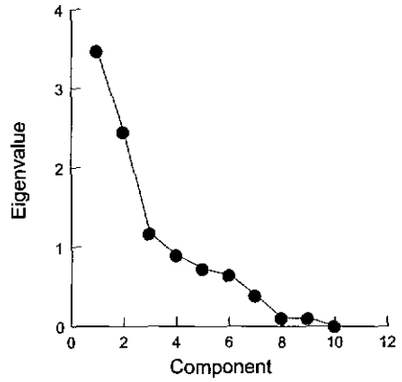

Figure 17.2 An example of a scree plot from the unrotated solution based on a correlation matrix for the data from Lovett *et al.* (2000) – see Box 17.1.

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. It 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. Nonlinear 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, Tabachnick & 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 similar to each other but different from other sites 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.

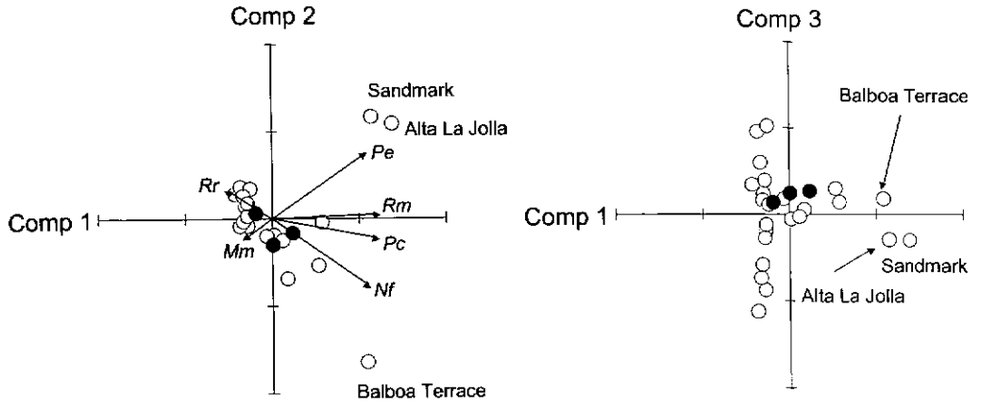

Figure 17.3 PCA scaling (ordination) plots of the 28 sites from Bolger *et al.* (1997) based on a correlation matrix of association between rodent species abundances. Solid circles are mainland control sites, open circles are urban fragments, all axes range from −4 to +4. The left-hand plot is a biplot and vectors of loadings for six of the species have been included, scaled by three. *Rr* is *Rattus rattus*, *Mm* is *Mus musculus*, *Pc* is *Peromyscus californicus*, *Pe* is *Peromyscus eremicus*, *Rm* is *Reithrodontomys megalotis* and *Nf* is *Neotoma fuscipes*.

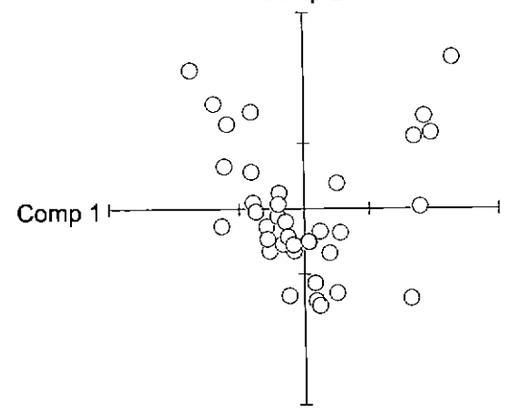

Figure 17.4 PCA scaling (ordination) plot of the 37 sites from Mac Nally (1988) based on a correlation matrix of association between bird species abundances. Axes range from −3 to +3. Note the arch ("horseshoe") pattern in the plot.

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 (left-hand plot in Figure 17.3; 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 17.3 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 initially to 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):

$$\mathbf{b} = \mathbf{U}\mathbf{b}_z \qquad (17.4)$$

In Equation 17.4, $\mathbf{b}$ 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 $\mathbf{b}$ are standardized coefficients and relate to standardized predictor variables. Covariances could be used with just centered 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 kth principal component is (Chaterjee & Price 1991, Jackson 1991):

$$s_{b_k} = \sqrt{\frac{MS_{Residual}}{l_k}} \quad (17.5)$$

In Equation 17.5, $MS_{Residual}$ is from the linear regression on the p principal components. So the standard errors are inversely proportional to 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 $SS_{Regression}$, 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 (1994) 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:

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

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 $(n-1)$ and $(p-1)$, 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:

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

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**):

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

In Equation 17.8, $\mathbf{U}^*$ 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

Box 17.4 Worked example of correspondence analysis (CA): habitat fragmentation and rodents

In this example, we will treat the abundances of each species in each fragment/site from Bolger *et al.* (1997) as a two way contingency table. Although the fragment and mainland sites were very different size, there was little difference in the pattern in the final joint plot of sites and species using raw abundances compared with that based on standardized ("relativized") abundances so that the total abundance at each site was one. Remember that CA partitions the total χ^2 after standardizing by row and column totals so its not surprising that data standardizations do not have much effect. We will just present the analyses of the unstandardized data.

The χ^2 statistic for independence of species and sites is 1722.777 (216 df, $P<0.001$) and total inertia is 1722.777 / 1002 = 1.719. We used the program PC-ORD to run a CA on these data. The total inertia was 1.702, slightly different from above and reflecting the different precision used by different software and possibly the difference between reciprocal averaging and the matrix approach to CA. Don't be surprised by minor variations in output for CA from different programs.

The CA extracted a total of eight eigenvectors (number of species minus one) and the first two explained over 70% of the total inertia.

Axis (component)	Eigenvalue	Percentage intertia
1	0.746	43.41
2	0.459	26.70
3	0.288	16.73
4 to 8	0.227	13.17
Total inertia	1.719	100.00

The joint CA plot of sites and species is in Figure 17.5. We have not included scales on the axes because different software will scale the scores differently – the basic patterns should be similar, however. In contrast to the PCA scaling, where the native species were most influential, the CA scaling emphasizes the two introduced species because there tended to be more of these than expected at a number of sites. It is clear that El Mac and Acuna sites were most different from the remaining sites, with 54th Street also separating out. These sites are associated with the abundance of *R. rattus*, this species having higher than expected abundance at these sites. A number of sites (including 32nd Street Sth, 60th Street, Canon, Florida, Juan, Laurel, Titus and Washington) had similar scores and also had higher than expected abundance of *M. musculus*. Few sites showed marked differences between observed and expected numbers for native species. The mainland sites could not be distinguished from numerous fragments that were associated with the seven native species. It appears that the ecological gradient across these 28 sites is not long and there is no evidence of an arching effect in the scaling plot.

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 the percentage 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.

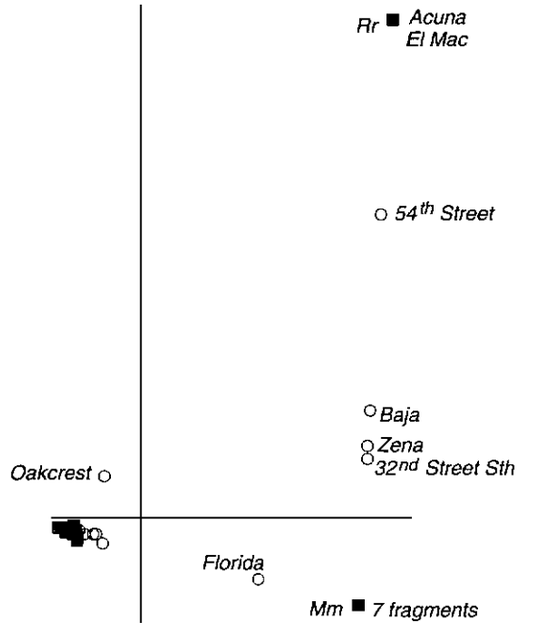

Figure 17.5 CA joint plot of the 28 sites from Bolger *et al.* (1997). Circles are sites (open are fragments, filled are mainland) and shaded squares are species. Labels for species are as in Figure 17.3. Note that some sites have the same scores on both axes and appear as a single point. Some sites also have the identical axis scores to some species. Axis scales have been omitted since different software scales scores differently.

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 and 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. non-metric 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

Box 17.5 | Worked example of canonical correlation analysis: chemistry of forested watersheds

We were interested in testing for correlations between the set of ten chemical variables (Box 17.1) and the set of four watershed variables (maximum elevation, site elevation, stream length and watershed area) for the 39 stream sites in New York state studied by Lovett *et al.* (2000). We omitted the acidified Winnisook site with its extreme concentration of H. We also tried to minimize collinearities by not including highly correlated variables within either set. For the chemical variables, we omitted total N as it was highly correlated with NO_3 and for the watershed

variable, we omitted stream length as it was highly correlated with catchment area. Three of the chemical variables (dissolved organic C, Cl^-, H) and catchment area were transformed to $\log_{10}$ to correct skewness.

Only summaries of the extensive output from statistical software will be presented. Three canonical variate pairs were extracted (only three watershed variables). The correlations for the pairs of canonical variates were as follows.

Canonical variate pair	1	2	3
Canonical correlation	0.874	0.733	0.524

Bartlett's tests of these correlations were as follows.

All three canonical correlations: χ^2 statistic = 77.387, df = 27, $P < 0.001$.
The two sets of variables are not independent so at least the first canonical variate pair is significantly correlated.
Canonical correlations 2 and 3: χ^2 statistic = 33.258, df = 16, $P = 0.007$.
Once the most correlated canonical variate pair (number one) is removed, at least the second canonical variate pair is also correlated, but more weakly.
Canonical correlation 3: χ^2 statistic = 9.775, df = 7, $P = 0.202$.
After removing the first two pairs, the third canonical variate pair is not correlated. In summary, only the first two canonical variate pairs are significantly correlated.

The canonical loadings for the chemical variables (correlations between the canonical variates and the ten chemical variables) were as follows.

Variable	Variate 1	Variate 2
NO_3	−0.740	−0.007
Total organic N	0.098	−0.563
NH_4	−0.058	−0.287
$\log_{10}$ dissolved organic C	−0.262	−0.311
SO_4	0.540	−0.339
$\log_{10}$ Cl	0.616	−0.358
Ca	0.192	0.110
Mg	0.641	−0.551
$\log_{10}$ H	0.331	0.016

The canonical loadings for the watershed variables (correlations between the canonical variates and the four watershed variables) were as follows.

Variable	Variate 1	Variate 2
Max. elevation	−0.862	0.073
Site elevation	−0.699	0.477
$\log_{10}$ area	0.139	0.393

The strong canonical correlation for the first canonical variate pair represents a correlation between a variate contrasting NO_3 with SO_4, $\log_{10}$ Cl and Mg and a variate combining maximum and sample elevation. Sites with higher elevations have greater concentrations of NO_3 and lower concentrations of SO_4, Cl and Mg.

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:

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

In Equation 17.9, **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 in 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 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 eigenvectors 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, percentage of 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 sampling unit. For example, in the study of rodents in habitat fragments, Bolger *et al.* (1997) also recorded the area of the fragment, the percentage 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 not only in scaling the sampling unit and species, such as with CA, but also in examining how the relative positions of sampling unit and species are related to the values of the additional covariates for each sampling unit. 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 sampling unit 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 sampling unit 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 sampling unit 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 sampling unit scores, as weighted averages of species scores, and those sampling unit 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 sampling unit 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 sampling unit and species and vectors representing the correlations between the

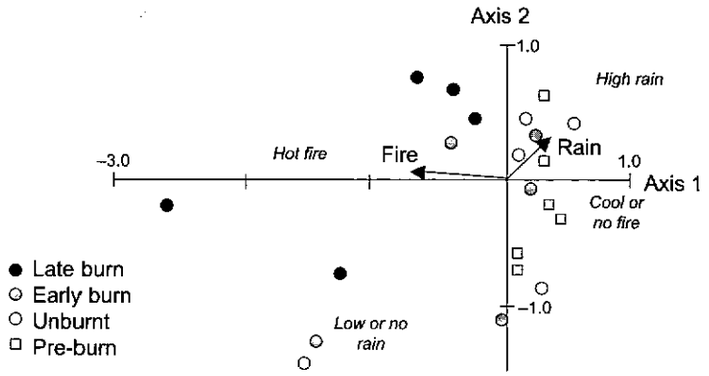

Figure 17.6 CCA biplot for species abundances of beetles in the wet season for four combinations of fire treatment and pre- and post-burn, redrawn from Blanche *et al.* (2001). Pre-burn years were 1988–89 and post-burn years were 1990–94. Vectors for rainfall prior to sampling and fire intensity are included.

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 sampling unit 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 sampling unit 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 to 1994). Abundances of ground-dwelling beetles, sorted to family and species, in each of three replicate 15–20 km^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 fire intensity (Figure 17.6). Treatment–year combinations favored 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 (years), 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 sampling unit and species scaling (ordination). An alternative approach is to scale the sampling unit 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

Box 17.6 Worked example of canonical correspondence analysis (CCA): habitat fragmentation and rodents

We repeated the CA on the rodent data from Bolger *et al.* (1997), but now constrained the axes to be correlated with environmental variables that were also recorded for each site. There were high correlations between the percentage of the area covered with shrubs and the age of the fragment and between the distance to the nearest large source canyon and the distance to the nearest canyon fragment of equal or greater size. To avoid problems with collinearity, only three variables were included: the area of the fragment (ha), the distance to the nearest large source canyon (m) and the age of the fragment (years). Only the 25 habitat fragments were used because the three mainland sites did not have values for the distance measure or age.

The fragments were very different in size so as for the CA presented in Box 17.4, we compared the scaling pattern based on raw abundances and also based on abundances standardized so that all sites had a total abundance of one. The broad pattern in the biplot was the same for both forms of the data so we just present the results for the raw data.

We used the program PC-ORD to do a CCA with the site and species scores standardized to zero mean and unit variance and scaled using the compromise approach between species scores positioned at centroid of site scores and vice versa. The total inertia in the data was 1.702 (as with the CA) and three principal axes were derived.

	Axis 1	Axis 2	Axis 3
Eigenvalue	0.595	0.083	0.039
Percentage inertia (variance)	35.0	4.9	2.3

The CCA biplot (Figure 17.7) shows that the introduced mouse *M. musculus* is associated with older fragments that are further away from source canyons. The other introduced species, *R. rattus*, is more common on small fragments, occurring in the opposite quadrant from the vector for area. Three native taxa (*P. eremicus, M. californianus, N. lepida*) were also more associated with larger fragments. The remaining native taxa occurred more commonly on younger fragments that were closer to source canyons.

The correlations between each CCA axis and the environmental variables showed that the first axis mainly represented fragment age and to a lesser extent distance to nearest source – the vector for age is longer than that for distance in Figure 17.7. This axis is negatively correlated with area – fragments with high scores on this axis are smaller. Axis 2 is positively correlated with all three variables, but more so with area.

Variable	Axis 1	Axis 2
Area	−0.458	0.887
Distance	0.480	0.439
Age	0.806	0.532

A CA on the same 25 fragments revealed a similar pattern as the CA on all 28 sites in Box 17.4.

	Axis 1	Axis 2	Axis 3
Eigenvalue	0.743	0.459	0.279
Percentage inertia (variance)	43.6	26.9	16.4

The first axis of the CA explained 44% of the total inertia, more than for the CCA on the same data with the environmental variables included (35%). The second and third axes also contributed more than in the CCA. This is usually the case when comparing CCA and CA results for the same data set (Jongman *et al.* 1996). The joint plot (Figure 17.8) was almost identical to that in Figure 17.5, indicating that the removal of the mainland sites had little effect on the results of the CA. In contrast to the CCA, Acuna, El Mac and 54th Street stand out as different from the other fragments. Their association with *R. rattus* is also stronger than in the constrained ordination.

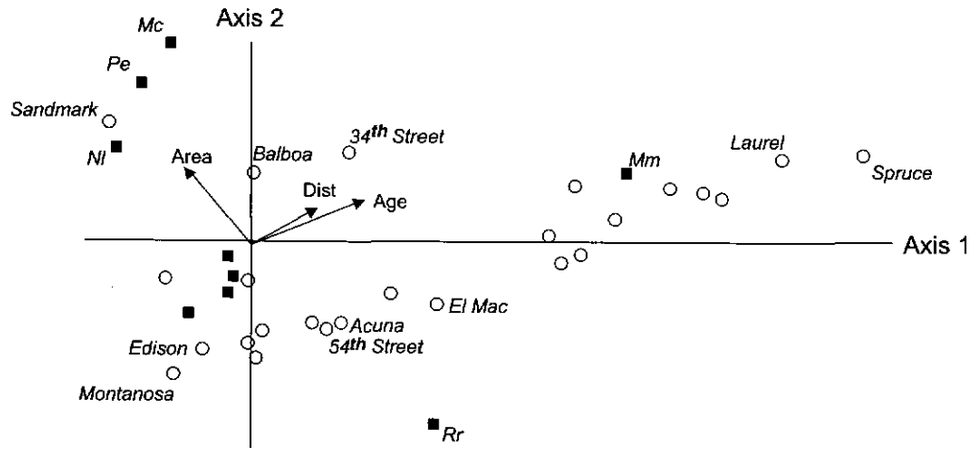

Figure 17.7 CCA biplot of 25 fragment sites from Bolger *et al.* (1997) using LC scores for sites and final scores for species. Circles are sites and shaded squares are species. Labels for species as in Figure 17.3. Axis scales have been omitted since different software scales scores differently.

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 non-metric multidimensional scaling, will be described in Chapter 18.

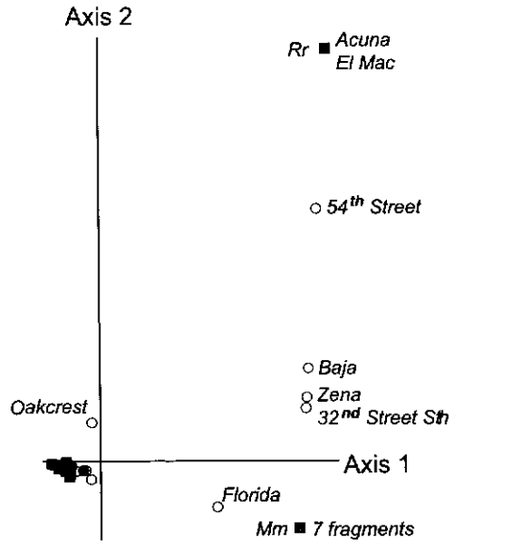

Figure 17.8 CA joint plot of the same 25 fragment sites from Bolger *et al.* (1997) as in Figure 17.7. Note that some sites have the same scores on both axes and appear as a single point. Some sites also have the identical axis scores to some species. The seven fragments associated with *M. musculus* (bottom right quadrant) are Laurel, Canon, Washington, 60th, Juan, Titus and 32nd Street Nth. Axis scales have been omitted since different software scales scores differently.

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 partialing 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 partialing 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. Borcard *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

17.8.1 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 and you do not wish differences in variance for each variable to have any influence on the analysis.
- Eigenvector coefficients and component loadings indicate the contribution of each variable to each component. They should be interpreted in conjunction with their standard errors.
- CA jointly scales objects and variables, based on counts, and emphasizes proportional representation of variables (e.g. species) in objects (e.g. sampling units).
- RDA and CCA constrain the scaling of objects and variables to a linear combination of covariates; for ecological data, this directly incorporates environmental variables in the scaling/ordination of species abundance data.

17.8.2 Hints for analysis

- Secondary rotation of components after an initial PCA will often improve simple structure and interpretability of components.
- Transformation of variables may improve linear relationships between variables and improve the effectiveness of component extraction in PCA.
- Examination of residuals from a PCA can help assess whether the number of retained components is adequate.
- Biplots can be used to represent scaling of objects in PCA, RDA, CA and CCA with correlations of original variables to each component indicated by vectors. It is not the closeness of the end of the vector to objects in the configuration that is important, but the length and angle of the vectors relative to the axes.

Chapter 18

Multidimensional scaling and cluster analysis

In the previous chapter, we were mainly interested in *R*-mode analyses that were based on associations between variables and scaled objects indirectly, although correspondence analysis scaled both objects and variables simultaneously. In this chapter, the primary focus is *Q*-mode analyses that directly scale objects based on similarities or dissimilarities between them. The techniques based on dissimilarities attempt to display the dissimilarities between objects graphically, with the distance between objects on the plot (inter-object distances) representing their relative dissimilarity. The scores for objects on the axes of these scaling plots can be used as variables in subsequent analyses so the techniques in this chapter are also methods for variable reduction. Remember that objects represent sampling or experimental units, such as plots, organisms, aquaria, sites, etc.

Some of the dissimilarity measures for dichotomous and continuous variables were outlined in Chapter 15 (and see Legendre & Legendre 1998 for a much more complete treatment) and all of those dissimilarities can be used with the analyses in this chapter. However, the choice of dissimilarities is a crucial one and different dissimilarities can result in very different patterns in, and interpretations of, the analyses we will describe. Additionally, the form of transformation and/or standardization of variables and/or objects, combined with the dissimilarity measure, can also be very influential.

18.1 Multidimensional scaling

Multidimensional scaling (MDS) refers to a broad class of procedures that scale objects based on a reduced set of new variables derived from the original variables (Cox & Cox 1994). As the name suggests, MDS is specifically designed to graphically represent relationships between objects in multidimensional space. The objects are represented on a plot with the new variables as axes and the relationship between the objects on the plot should represent their underlying dissimilarity. The methods we described in Chapter 17 achieve this scaling indirectly, although MDS is more commonly based on similarities or dissimilarities between objects and was termed "similarities MDS" by Jackson (1991).

The basic data structure we will use in this chapter is similar to that from Chapter 17, a data matrix of i equals 1 to n objects by j equals 1 to p variables. Any two objects will be identified as h and i (*sensu* Legendre & Legendre 1998). The dissimilarities between objects calculated from our data are termed d, so that the dissimilarity between any two objects is d_{hi}. We will call the distance between any two objects (inter-object distances) in the scaling (configuration) plot $\tilde{d}_{hi}$ and it is usually measured as simple Euclidean distance. Unfortunately, there is some inconsistency in the symbols used for dissimilarity and inter-object distance in the literature, with δ commonly used for dissimilarity. This seems inappropriate as Greek letters are usually reserved for unknown parameters.