Chapter 11 *Canonical analysis*

11.0 Principles of canonical analysis

Canonical analysis is the simultaneous analysis of two, or possibly several data tables. Canonical analyses allow ecologists to perform *direct comparisons* of two data matrices (also called "direct gradient analysis"; Fig. 10.4, Table 10.1). Typically, one may be interested in the relationship between a first table describing species composition and a second table containing environmental descriptors, observed *at the same locations*; or two tables of environmental descriptors, e.g. a table about the chemistry of lakes and another about drainage basin geomorphology.

Indirect comparison

Direct comparison

In *indirect comparison* (also called "indirect gradient analysis"; Fig. 10.4), the matrix of explanatory variables **X** does not intervene in the calculation producing the ordination of **Y**. Correlation or regression of the ordination vectors on **X** are computed *a posteriori*. In *direct comparison analysis* (canonical analysis) on the contrary, matrix **X** intervenes in the calculation, forcing the ordination vectors to be maximally related to combinations of the variables in **X**. This description applies to all forms of canonical analysis and in particular to the asymmetric forms described in Sections 11.1 to 11.3.

There is a parallel in cluster analysis, when clustering results are constrained to be consistent with explanatory variables in multivariate regression trees (MRT, Section 8.11) or with structural relationships among observations, either temporal (Subsection 12.6.4) or spatial (Subsection 13.3.2), which are inherent to the sampling design. In constrained clustering or canonical ordination, the results differ in most instances from those of unconstrained analysis and are, hopefully, more readily interpretable. Furthermore, direct comparison analysis allows one to directly test *a priori* ecological hypotheses by (1) bringing out *all* the variance of **Y** that is related to **X** and (2) allowing formal tests of these hypotheses to be performed, as detailed below. Further examination of the unexplained variability may help generate new hypotheses, to be tested using new field observations (Section 13.5).

Canonical form

In mathematics, a *canonical form* (from the Greek κανων, pronounced "kanôn", rule) is the simplest and most comprehensive form to which certain functions, relations, or expressions can be reduced without loss of generality. For example, the

canonical form of a covariance matrix is its matrix of eigenvalues. In general, methods of canonical analysis use eigenanalysis (i.e. calculation of eigenvalues and eigenvectors), although some extensions of canonical analysis have been described that use multidimensional scaling (nMDS) algorithms (Section 9.4).

There are two main families of canonical ordination methods: asymmetric and symmetric. In the asymmetric forms of analysis, there is a response data set and an explanatory data set, which are represented by $\mathbf{Y}$ and $\mathbf{X}$, respectively, in this chapter. The asymmetric methods are redundancy analysis (RDA), canonical correspondence analysis (CCA), and linear discriminant analysis (LDA). In contrast, symmetric methods are used in cases where the two data sets, called $\mathbf{Y}_1$ by $\mathbf{Y}_2$ to mark the symmetry, play the same role in the study; this means that an analysis of $\mathbf{Y}_1$ by $\mathbf{Y}_2$ produces the same result as an analysis of $\mathbf{Y}_2$ by $\mathbf{Y}_1$. These methods include canonical correlation analysis (CCorA), co-inertia analysis (CoIA), Procrustes analysis (Proc), and some others.

Interrelationships among the variables involved in canonical analysis may be represented by the following partitioned covariance matrix, resulting from the concatenation of the $\mathbf{Y}$ (or $\mathbf{Y}_1$, order $n \times p$) and $\mathbf{X}$ (or $\mathbf{Y}_2$, $n \times m$) data sets. The joint dispersion matrix $\mathbf{S_{Y+X}}$ contains blocks that are identified as follows for convenience:

$$\mathbf{S_{Y+X}} = \left[\begin{array}{ccc|ccc} s_{y_1,y_1} & \cdots & s_{y_1,y_p} & s_{y_1,x_1} & \cdots & s_{y_1,x_m} \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ s_{y_p,y_1} & \cdots & s_{y_p,y_p} & s_{y_p,x_1} & \cdots & s_{y_p,x_m} \\ \hline s_{x_1,y_1} & \cdots & s_{x_1,y_p} & s_{x_1,x_1} & \cdots & s_{x_1,x_m} \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & \vdots & & \vdots \\ s_{x_m,y_1} & \cdots & s_{x_m,y_p} & s_{x_m,x_1} & \cdots & s_{x_m,x_m} \end{array}\right] = \left[\begin{array}{cc} \mathbf{S_{YY}} & \mathbf{S_{YX}} \\ \mathbf{S_{XY}} & \mathbf{S_{XX}} \end{array}\right] = \left[\begin{array}{cc} \mathbf{S_{YY}} & \mathbf{S_{YX}} \\ \mathbf{S'_{YX}} & \mathbf{S_{XX}} \end{array}\right] \qquad \textbf{(11.1)}$$

Submatrices $\mathbf{S_{YY}}$ (order $p \times p$) and $\mathbf{S_{XX}}$ ($m \times m$) concern each of the two sets of descriptors, respectively, whereas $\mathbf{S_{YX}}$ ($p \times m$) and its transpose $\mathbf{S'_{YX}} = \mathbf{S_{XY}}$ ($m \times p$) account for the covariances among the descriptors of the two groups, as in eq. 4.27.

Asymmetric, canonical analysis

Asymmetric canonical analysis combines the concepts of ordination and regression. It involves a response matrix $\mathbf{Y}$ and an explanatory matrix $\mathbf{X}$. As it was the case with the simple ordination methods (Chapter 9 and Fig. 11.1a), the asymmetric methods of canonical analysis produce a single ordination of the objects, which may be plotted in a scatter diagram. With the symmetric methods on the contrary, two different ordinations of the objects are produced, one for each data set; see below.

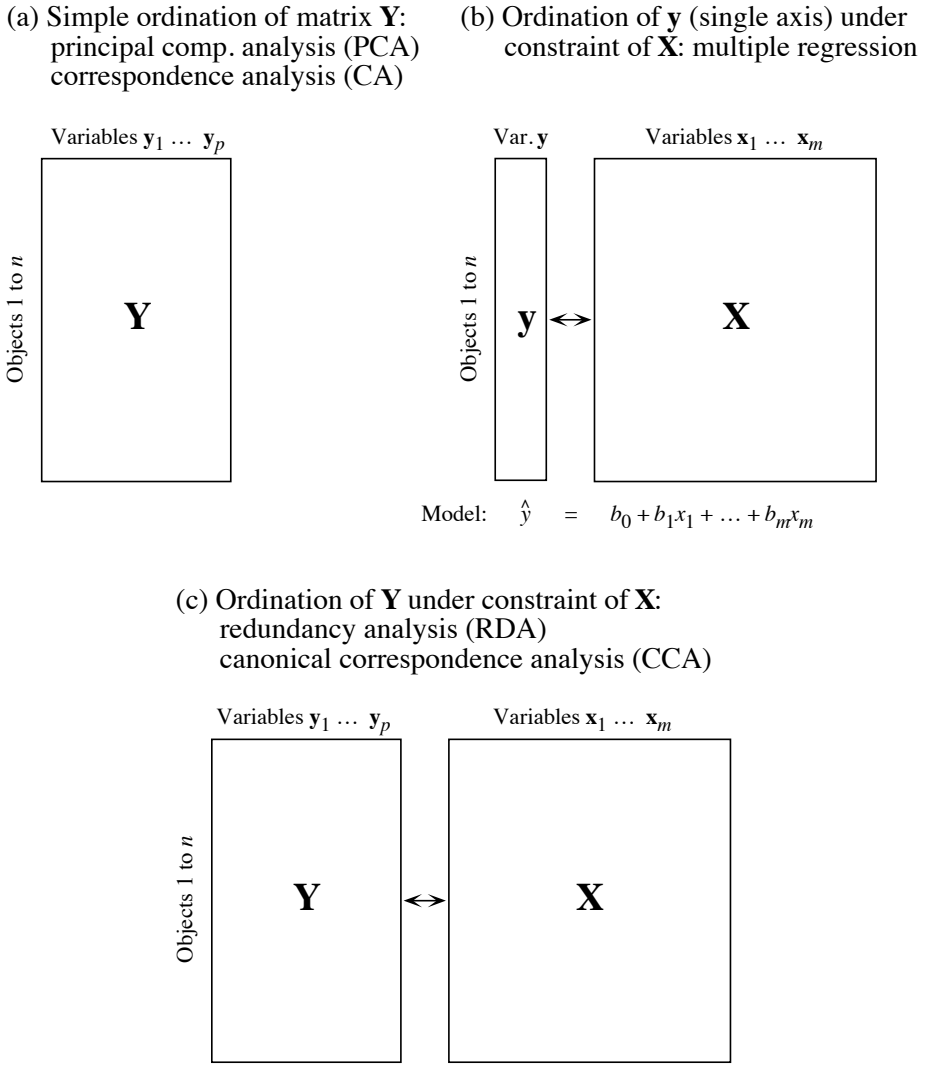


Figure 11.1 Relationships between (a) ordination, (b) regression, and (c) two asymmetric forms of canonical analysis (RDA and CCA). In (c), each canonical axis of **Y** is constrained to be a linear combination of the explanatory variables **X**.

Redundancy analysis (RDA, Section 11.1) and canonical correspondence analysis (CCA, Section 11.2) are related to multiple linear regression. In Subsection 10.3.3, multiple regression was described as a method for modelling a response variable **y** using a set of explanatory variables assembled into a data table **X**. Another aspect of regression analysis must be stressed: while the original response variable **y** provides,

by itself, an ordination of the objects in one dimension, the vector of fitted values (eq. 10.15)

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_p x_{ip}$$

creates a new one-dimensional ordination of the same objects (Fig. 11.1b). The ordinations corresponding to $\mathbf{y}$ and $\hat{\mathbf{y}}$ differ; the square of their correlation is the coefficient of determination (eq. 10.20) of the multiple regression model:

$$R^2_{\mathbf{y}|\mathbf{X}} = [r(\mathbf{y}, \hat{\mathbf{y}})]^2 \qquad \textbf{(11.2)}$$

So, multiple regression creates a correspondence between ordinations $\mathbf{y}$ and $\hat{\mathbf{y}}$, because ordination $\hat{\mathbf{y}}$ is constrained to be optimally (in the least-squares sense) and linearly related to the variables in $\mathbf{X}$. The constraint implemented in multiple regression maximizes R^2. The asymmetric methods of canonical analysis share this property.

Asymmetric canonical analysis combines the properties of two families of methods, i.e. ordination and regression (Fig. 11.1c). It produces ordinations of $\mathbf{Y}$ that are constrained to be linearly related to a second set of variables $\mathbf{X}$, and the results are plotted in reduced space. The way in which the relationship between $\mathbf{Y}$ and $\mathbf{X}$ is established differs among methods of asymmetric canonical analysis.

- In redundancy analysis (RDA, Section 11.1), each canonical ordination axis corresponds to a direction, in the multivariate scatter of objects, that is maximally related to a linear combination of the explanatory variables $\mathbf{X}$. A canonical axis is thus similar to a principal component (Box 9.1). Two ordinations of the objects may be plotted along the canonical axes: (1) linear combinations of the $\mathbf{Y}$ variables (matrix $\mathbf{F}$, eq. 11.17), as in PCA, and (2) linear combinations of the fitted $\hat{\mathbf{Y}}$ variables (matrix $\mathbf{Z}$, eq. 11.18), which are thus also linear combinations of the $\mathbf{X}$ variables. RDA preserves the Euclidean distances among objects in matrix $\hat{\mathbf{Y}}$, which contains values of $\mathbf{Y}$ fitted by regression to the explanatory variables $\mathbf{X}$ (Fig. 11.2); variables in $\hat{\mathbf{Y}}$ are therefore linear combinations of the $\mathbf{X}$ variables.

- Canonical correspondence analysis (CCA, Section 11.2) is similar to RDA. The difference is that CCA preserves the χ^2 distance (as in correspondence analysis), instead of the Euclidean distance among objects in matrix $\hat{\mathbf{Y}}$. Calculations are a bit more complex since matrix $\hat{\mathbf{Y}}$ contains fitted values obtained by weighted linear regression of matrix $\overline{\mathbf{Q}}$ of correspondence analysis (eq. 9.24) on the explanatory variables $\mathbf{X}$. As in RDA, two ordinations of the objects may be plotted.

- In linear discriminant analysis (Section 11.3), the objects are divided into k groups, described by a qualitative descriptor (factor) forming the response matrix $\mathbf{Y}$. The method seeks linear combinations of explanatory variables (matrix $\mathbf{X}$) that explain the classification in $\mathbf{Y}$ by maximizing the dispersion of the centroids of the k groups. This is obtained by maximizing the ratio of the among-object-group dispersion over the pooled within-object-group dispersion (eq. 11.33).

Symmetric, canonical analysis

The *symmetric forms of canonical analysis* described in this book are the following:

- In canonical correlation analysis (CCorA, Section 11.4), the canonical axes maximize the correlation between linear combinations of the two sets of variables $\mathbf{Y}_1$ and $\mathbf{Y}_2$. This is obtained by maximizing the squared among-variable-set correlations (Table 11.10). Two different ordinations of the objects are obtained, one for data set $\mathbf{Y}_1$ and the other for $\mathbf{Y}_2$.

- Co-inertia analysis (CoIA) and Procrustes analysis (Proc) (Section 11.5) search for common structures between two data sets $\mathbf{Y}_1$ and $\mathbf{Y}_2$ describing the same objects. Each object has two representations in the joint plot, one from $\mathbf{Y}_1$ and the other from $\mathbf{Y}_2$.

The application of the various methods of canonical analysis to ecological data was briefly discussed in Section 10.2. In summary, when one of the data sets (**Y**) is to be explained by another (**X**), the asymmetric forms of canonical analysis should be used; the methods are redundancy analysis (RDA) and canonical correspondence analysis (CCA) when **Y** is a full table of response variables, and linear discriminant analysis (LDA) when **Y** contains a classification of the objects. RDA is used when the **X** variables display linear relationships with the **Y** variables, whereas CCA can be used in the cases where correspondence analysis (CA, Section 9.2) would be appropriate for an ordination of **Y** alone. Linear discriminant analysis is applicable when the response data set contains a classification of the objects or an ANOVA factor; in ecology, LDA is used mostly to discriminate among groups of sites using descriptors of the physical environment (Section 11.3). In contrast, canonical correlation analysis (CCorA), co-inertia analysis (CoIA) and Procrustes analysis (Proc) are used to relate two data sets describing the same objects in a correlative framework (Sections 11.4 and 11.5).

Canonical analysis has become an instrument of choice for ecological analysis. A bibliography on the applications of canonical analysis to ecology, covering the period 1986 to 1996, contains a total of 804 entries (Birks *et al.*, 1998). CCorA and discriminant analysis are available in most commercial statistical packages. For RDA, CCA, CoIA and Proc, one must rely on specialized ordination packages and R functions. CANOCO (ter Braak, 1988b) was the first ordination package that made RDA and CCA available to users. These methods are also available in PC-ORD and SYN-TAX 2000. See Section 11.7.

11.1 Redundancy analysis (RDA)

Redundancy analysis (RDA) is the direct extension of multiple regression to the modelling of multivariate response data. The analysis is asymmetric: **Y** ($n \times p$) is a table of response variables and **X** ($n \times m$) is a table of explanatory variables. In RDA, the ordination of **Y** is constrained in such a way that the resulting ordination axes (matrix **Z** below) are linear combinations of the variables in **X**. The difference between

RDA and canonical correlation analysis (CCorA, Section 11.4) is the same as that between simple linear regression (asymmetric analysis) and linear correlation analysis (symmetric); see Box 10.1.

In RDA, the ordination axes are obtained by principal component analysis (PCA, Section 9.1) of a matrix $\hat{\mathbf{Y}}$, computed by fitting the **Y** variables to **X** by multivariate linear regression (details in Subsection 11.1.1). So, in scaling type 1 plots (Subsection 11.1.3), RDA preserves the Euclidean distance among objects (D_1, Chapter 7): the ordination of the points in matrix **Z** is a PCA rotation of the points in $\hat{\mathbf{Y}}$. The ordination axes in **Z** differ, of course, from the principal components that could be computed directly from the **Y** data table because they are constrained to be linear combinations of the variables in **X**. Prior to RDA, the data in **Y** must be at least centred, or transformed following the same principles as in PCA.

1 — Simple RDA

Canonical redundancy analysis was first described by Rao (1964). In his 1973 book (p. 594-595), he proposed the topic to readers as an exercise at the end of his Chapter 8 on multivariate analysis. Rao called the method *Principal components of instrumental variables*. RDA was later rediscovered by Wollenberg (1977) who called the method *Redundancy analysis* by reference to the *redundancy index* of Stewart & Love (1968), which is the proportion of the variance of the response data matrix **Y** that is accounted for by the explanatory matrix **X**. Redundancy is synonymous with explained variance (Gittins, 1985). In his paper, Wollenberg did not refer to Rao's paper (1964) and book (1973). Wollenberg's equation, which only applied to correlation matrices, was less general than that of Rao which involved covariance matrices in general.

Redundancy

Redundancy analysis (RDA) of a response matrix **Y** (with n objects and p variables) by an explanatory matrix **X** (with n objects and m variables) is called simple RDA in Subsections 11.1.1 to 11.1.5, by opposition to partial RDA, described in Subsections 11.1.6 to 11.1.10, which involves a matrix of covariables **W**. Simple RDA involves two computational steps (Fig. 11.2). In the algebraic development that follows, the columns of matrices **Y** and **X** are centred to have means of 0. In computer software, the columns of **X** may be standardized for programming convenience, but this has no effect on the results of the analysis since the matrix of fitted values $\hat{\mathbf{Y}}$ is identical when computed from centred or standardized **X** variables. As in PCA, the variables in **Y** should be standardized if they are not dimensionally homogeneous (e.g. if they are a mixture of temperatures, concentrations, and pH values). Transformations applicable to community composition data (presence-absence or abundance) are described in Section 7.7. As in multiple regression analysis, matrix **X** can contain explanatory variables of different mathematical types: quantitative, multi-state qualitative (e.g. ANOVA factors), or binary variables; see the last five paragraphs of Subsection 10.3.3. If present, collinearity among the **X** variables should be reduced prior to RDA using the methods described for multiple regression in Subsection 10.3.3. Chapters 13 and 14 will show how different expressions of spatial relationships can be used as the explanatory matrix **X** in RDA.

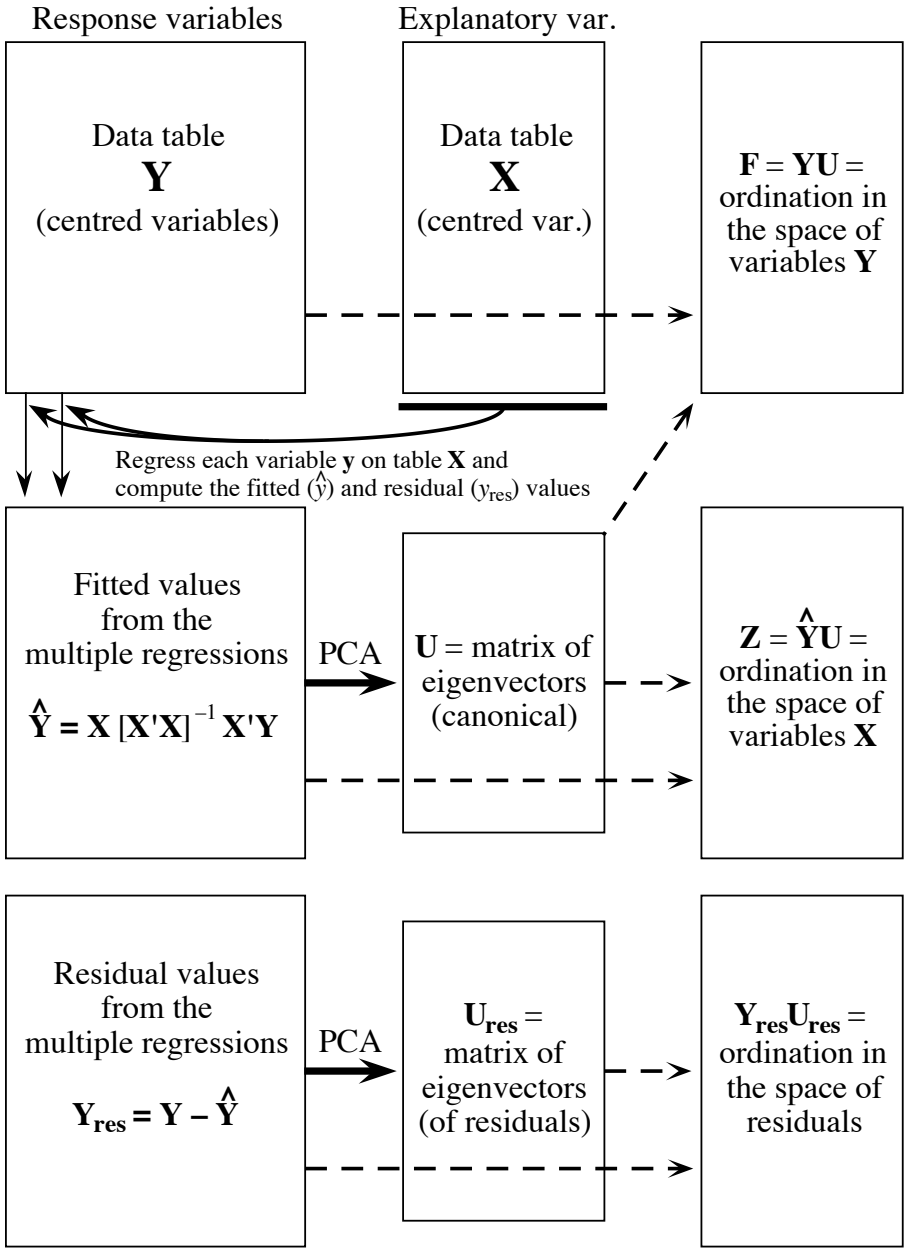


Figure 11.2 Redundancy analysis may be understood as a two-step process: (1) regress each variable in **Y** on all variables in **X** and compute the fitted values; (2) carry out a PCA of the matrix of fitted values to obtain the eigenvalues and eigenvectors. Two ordinations are obtained, one (**F** = **YU**) in the space of the response variables **Y**, the other (**Z** = $\hat{\mathbf{Y}}$**U**) in the space of the explanatory variables **X**. Another PCA ordination can be computed for the matrix of residuals.

The variable distributions should be examined for normality at this stage, as well as bivariate plots within and between the sets **Y** and **X**. Because RDA is a linear model based on multiple linear regression, data transformations (Section 1.5) should be applied as needed to linearize the relationships and make the frequency distributions as symmetric as possible, thus reducing the effect of outliers.

- Step 1 is a *multivariate linear regression* of **Y** on **X** (eq. 10.16), which produces a matrix of fitted values $\hat{\mathbf{Y}}$ through the linear equation:

$$\hat{\mathbf{Y}} = \mathbf{X}\,[\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\mathbf{Y} \qquad \textbf{(11.3)}$$

This is equivalent to a series of multiple linear regressions of the individual variables of **Y** on **X** to calculate vectors of fitted values followed by stacking these column vectors side by side into matrix $\hat{\mathbf{Y}}$. In principle, model II regression should be used when the explanatory variables **X** are *random*, by opposition to *controlled* (Subsection 10.3.2). Ordinary least squares (OLS) are used in eq. 11.3 because, among the model II regression methods, OLS produces fitted values with the smallest error for given values of the predictors (Table 10.4). For efficiency reasons in computer software, matrix $\hat{\mathbf{Y}}$ may be computed through QR decomposition instead of eq. 11.3.

- Step 2 is a principal component analysis of $\hat{\mathbf{Y}}$. This PCA produces the canonical eigenvalues and eigenvectors, as well as matrix **Z** containing the canonical axes (object ordination scores, like matrix **F** in PCA). That step is performed to obtain reduced-space ordination diagrams displaying the objects, response variables, and explanatory variables for the most important axes of the canonical relationship. The PCA step is pertinent only if a significant canonical relationship has been found between **Y** and **X** through an appropriate test of significance (Subsection 11.1.2).

Like the fitted values of a multiple linear regression, which are linear combinations of the explanatory variables, the canonical axes (object ordination scores) are also linear combinations of the explanatory variables in **X**. That RDA axes are linear combinations of the explanatory variables is the fundamental property of RDA (ter Braak, 1987c; ter Braak and Prentice, 1988). Individual canonical axes can be tested for significance to determine which ones are important enough to warrant consideration, plotting, and detailed analysis.

2 — Statistics in simple RDA

After step 1 of RDA, one can compute the following informative statistics.

Redundancy statistic (R^2)

1. From matrices **Y** and $\hat{\mathbf{Y}}$, one can calculate the canonical R^2, which Miller and Farr (1971) called the *bimultivariate redundancy statistic*. This statistic measures the strength of the linear relationship between **Y** and **X**:

$$R^2_{\mathbf{Y}|\mathbf{X}} = \frac{\mathrm{SS}\,(\hat{\mathbf{Y}})}{\mathrm{SS}\,(\mathbf{Y})} \qquad \textbf{(11.4)}$$

where SS ($\hat{\mathbf{Y}}$) is the total sum of squares (or sum of squared deviations from the means) of $\hat{\mathbf{Y}}$ and SS($\mathbf{Y}$) is the total sum of squares of $\mathbf{Y}$. The canonical R^2 is constructed in the same way and has the same meaning as the R^2 statistic in multiple regression (eq. 10.20): it is the proportion of the variation of $\mathbf{Y}$ explained by a linear model of the variables in $\mathbf{X}$.

Note: in the absence of relationship between $\mathbf{Y}$ and $\mathbf{X}$, the expected value of R^2 in multiple regression and in RDA is not 0 but $m/(n-1)$. This is because a matrix $\mathbf{X}$ containing $m = (n-1)$ columns of random numbers produces an R^2 of 1; this surprising fact can easily verify numerically by computing a multiple regression or a RDA with a matrix $\mathbf{X}$ containing $(n-1)$ columns of random numbers. Hence, the expected value (E) of the R^2 produced by a single explanatory variable made of random numbers is $E(R^2) = 1/(n-1)$, and $E(R^2) = m/(n-1)$ for m explanatory variables. This is illustrated in the numerical simulation results presented by Peres-Neto *et al*. (2006).

Adjusted R^2

2. The adjusted R^2 (R^2_a) is computed as in eq. 10.21 (Ezekiel, 1930):

$$R^2_a = 1 - (1 - R^2_{\mathbf{Y}|\mathbf{X}}) \frac{(n-1)}{(n-m-1)} \qquad \textbf{(11.5)}$$

where m is the number of explanatory variables in $\mathbf{X}$ or, more precisely, the rank of the variance-covariance matrix of $\mathbf{X}$.

F-statistic
Overall test

3. The F-statistic for the overall test of significance is constructed as follows (Miller, 1975):

$$F = \frac{R^2_{\mathbf{Y}_{stand}|\mathbf{X}} / mp}{(1 - R^2_{\mathbf{Y}_{stand}|\mathbf{X}}) / (n-m-1)\, p} \qquad \textbf{(11.6)}$$

This statistic is used to perform the overall test of significance of the canonical relationship. The null hypothesis of the test is H_0: the strength of the linear relationship, measured by the canonical R^2, is not larger than the value that would be obtained for unrelated $\mathbf{Y}$ and $\mathbf{X}$ matrices of the same sizes.

When the variables of $\mathbf{Y}$ are standardized ($\mathbf{Y}_{stand}$) and the error distribution is normal, the F-statistic (eq. 11.6) can be tested for significance using the Fisher-Snedecor F-distribution with degrees of freedom $\nu_1 = mp$ and $\nu_2 = p(n-m-1)$. p is the number of response variables in $\mathbf{Y}$. Because m parameters were estimated for each of the p multiple regressions used to compute the vectors of fitted values forming the p columns of $\hat{\mathbf{Y}}$, a total of mp parameters were estimated. This is why there are $\nu_1 = mp$ degrees of freedom attached to the numerator of F. Each multiple regression equation has residual degrees of freedom equal to $(n-m-1)$, so the total number of degrees of freedom of the denominator, ν_2, is p times $(n-m-1)$. Miller (1975) conducted numerical simulations in the multivariate normal case, with combinations of m and p

from 2 to 15 and sample sizes of $n = 30$ to 160. He showed that eq. 11.6 produced distributions of F values that were very close to theoretical F-distributions with the same numbers of degrees of freedom. Additional simulations conducted by Legendre *et al*. (2011, Appendix A) confirmed that the parametric test of significance had correct levels of type I error when **Y** was standardized. This was not the case, however, for non-standardized matrices of response variables **Y** generated with equal or unequal population variances, especially when the error was not normal. Permutation tests always had correct levels of type I error in these simulations. The effect of correlations among the standardized response variables in **Y** on the validity of the parametric test remains to be investigated.

In many instances, the response variables should not be standardized prior to RDA. With community composition data (species abundances), for example, the variances of the species should be preserved in most analyses since abundant and rare species do not play the same roles in ecosystems. A permutation test should always be used in that case. For permutation tests, one can simplify eq. 11.6 of the F-statistic by eliminating the constant p from the numerator and denominator:

Permutation test

$$F = \frac{R^2_{\mathbf{Y}|\mathbf{X}} / m}{(1 - R^2_{\mathbf{Y}|\mathbf{X}}) / (n - m - 1)} \qquad \textbf{(11.7)}$$

While the numerator and denominator of eq. 11.6 indicate the numbers of degrees of freedom for a correct parametric test of F, eliminating p from both does not change the computed value of F. Equation 11.7 is the one used for permutation tests in programs of canonical analysis such as CANOCO and VEGAN's ***rda()***. Actually, the degrees of freedom can be entirely eliminated from statistic equations used in permutation tests since they are invariant across all permutations of the data. However, most computer programs and functions that carry out permutation tests display them to allow comparison with the F-statistic used in parametric tests.

Test of individual axes

4. Individual canonical axes can be tested for significance. Since one deals with complex, multivariate data influenced by many factors, several independent structures may coexist in the response data. If these structures are linearly independent, they should appear on different canonical axes. The results of the tests of individual axes allow researchers to determine which of the canonical axes represent variation that is more structured than random. Canonical axes that do not explain more variation than random should be identified since they do not need to be further considered in the interpretation of the results.

Two methods, called the *forward* and *marginal* testing procedures, can be used for testing individual axes. The forward method was developed by Cajo J. F. ter Braak and implemented in the CANOCO package since version 3.10 (ter Braak, 1990). The marginal method was developed by Jari Oksanen for the ***permutest.cca()*** function of the VEGAN R package; that function carries out tests of significance of the canonical axes when users call the ***anova.cca()*** function with parameter *by="axis"*, after

canonical analysis by functions ***rda()*** or ***cca()***. In a simulation study, Legendre *et al*. (2011) showed that these two methods had correct levels of type I error and comparable powers. This latter study also investigated a third method, the simultaneous test of all canonical axes, which was shown to be invalid.

The null hypothesis for the test of significance of the j^{th} canonical axis is H_0: the linear dependence of the response variables **Y** on the explanatory variables **X** is less than *j*-dimensional. More informally, the null hypothesis is that the j^{th} axis under test explains no more variation than a random axis of the same order (*j*), given the variation explained by the previously tested axes. The test of individual canonical axes can also be carried out in partial RDA (Subsection 11.1.6), a form of RDA that incorporates a matrix of covariables **W**.

3 — The algebra of simple RDA

The eigenanalysis equation for redundancy analysis, which is an asymmetric form of analysis, can be obtained from eq. 11.48 of canonical correlation analysis (CCorA, Section 11.4), which is a symmetric form of analysis, by changing the $\mathbf{S}_{\mathbf{YY}}^{-1}$ matrix (called $\mathbf{S}_{11}^{-1}$ in eq. 11.48) into an identity matrix **I**. The latter does not have to be written after matrix $\mathbf{S}_{\mathbf{YX}}'$ and thus disappears from the equation (Rao, 1973; ter Braak, 1987c):

$$(\mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}_{\mathbf{YX}}' - \lambda_k\mathbf{I})\,\mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.8)}$$

The covariance relationships among the explanatory variables, $\mathbf{S}_{\mathbf{XX}}^{-1}$, remains included in the equation. Equation 11.8 differs from the original formulations by Rao (1964, 1973) and Wollenberg (1977), but it produces the same canonical eigenvalues.

Equation 11.8 is the end result of carrying out the two steps described in the previous subsection, which characterize RDA: (1) a multivariate regression of **Y** on **X** to obtain a matrix of fitted values $\hat{\mathbf{Y}}$, followed by (2) a PCA of that matrix of fitted values. The asymmetric nature of RDA comes from the fact that multivariate regression (eqs. 10.16 and 11.3) is an asymmetric analysis, just as its univariate counterpart, multiple linear regression, where **y** is the response vector and **X** is the explanatory matrix. The developments that follow show that these two computational steps produce eq. 11.8.

1) For *each* response variable in matrix **Y**, compute a multiple linear regression on all variables in matrix **X**. For each regression, the coefficients are computed as follows (eq. 2.19):

$$\mathbf{b} = [\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\mathbf{y}$$

The matrix containing all regression coefficients can be obtained by a single matrix operation (equation without number above eq. 10.16 in Subsection 10.3.3):

$$\mathbf{B} = [\mathbf{X}'\mathbf{X}]^{-1}\,\mathbf{X}'\mathbf{Y} \tag{11.9}$$

where $\mathbf{B}$ ($m \times p$) is the matrix of regression coefficients of all p response variables $\mathbf{Y}$ on the m explanatory variables $\mathbf{X}$.

As in multiple regression, the fitted values $[\hat{y}]$ can be computed by a single matrix operation:

$$\hat{\mathbf{Y}} = \mathbf{X}\,\mathbf{B} \tag{11.10}$$

This is the multivariate extension of eq. 10.1. Replacing $\mathbf{B}$ by the expression from eq. 11.9, eq. 11.10 becomes:

$$\hat{\mathbf{Y}} = \mathbf{X}\,[\mathbf{X}'\,\mathbf{X}]^{-1}\,\mathbf{X}'\,\mathbf{Y} \tag{11.11}$$

which is the multivariate linear regression equation (eq. 10.16). Because the variables in $\mathbf{X}$ and $\mathbf{Y}$ were centred on their respective means, there are no intercept parameters in the column vectors of regression coefficients forming $\mathbf{B}$, and the column vectors in $\hat{\mathbf{Y}}$ are also centred.

2) The covariance matrix corresponding to the table of fitted values $\hat{\mathbf{Y}}$ is computed using eq. 4.6:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = [1/(n-1)]\;\hat{\mathbf{Y}}'\hat{\mathbf{Y}} \tag{11.12}$$

Replacing $\hat{\mathbf{Y}}$ by the expression from eq. 11.11, eq. 11.12 becomes:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = [1/(n-1)]\,\mathbf{Y}'\,\mathbf{X}\,[\mathbf{X}'\,\mathbf{X}]^{-1}\,\mathbf{X}'\,\mathbf{X}\,[\mathbf{X}'\,\mathbf{X}]^{-1}\,\mathbf{X}'\,\mathbf{Y} \tag{11.13}$$

This equation reduces to:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = \mathbf{S}_{\mathbf{YX}}\mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}'_{\mathbf{YX}} \tag{11.14}$$

where $\mathbf{S}_{\mathbf{YY}}$ is the ($p \times p$) covariance matrix among the response variables, $\mathbf{S}_{\mathbf{XX}}$ the ($m \times m$) covariance matrix among the explanatory variables (it is actually a matrix $\mathbf{R}_{\mathbf{XX}}$ when the $\mathbf{X}$ variables have been standardized), and $\mathbf{S}_{\mathbf{YX}}$ is the ($p \times m$) covariance matrix among the variables of the two sets; the order of its transpose $\mathbf{S}'_{\mathbf{YX}} = \mathbf{S}_{\mathbf{XY}}$ is ($m \times p$). If the $\mathbf{Y}$ variables had also been standardized, this equation would read $\mathbf{R}_{\mathbf{YX}}\mathbf{R}_{\mathbf{XX}}^{-1}\mathbf{R}'_{\mathbf{YX}}$, which is the multivariate form of the equation for the coefficient of multiple determination (eq. 4.31).

3) The matrix of fitted values $\hat{\mathbf{Y}}$ is subjected to principal component analysis to reduce the dimensionality of the solution. This corresponds to solving the eigenvalue problem:

$$(\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} - \lambda_k \mathbf{I})\, \mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.15)}$$

which, using eq. 11.14, translates into:

$$(\mathbf{S}_{\mathbf{YX}} \mathbf{S}_{\mathbf{XX}}^{-1} \mathbf{S}_{\mathbf{YX}}' - \lambda_k \mathbf{I})\, \mathbf{u}_k = \mathbf{0} \qquad \textbf{(11.16)}$$

This is the equation for redundancy analysis (eq. 11.8). Different programs may express the eigenvalues in different ways: raw eigenvalues, fractions of the total variance in **Y**, or percentages; see Tables 11.2 and 11.4 for examples.

The matrix containing the normalized canonical eigenvectors $\mathbf{u}_k$ is called **U**. The eigenvectors give the contributions of the descriptors in matrix $\hat{\mathbf{Y}}$ to the various canonical axes. Matrix **U**, of size $(p \times p)$, contains only $\min[p, m, n-1]$ eigenvectors with non-zero eigenvalues, since the number of canonical eigenvectors cannot exceed the minimum of p, m and $(n-1)$:

- It cannot exceed p, which is the dimension of the reference space of matrix **Y**. This is obvious in multiple regression where matrix **Y** contains a single variable; the ordination given by the fitted values $\hat{y}$ is one-dimensional.

- It cannot exceed m, which is the number of variables in **X**. Consider an extreme example: if **X** contains a single explanatory variable ($m = 1$), regressing all p variables in **Y** on this single explanatory variable produces p fitted vectors $\hat{y}$ which all point in the same direction of the p-dimensional space; a principal component analysis of matrix $\hat{\mathbf{Y}}$ of these fitted vectors can only produce one common (canonical) axis.

- It cannot exceed $(n-1)$, which is the maximum number of dimensions required to represent n points in Euclidean space.

The canonical coefficients in the normalized matrix **U** give the contributions of the variables of $\hat{\mathbf{Y}}$ to the canonical axes. They should be interpreted as in PCA. Matrix **U** is used to produce scaling 1 biplot or triplot diagrams, described below. For scaling 2 plots, **U** is rescaled in such a way that the length of each eigenvector is $\sqrt{\lambda_k}$.

If **X** and **Y** are made to contain the same data (i.e. $\mathbf{X} = \mathbf{Y}$), eq. 11.16 becomes $(\mathbf{S}_{\mathbf{YY}} - \lambda_k \mathbf{I})\, \mathbf{u}_k = \mathbf{0}$, which is the equation for principal component analysis (eq. 9.1). The result of RDA is then a principal component analysis of data table **Y**, a fact that was pointed out by Rao (1964, 1973) and by Wollenberg (1977). Another way to look at this point is to say that a RDA of **Y** by **Y** is a PCA of **Y** because $\hat{\mathbf{Y}} = \mathbf{Y}$ in that case.

Additional computations must be done to produce the RDA triplot diagram (below), which contains three types of elements: response variables (e.g. species), objects (e.g. sites), and explanatory variables.

4) The ordination of objects in the space of the response variables **Y** is obtained directly from the centred matrix $\mathbf{Y}_c$, using the standard equation for principal components (matrix **F**, eq. 9.4) and matrix **U** of the eigenvectors $\mathbf{u}_k$ found in eq. 11.16:

$$\mathbf{F} = \mathbf{Y}_c\mathbf{U} \qquad \textbf{(11.17)}$$

Site scores

The ordination vectors (columns of **F**) defined in eq. 11.17 are called the vectors of "site scores". They have variances that are close, but not equal to the corresponding eigenvalues. How to represent matrix **F** in biplots is discussed in point 8 (below).

5) Likewise, the ordination of objects in space **X** is obtained as follows:

$$\mathbf{Z} = \hat{\mathbf{Y}}\,\mathbf{U} = \mathbf{X}\mathbf{B}\mathbf{U} \qquad \textbf{(11.18)}$$

Fitted site scores

As stated above, the vectors in matrix $\hat{\mathbf{Y}}$ are centred on their respective means. The right-hand part of eq. 11.18, obtained by replacing $\hat{\mathbf{Y}}$ by it value in eq. 11.10, shows that this ordination is a linear combinations of the **X** variables. For that reason, these ordination vectors (columns of matrix **Z**) are also called "fitted site scores", or "sample scores that are linear combinations of environmental variables" in program CANOCO. The ordination vectors, defined in eq. 11.18, have variances equal to the corresponding eigenvalues. The representation of matrix **Z** in biplots is discussed in point 8 (below).

The "site scores" of eq. 11.17 are obtained by projecting the original data (matrix **Y**) onto axis *k*; they approximate the observed data, which contain residuals ($\mathbf{Y} = \hat{\mathbf{Y}} + \mathbf{Y}_{res}$, Fig. 11.2). In contrast, the "fitted site scores" of eq. 11.18 are obtained by projecting the fitted values of the multiple regressions (matrix $\hat{\mathbf{Y}}$) onto axis *k*; they approximate the fitted data. Either set may be used in biplots; different programs offer one or the other as the default option. These plots may look very different, so users must decide which one they want to obtain and report in published papers. The practical difference between "site scores" and "fitted site scores" is further discussed in the second example below.

6) The correlation r_k between the ordination vectors in spaces **Y** (from eq. 11.17) and **X** (from eq. 11.18) for dimension *k* is called the "species-environment correlation". It measures the strength of the relationship between the two data sets as expressed by each canonical axis *k*. It should be interpreted with caution because a canonical axis with high species-environment correlation may explain but a small fraction of the variation in **Y**, which is given by the amount (or proportion) of variance of matrix **Y** explained by each canonical axis; see example in Table 11.2.

7) The last important information needed for interpretation is the contribution of the explanatory variables **X** to the canonical ordination axes. Either the regression or the correlation coefficients may be considered:

- Matrix **C** of the canonical coefficients,

$$\mathbf{C} = \mathbf{B}\,\mathbf{U} \qquad \textbf{(11.19)}$$

gives directly the weights of the explanatory variables **X** in the formation of the matrix of fitted site scores. The ordination of objects in the space of the explanatory variables can be found directly by computing **XC**; these vectors of site scores are the same as in eq. 11.18. The coefficients in the columns of matrix **C** are identical to the regression coefficients of the ordination scores from eq. 11.18 on the matrix of standardized explanatory variables **X**; they may thus be interpreted in the same way.

• Correlations may also be computed between the variables in **X**, on the one hand, and the ordination vectors, in either space **Y** (from eq. 11.17) or space **X** (from eq. 11.18), on the other hand. The correlations between **X** and the ordination vectors in space **X**, $\mathbf{R_{XZ}} = \text{cor}(\mathbf{X}, \mathbf{Z})$, are used to represent the explanatory variables in biplots.

Biplot
Triplot

8) In RDA, one can draw *biplot diagrams*, called *biplots*, which contain two sets of points as in PCA (Subsection 9.1.4), or *triplot diagrams* (*triplots*) which contain three sets: the site scores (matrices **F** or **Z**, from eqs. 11.17 and 11.18), the response variables from **Y,** and the explanatory variables from **X**. Each pair of sets of points may be drawn in a biplot. Biplots help interpret the ordination of objects in terms of **Y** and **X**. When there are too many objects, or too many variables in **Y** or **X**, separate ordination diagrams for the response and explanatory variables may be drawn and presented side by side. The construction of RDA biplot diagrams is explained in detail in ter Braak (1994); his conclusions are summarized here. As in PCA, two main types of scalings may be used (Table 9.2):

Scalings in RDA

RDA scaling type 1. — The eigenvectors in matrix **U**, representing the scores of the response variables along the canonical axes, are scaled to lengths 1. The site scores in space **X** are obtained from equation $\mathbf{Z} = \hat{\mathbf{Y}}\mathbf{U}$ (eq. 11.18); these vectors have variances equal to λ_k. The site scores in space **Y** are obtained from equation $\mathbf{F} = \mathbf{YU}$; the variances of these vectors are usually slightly larger than λ_k because **Y** contains both the fitted and residual components and has thus more total variance than $\hat{\mathbf{Y}}$. Matrices **Z** and **U**, or **F** and **U**, can be used together in biplots because the products of the eigenvectors with the site score matrices reconstruct the original matrices perfectly: $\mathbf{ZU}' = \hat{\mathbf{Y}}$ and $\mathbf{FU}' = \mathbf{Y}$, as in PCA (Subsection 9.1.4).

Matrix of biplot scores

In scaling type 1, a quantitative explanatory variable **x** is represented in the biplot or triplot using the vector of correlations of **x** with the fitted site scores, $\mathbf{r_{xZ}} = \text{cor}(\mathbf{x}, \mathbf{Z})$, modified by multiplying each correlation by $\sqrt{\lambda_k / \text{Total variance in } \mathbf{Y}}$ where λ_k is the eigenvalue of the corresponding axis *k*. The whole *matrix of biplot scores* in scaling type 1 ($\mathbf{BS}_1$) for the explanatory variables is computed as follows:

$$\mathbf{BS}_1 = (\text{Total variance in } \mathbf{Y})^{-1/2} \mathbf{R_{XZ}} \boldsymbol{\Lambda}^{1/2} \qquad \textbf{(11.20)}$$

This correction accounts for the fact that, in this scaling, the variances of the site scores differ among axes. The correlation matrix $\mathbf{R_{XZ}}$ was obtained in calculation step 7.

The consequences of this scaling, for PCA, are summarized in the central column of Table 9.2. The graphs resulting from this scaling, called *distance biplots* or *triplots*,

focus the interpretation on the ordination of objects because the distances among objects approximate their Euclidean distances in the spaces corresponding to matrices $\mathbf{Y}$ or $\hat{\mathbf{Y}}$.

Distance triplot

The main features of a distance biplot or triplot are the following: (1) Distances among objects in a biplot are approximations of their fitted Euclidean distances. (2) Projecting an object at right angle on a response variable **y** approximates the fitted value (e.g. abundance) of the object along that variable, as in Fig. 9.3a. (3) The angles among variables **y** are meaningless. (4) The angle between two variables **x** and **y** in the biplot reflect their correlation. (5) Binary explanatory variables **x** may be represented as the centroids of the objects possessing state "present" or "1" for that variable. Examples are given in Subsection 11.1.4. Since a centroid represents a "mean object", its relationship to a variable **y** is found by projecting it at right angle on the variable, as for an object. Distances among centroids, and between centroids and individual objects, approximate Euclidean distances.

RDA scaling type 2. — Alternatively, one obtains response variable scores by rescaling the eigenvectors in matrix $\mathbf{U}$ to lengths $\sqrt{\lambda_k}$, using the transformation $\mathbf{U\Lambda}^{1/2}$ as in PCA (eq. 9.10). The site scores in space $\mathbf{X}$ obtained for scaling 1 (eq. 11.18) are rescaled to unit variances using the transformation $\mathbf{Z\Lambda}^{-1/2}$; this is the same transformation as used in PCA (eq. 9.14) to obtain matrix $\mathbf{G}$ of site scores in scaling 2. Likewise, the site scores in space $\mathbf{Y}$ obtained for scaling 1 are rescaled using the transformation $\mathbf{F\Lambda}^{-1/2}$; the variances of these vectors are usually slightly larger than 1 for the reason explained in the case of scaling 1. Matrices $\mathbf{Z\Lambda}^{-1/2}$ and $\mathbf{U\Lambda}^{1/2}$, or $\mathbf{F\Lambda}^{-1/2}$ and $\mathbf{U\Lambda}^{1/2}$, can be used together in biplots because the products of the eigenvectors with the site score matrices reconstruct the original matrices perfectly: $\mathbf{Z\Lambda}^{-1/2}\mathbf{\Lambda}^{1/2}\mathbf{U}' = \hat{\mathbf{Y}}$ and $\mathbf{F\Lambda}^{-1/2}\mathbf{\Lambda}^{1/2}\mathbf{U}' = \mathbf{Y}$, as in PCA (Subsection 9.1.4).

In scaling type 2, a quantitative explanatory variable **x** is represented in the biplot using the vector of correlations of **x** with the fitted site scores, $\mathbf{r_{xZ}} = \text{cor}(\mathbf{x}, \mathbf{Z})$, obtained in calculation step 7, without further transformation. The matrix of biplot scores ($\mathbf{BS}_2$) for the explanatory variables is then:

$$\mathbf{BS}_2 = \mathbf{R_{XZ}} = \text{cor}(\mathbf{X}, \mathbf{Z}) \qquad \textbf{(11.21)}$$

Note that $\text{cor}(\mathbf{X}, \mathbf{Z\Lambda}^{-1/2})$ produces the same correlations as $\text{cor}(\mathbf{X}, \mathbf{Z})$.

The consequences of this scaling, for PCA, are summarized in the right-hand column of Table 9.2. The graphs resulting from this scaling, called *correlation biplots* or *triplots*, focus on the relationships among the response variables (matrix $\mathbf{Y}$ or $\hat{\mathbf{Y}}$).

Correlation triplot

The main features of a correlation biplot or triplot are the following: (1) Distances among objects in the biplot *are not* approximations of their fitted Euclidean distances. (2) Projecting an object at right angle on a response variable **y** approximates the fitted value (e.g. abundance) of the object along that variable. (3) The angle between two variables **x** and **y** in the biplot reflects their correlation. (4) Projecting an object at right angle on a variable **x** approximates the value of that object along the variable.

Table 11.1 Maximum number of non-zero eigenvalues and corresponding eigenvectors that may be obtained from canonical analysis of a matrix of response variables $\mathbf{Y}(n \times p)$ and a matrix of explanatory variables $\mathbf{X}(n \times m)$ using redundancy analysis (RDA) or canonical correspondence analysis (CCA).

	Canonical eigenvalues and eigenvectors	Non-canonical eigenvalues and eigenvectors
RDA	$\min[p, m, n-1]$	$\min[p, n-1]$
CCA	$\min[(p-1), m, n-1]$	$\min[(p-1), n-1]$

(5) Binary explanatory variables may be represented as described above. Their interpretation is done in the same way as in scaling type 1, except for the fact that the distances in the biplot among centroids, and between centroids and individual objects, do not approximate Euclidean distances.

The type of scaling depends on the purpose of the plot: displaying the distances among objects or the correlations among variables. When most explanatory variables are binary, scaling type 1 is probably the most interesting; when most of the variables in set **X** are quantitative, one may prefer scaling type 2. When the first two eigenvalues are nearly equal, the two scalings lead to very similar plots.

9) Redundancy analysis usually does not completely explain the variation in the response variables (matrix **Y**). During the regression step (Fig. 11.2), regression residuals may be computed for each variable **y**; the residuals are the differences between the observed values y_{ij} in matrix **Y** and the corresponding fitted values $\hat{y}_{ij}$ in matrix $\hat{\mathbf{Y}}$. The matrix of residuals ($\mathbf{Y}_{\text{res}}$ in Fig. 11.2) is also a matrix of size $(n \times p)$. Residuals may be analysed by principal component analysis, leading to $\min[p, n-1]$ non-canonical eigenvalues and eigenvectors (Fig. 11.2, bottom). So, the full analysis of matrix **Y** (i.e. the analysis of fitted values and residuals) may lead to more eigenvectors than a principal component analysis of matrix **Y**: there is a maximum of $\min[p, m, n-1]$ non-zero canonical eigenvalues and corresponding eigenvectors, plus a maximum of $\min[p, n-1]$ non-canonical eigenvalues and eigenvectors, the latter being computed from the matrix of residuals (Table 11.1). When the variables in **X** are good predictors of the variables in **Y**, the canonical eigenvalues may be larger than the first non-canonical eigenvalues, but this is not always the case. If the variables in **X** are not good predictors of **Y**, the first non-canonical eigenvalues, computed on the residuals, may be larger than their canonical counterparts.

In the case where **Y** contains a single response variable, redundancy analysis is simply a multiple linear regression analysis. This is why variation partitioning

(Subsection 11.1.11) can be obtained for a single response variable using an R function, ***varpart()***, which was designed for the analysis of multivariate response data.

Different algorithms can be used in computer programs to compute RDA. One may go through the multiple regression and principal component analysis steps described in Fig. 11.2, or calculate the matrix corresponding to $\mathbf{S_{YX}S_{XX}^{-1}S_{YX}'}$ in eq. 11.8 and decompose it into eigenvalues and eigenvectors using standard eigen-analysis (Section 2.9). Computation of the matrix of fitted values $\hat{\mathbf{Y}}$ can be done by QR decomposition, as explained in Subsection 11.1.1, and eigen-decomposition can be replaced by singular value decomposition (SVD, Section 2.11) as shown for PCA (Subsection 9.1.9). Instead of eigen-decomposition or SVD, an iterative algorithm is used in the program CANOCO to calculate the first four canonical eigenvalues and eigenvectors (ter Braak, 1987c).

4 — Numerical examples, simple RDA

As a first example, consider again the data presented in Table 10.6. For RDA, the first five variables were assembled into matrix **Y** whereas the three spatial variables made up matrix **X**. The **Y** variables were standardized at the beginning of the calculations because they were dimensionally heterogeneous. The results of RDA are presented in Table 11.2. There are min[5, 3, 19] = 3 canonical eigenvectors in this example, and 5 non-canonical PCA axes computed from the residuals. This is a case where the canonical analysis is not very successful: the three canonical eigenvalues account together for only 28% (R^2 = 0.2807) of the variation present in the standardized response data **Y**. The first non-canonical eigenvalues are larger than any of the canonical eigenvalues. The correlations shown in Table 11.2 between the two sets of ordination axes (matrices **F** and **Z**) are rather weak. The ordination of objects along the canonical axes (calculation steps 4 and 5 of the previous subsection) as well as the contributions of the explanatory variables to the canonical ordination axes (calculation step 6) are not reported in the table.

A second example was constructed to illustrate the calculation and interpretation of redundancy analysis. In this artificial example, fish have been observed at 10 sites along a transect perpendicular to the beach of a tropical island, with water depths going from 1 to 10 m (Table 11.3). The first three sites are on sand while the other sites alternate between coral and "other substrate". The first six species avoid the sandy area, possibly because there is little food for them there, whereas the last three are ubiquitous. The sums of abundances for the 9 species are in the last row of the table. Species 1 to 6 come in three successive pairs, with distributions forming opposite gradients of abundance between sites 4 and 10. Species 1 and 2 are not associated with a single type of substrate. Species 3 and 4 are found in the coral areas only while species 5 and 6 are found on other substrates only (coral debris, turf, calcareous algae, etc.). The distributions of abundances of the ubiquitous species (7 to 9) have been produced using a random number generator, fitting the frequencies to a predetermined sum; these species will only be used to illustrate CCA in Section 11.2.

Table 11.2 Results of redundancy analysis (selected output). Matrix **Y** contained the first five variables of Table 10.6 and matrix **X**, the last three.

	Canonical axes			Non-canonical axes				
	I	II	III	IV	V	VI	VII	VIII
Eigenvalues (with respect to total variance of the standardized variables in **Y** = 5)								
	0.8044	0.5864	0.0124	1.4517	1.1165	0.5469	0.3715	0.1101
Fraction of total variance in **Y**								
	0.1609	0.1173	0.0025	0.2903	0.2233	0.1094	0.0743	0.0220
Correlations between the ordination vectors in spaces **Y** and **X**								
	0.7996	0.5936	0.1301					
Normalized eigenvectors (the rows correspond to the five standardized variables in matrix **Y**)								
1	0.2977	0.6173	–0.3441	–0.3345	0.5904	–0.1631	–0.5570	–0.4502
2	–0.6286	0.3455	0.0471	0.1753	0.5936	–0.4738	0.1769	0.6010
3	0.1664	0.4049	0.8922	–0.7254	–0.0735	0.3017	–0.2000	0.5808
4	0.6414	–0.2740	0.0928	0.4459	–0.2856	–0.1018	–0.7857	0.3031
5	0.2778	0.5105	–0.2735	0.3638	0.4605	0.8047	–0.0331	0.0832

RDA was computed using the first six species as matrix **Y**. Had the data been real, they would have been subjected to a Hellinger, chord, or chi-square transformation (Section 7.7) prior to RDA, because of the large proportion of zeros in the data. This is not done here in order to simplify the task of readers who would like to replicate the results. These same data, augmented with species 7 to 9, will be analysed using CCA in Section 11.2. Comparison of the RDA results about species 1 to 6 (Tables 11.4 and Fig. 11.3), on the one hand, to the CCA results about species 1 to 9 (Table 11.7 and Fig. 11.9), on the other hand, allows some comparison of the two methods.

The **Y** variables were not standardized: species abundances do not require standardization since they are all in the same physical dimensions. In most ecological studies, it is important to preserve the variances of the individual species in the analyses because abundant and rare species play different roles in ecosystems. Among the **X** variables, the three binary variables coding for substrate types form a collinear group. Including all three in the cross-product matrix [**X**'**X**] would prevent its inversion because the matrix would be singular (Section 2.8); this would jeopardize

Table 11.3 Artificial data set representing observations (fish abundances) at 10 sites along a tropical reef transect. The variables are further described in the text.

Site No.	Sp. 1	Sp. 2	Sp. 3	Sp. 4	Sp. 5	Sp. 6	Sp. 7	Sp. 8	Sp. 9	Depth (m)	Substrate type Coral	Sand	Other
1	1	0	0	0	0	0	2	4	4	1	0	1	0
2	0	0	0	0	0	0	5	6	1	2	0	1	0
3	0	1	0	0	0	0	0	2	3	3	0	1	0
4	11	4	0	0	8	1	6	2	0	4	0	0	1
5	11	5	17	7	0	0	6	6	2	5	1	0	0
6	9	6	0	0	6	2	10	1	4	6	0	0	1
7	9	7	13	10	0	0	4	5	4	7	1	0	0
8	7	8	0	0	4	3	6	6	4	8	0	0	1
9	7	9	10	13	0	0	6	2	0	9	1	0	0
10	5	10	0	0	2	4	0	1	3	10	0	0	1
Sum	60	50	40	30	20	10	45	35	25				

the calculation of the regression coefficients (eq. 11.9) and of the matrix of fitted values $\hat{\mathbf{Y}}$ (eq. 11.11). It is not necessary, however, to eliminate one of the dummy variables: in well-designed programs for canonical analysis, the last dummy variable is automatically eliminated from the calculations leading to $\hat{\mathbf{Y}}$, but its position in the ordination diagram is estimated in the final calculations. A group of dummy variables coding for a qualitative variable, like the substrate types here, can be replaced by a single factor-type variable in R functions such as VEGAN's ***rda()***.

Results of the analysis are presented in Table 11.4. Scaling type 1 was selected for the biplot in order to illustrate the extra calculation step required to transform the correlations into biplot scores for scaling type 1. The data could have produced 3 canonical axes and up to 6 non-canonical eigenvectors. In this example, only 4 of the 6 non-canonical axes had variances larger than 0. An overall test of significance (Subsection 11.1.2) showed that the canonical relationship between matrices **X** and **Y** was very highly significant ($p = 0.001$ after 999 permutations). The canonical axes explained 66%, 22% and 8% of the variance of the response data, respectively, for a total R^2 of 0.9597 and $R^2_a = 0.9396$. The three canonical axes were all significant ($p < 0.05$) and displayed strong species-environment correlations ($r = 0.999$, 0.997, and 0.980, respectively).

In Table 11.4, the eigenvalues are first shown with respect to the total variance of matrix **Y**, as is customary in principal component analysis. They are also presented as proportions of the total variance of **Y**; these are the eigenvalues provided by CANOCO for PCA and RDA. The species and sites are scaled for a distance triplot (RDA scaling type 1). The eigenvectors, normalized to length 1, provide the "species scores". The

Table 11.4 Results of redundancy analysis of the data in Table 11.3 (selected output). Matrix **Y**: species 1 to 6. Matrix **X**: depth and substrate classes.

	Canonical axes			Non-canonical axes			
	I	II	III	IV	V	VI	VII
Eigenvalues (with respect to total variance of **Y** = 112.88889)							
	74.52267	24.94196	8.87611	4.18878	0.31386	0.03704	0.00846
Fraction of total variance of **Y**							
	0.66014	0.22094	0.07863	0.03711	0.00278	0.00033	0.00007
Cumulative fraction of total variance of **Y** accounted for by axes 1 to *k*							
	0.66014	0.88108	0.95971	0.99682	0.99960	0.99993	1.00000
Normalized eigenvectors ("species scores"): mat. **U** for canonical, $\mathbf{U}_{\mathbf{res}}$ for non-canonical portions (Fig. 11.2)							
Species 1	0.30127	–0.64624	0.39939	–0.00656	–0.40482	0.70711	–0.16691
Species 2	0.20038	–0.47265	–0.74458	0.00656	0.40482	0.70711	0.16691
Species 3	0.74098	0.16813	0.25690	–0.68903	–0.26668	0.00000	0.67389
Species 4	0.55013	0.16841	–0.26114	0.58798	0.21510	0.00000	0.68631
Species 5	–0.11588	–0.50594	0.29319	0.37888	–0.66624	0.00000	0.12373
Species 6	–0.06292	–0.21535	–0.25679	–0.18944	0.33312	0.00000	–0.06187
Matrix **Z** for the canonical part ("fitted site scores", eq. 11.18) and **F** for the non-canonical part (eq. 9.4)							
Site 1	–6.79498	5.49498	2.24897	0.24712	1.14353	0.23570	0.01271
Site 2	–6.96197	5.91719	0.63774	0.00000	0.00000	–0.47140	0.00000
Site 3	–7.12895	6.33941	–0.97349	–0.24712	–1.14353	0.23570	–0.01271
Site 4	–3.55205	–6.52301	4.39356	2.14250	–0.28230	0.00000	0.00141
Site 5	12.69996	0.24686	3.17159	–3.80923	–0.14571	0.00000	0.10360
Site 6	–3.88603	–5.67858	1.17109	0.71417	–0.09410	0.00000	0.00047
Site 7	12.36599	1.09129	–0.05088	0.22968	0.08889	0.00000	–0.22463
Site 8	–4.22000	–4.83415	–2.05138	–0.71417	0.09410	0.00000	–0.00047
Site 9	12.03201	1.93572	–3.27335	3.57956	0.05682	0.00000	0.12103
Site 10	–4.55398	–3.98972	–5.27384	–2.14250	0.28230	0.00000	–0.00141
Correlations of environmental variables with the **Z** site scores							
Depth	0.42265	–0.55914	–0.71325				
Coral	0.98850	0.15079	–0.01178				
Sand	–0.55652	0.81760	0.14771				
Other subs.	–0.40408	–0.90584	–0.12715				
Biplot scores of environmental variables							
Depth	0.34340	–0.26282	–0.20000				
Coral	0.80314	0.07088	–0.00330				
Sand	–0.45216	0.38431	0.04142				
Other subs.	–0.32831	–0.42579	–0.03565				
Centroids, in the triplot, of the sites with code "1" for the BINARY environmental variables							
Coral	12.36599	1.09129	–0.05088				
Sand	–6.96197	5.91719	0.63774				
Other subs.	–4.05301	–5.25636	–0.44014				

"fitted site scores" (matrix **Z**) are obtained from eq. 11.18. They provide the ordination of the objects, computed from $\hat{\mathbf{Y}}$, in the space of the explanatory variables **X**. These axes are orthogonal to one another because they directly result from the PCA of $\hat{\mathbf{Y}}$. The "site scores" (matrix **F**, not shown) in the space of **Y** would be obtained by eq. 11.17. The columns of matrix **F** are, however, not orthogonal to one another because **Y** contains the "residual" components of the multiple regressions (Fig. 11.2). Both the "site scores" (matrix **F**) and "fitted site scores" (matrix **Z**) may be used in RDA triplots.

Correlations of the environmental variables with the ordination vectors can be obtained in two forms: with respect to either the "site scores" (eq. 11.17) or the "fitted site scores" (eq. 11.18). The latter set of correlations is used to draw triplots containing the sites as well as the variables from **Y** and **X**, as done in Fig. 11.3. There were three binary variables in Table 11.3. Each such variable may be represented by the centroid of the sites possessing state "1" for that variable (or else, the centroid of the sites possessing state "0"). These three variables are represented by both arrows (correlations) and symbols (centroids) in Fig. 11.3 to show the difference between these representations. In real-case triplots, only one of the two representations is used.

The fitted site scores in Table 11.4 have much larger ranges of values than the species scores and the biplot scores of environmental variables. Drawing triplots from these tables of values would produce graphs in which the arrows representing the species and environmental variables would be minute and clustered in the centre of the graph. Two strategies are used in computer software: either the tables of output results are modified to make the three sets of values to be drawn (species, sites, environmental variables) commensurable in the graph (this is the case in CANOCO and in VEGAN's function ***rda()***), or the output tables are those produced by the equations of Subsection 11.1.3 but the species and environmental variable arrows are drawn using a different scale than for the site scores (as done in Fig. 11.3).

5 — RDA and CCA of community composition data

Different approaches are available for the canonical analysis of community composition data (Fig. 11.4): the classical approaches (RDA and CCA), transformation-based RDA (tb-RDA), and distance-based RDA (db-RDA). The three approaches are discussed here in turn.

In the classical approach (Fig. 11.4a), the species-environment relationship is analysed by RDA (this section) or by CCA (Section 11.2). In the early applications of canonical analysis to community ecology, the latter was considered preferable for species data tables sampled in highly diversified regions ("long gradients"), which contain many zeros. This is the case, for example, when sampling communities along extensive spatial or temporal gradients, where the species composition may differ greatly between the two ends of the gradient. For groups of sites that were fairly homogeneous in species composition ("short gradients"), RDA was considered appropriate. A wider array of options is now available.

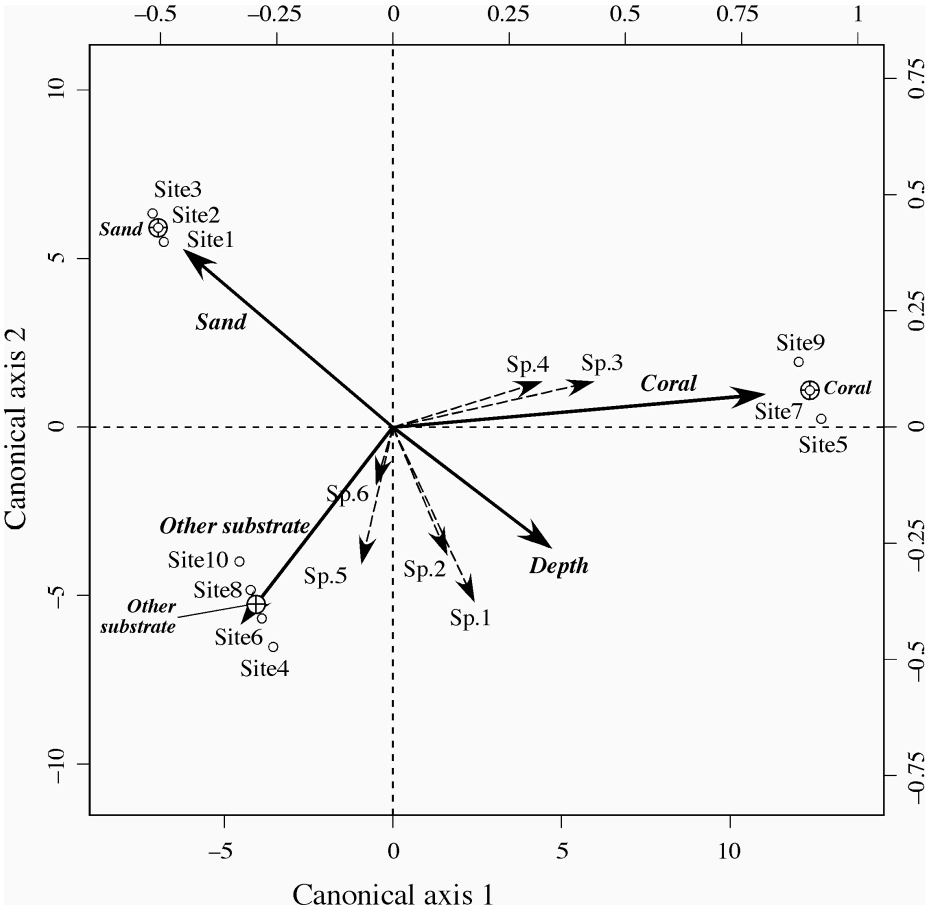


Figure 11.3 RDA triplot of the data in Table 11.3, scaling 1; the numerical results are in Table 11.4. Open circles represent the sites; the site numbers correspond to the site water depths (in m). Dashed arrows are the species. Full-line arrows represent the environmental variables. The sites are positioned in the diagram using the lower and left-hand scales, whereas the species and environmental variables are positioned using the top and right-hand scales. The "centroids of the sites with code 1 for the [three] binary environmental variables" are represented by crossed circles. Binary environmental variables are usually represented by *either* arrows *or* symbols, not both as in this triplot.

Like PCA (Fig. 9.8), RDA can be made to preserve some distance that is appropriate to study composition data along gradients, instead of the Euclidean distance. Figure 11.4b shows that composition data can be transformed using the transformations described in Section 7.7. This is the transformation-based RDA

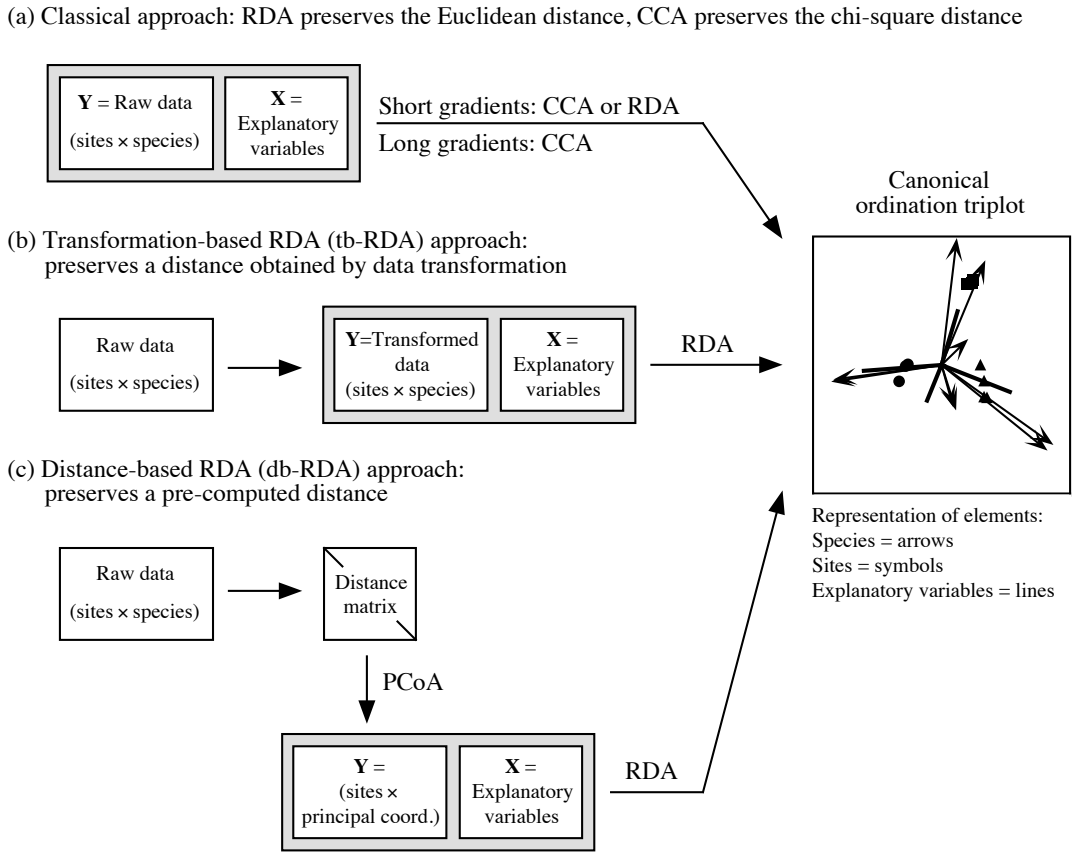


Figure 11.4 Comparison of (a) classical RDA and CCA, and (b and c) alternative approaches forcing RDA to preserve other distances adapted to community composition data. Modified from Legendre & Gallagher (2001).

tb-RDA

(Legendre & Gallagher, 2001), or tb-RDA, approach. RDA computed on data transformed by these equations will actually preserve the chord, profile, Hellinger, chi-square distance or chi-square metric among sites, depending on the transformation used.

One can also (Fig. 11.4c) compute one of the distance functions appropriate for community composition data (Table 7.4), carry out a principal coordinate analysis (PCoA) of the distance matrix, and use *all* the PCoA eigenvectors as input into a RDA.

db-RDA

This is the distance-based RDA, or db-RDA, approach advocated by Legendre & Anderson (1999).

The db-RDA approach must be used in analyses involving distance functions that cannot be obtained by a data transformation followed by RDA (tb-RDA). Among these are most of the coefficients designed for binary data, e.g. Jaccard ($\sqrt{1 - S_7}$) and Sørensen (D_{13} or $\sqrt{1 - S_8}$), as well as quantitative distance measures like the asymmetric Gower coefficient ($\sqrt{1 - S_{19}}$), the geodesic metric (D_4), Whittaker (D_9), Canberra (D_{10}), Clark (D_{11}), percentage difference (D_{14}), and mean character difference modified for species data D_{19}. Distance coefficients intended for other types of data, e.g. symmetric Gower ($\sqrt{1 - S_{15}}$), Estabrook-Rogers ($\sqrt{1 - S_{16}}$), and the generalized Mahalanobis distance for groups of observations, can also be used in canonical ordination through db-RDA. Published studies involving db-RDA include Anderson (1999), Geffen *et al*. (2004) and Lear *et al*. (2008).

6 — Partial RDA

Partial RDA is the analysis of response variables **Y** by explanatory variables **X** in the presence of additional explanatory variables, **W**, called covariables. In partial RDA, the linear effects of the explanatory variables **X** on the response variables **Y** are adjusted for the effects of the covariables **W**, as was done in partial linear regression (Subsection 10.3.5). Partial RDA was first proposed by Davies & Tso (1982, their Section 10.3).

In multiple regression, the partial regression of **y** on **X** in the presence of covariables **W** can be computed in two different ways that were described in Subsection 10.3.5. After computing the residuals of **y** on **W** (noted $\mathbf{y}_{res|\mathbf{W}}$) and the residuals of **X** on **W** (noted $\mathbf{X}_{res|\mathbf{W}}$), one could either (1) regress $\mathbf{y}_{res|\mathbf{W}}$ on $\mathbf{X}_{res|\mathbf{W}}$ or (2) regress **y** on $\mathbf{X}_{res|\mathbf{W}}$. The same partial regression coefficients were obtained in both cases. Between calculation methods, the vectors of fitted values only differed by the value of the intercept of the regression of **y** on $\mathbf{X}_{res|\mathbf{W}}$, which was also the mean of **y**. The R^2 of the first analysis was the partial R^2, whereas that of the second analysis was the semipartial R^2; their square roots were the partial and semipartial correlation coefficients described in Box 4.1.

The same two approaches can be used for partial RDA, which is the extension of partial linear regression to a multivariate response matrix **Y**. First, one computes the residuals of **Y** on **W** (noted $\mathbf{Y}_{res|\mathbf{W}}$) and the residuals of **X** on **W** ($\mathbf{X}_{res|\mathbf{W}}$). Then, one can compute either (1) the RDA of $\mathbf{Y}_{res|\mathbf{W}}$ by $\mathbf{X}_{res|\mathbf{W}}$ or (2) the RDA of **Y** by $\mathbf{X}_{res|\mathbf{W}}$. The two approaches produce the same canonical eigenvalues, eigenvectors and axes. In both approaches, the significance of the canonical axes can be tested using the forward and marginal methods described in Subsection 11.1.2 (paragraph 4). In partial RDA, the canonical axes (matrix **Z**) are linear combinations of the residuals of the explanatory variables **X**, $\mathbf{X}_{res|\mathbf{W}}$, and are orthogonal to the covariables in **W**. The R^2 obtained in the first approach is the partial canonical R^2, whereas that of the second analysis is the semipartial canonical R^2; these two statistics are described in the next subsection. In computer programs, it is customary to use as matrix **F** (eq. 11.17) the matrix obtained from the RDA of $\mathbf{Y}_{res|\mathbf{W}}$ by $\mathbf{X}_{res|\mathbf{W}}$, not the matrix computed in the RDA of **Y** by $\mathbf{X}_{res|\mathbf{W}}$.

Table 11.5 Algorithm for partial RDA in the R language. This is the skeleton of the algorithm used in the ***rda()*** function of the VEGAN package. (Jari Oksanen, personal communication.)

```
pRDA <- function(Y, X = NULL, W = NULL, scale.Y = FALSE)
{
    Y <- scale(as.matrix(Y), center = TRUE, scale = scale.Y)

    if (!is.null(W)) {                  # If covariables W are present
        W <- scale(as.matrix(W), center = TRUE, scale = FALSE)
        Y <- qr.resid(qr(W), Y)
    }
    if (!is.null(X)) {                  # If there are explanatory variables X
        X <- scale(as.matrix(X), center = TRUE, scale = FALSE)
        X <- cbind(X, W)
        Q <- qr(X)
        RDA <- svd(qr.fitted(Q, Y))
        RDA$w <- Y %*% RDA$v %*% diag(1/RDA$d)
        Y <- qr.resid(Q, Y)
    } else {                            # No explanatory variables X nor covariables W
        RDA <- NULL
    }
    RES <- svd(Y)                       # PCA of the residuals
    list(RDA = RDA, RES = RES)
}
```

Table 11.5 presents a very short algorithm for partial RDA, designed by Prof. Jari Oksanen (University of Oulu, Finland). This algorithm handles different cases. (1) If there are covariables (**W**) in the analysis, **Y** is regressed on **W** and residuals $\mathbf{Y}_{\text{res}|\mathbf{W}}$ are computed using QR decomposition (function ***qr()*** in R), which is faster than multivariate regression by matrix inversion (eqs. 10.16 and 11.11). (2) If there are explanatory variables (**X**), RDA is the eigen-decomposition (by SVD through function ***svd()*** in R, Section 2.11) of the fitted values of the multivariate regression of **Y** on **X**. If **X** and **W** are both present, regressing $\mathbf{Y}_{\text{res}|\mathbf{W}}$ on the column concatenation of **X** and **W** produces the same result as a partial regression of $\mathbf{Y}_{\text{res}|\mathbf{W}}$ on $\mathbf{X}_{\text{res}|\mathbf{W}}$ because $\mathbf{Y}_{\text{res}|\mathbf{W}}$ is orthogonal to **W**. (3) A PCA of the residuals is computed. (4) If there are neither explanatory variables **X** nor covariables **W** in the analysis, the result only contains a PCA of **Y** and no RDA is computed.

7 — Statistics in partial RDA

Partial *F*-statistic

For analysis in the presence of **W** containing q covariables (partial RDA), the partial F-statistic is constructed as follows (ter Braak & Smilauer, 2002):

$$F = \frac{\mathrm{SS}(\mathbf{Y}_{\mathrm{fit}}) / m}{\mathrm{SS}(\mathbf{Y}_{\mathrm{res}}) / (n - m - q - 1)} \qquad \textbf{(11.22)}$$

There are several ways of computing the sum of squares of the fitted values $\mathrm{SS}(\mathbf{Y}_{\mathrm{fit}})$ and residuals $\mathrm{SS}(\mathbf{Y}_{\mathrm{res}})$ in the partial RDA case. The most convenient are the following:

$$\mathrm{SS}(\mathbf{Y}_{\mathrm{fit}}) = \mathrm{SS}(\mathbf{Y}_{\mathrm{fit}|(\mathbf{X}+\mathbf{W})}) - \mathrm{SS}(\mathbf{Y}_{\mathrm{fit}|\mathbf{W}})$$

and

$$\mathrm{SS}(\mathbf{Y}_{\mathrm{res}}) = \mathrm{SS}(\mathbf{Y}) - \mathrm{SS}(\mathbf{Y}_{\mathrm{fit}|(\mathbf{X}+\mathbf{W})})$$

where (**X**+**W**) designates the concatenation of **X** and **W** in a single matrix; this is obtained by the operation ***cbind(*****X**,**W***)* in the R language. $\mathbf{Y}_{\mathrm{fit}}$ was noted $\hat{\mathbf{Y}}$ in eq. 11.3 which did not involve covariables **W**.

Semipartial R^2

The semipartial R^2, $R^2_{\mathbf{Y}|\mathbf{X}_{\mathrm{res}|\mathbf{W}}}$, is the proportion of explained variation with respect to the total variation in **Y**. This is the most widely used R^2 statistic in partial RDA because the denominator, which is the total variation in **Y**, forms a common basis for comparisons among analyses using different explanatory matrices **X** and different matrices of covariables **W**. It is the R^2 of the simple RDA of **Y** by $\mathbf{X}_{\mathrm{res}|\mathbf{W}}$:

$$R^2_{\mathbf{Y}|\mathbf{X}_{\mathrm{res}|\mathbf{W}}} = \frac{\mathrm{SS}(\mathbf{Y}_{\mathrm{fit}})}{\mathrm{SS}(\mathbf{Y})} \qquad \textbf{(11.23)}$$

Partial R^2

The partial R^2, $R^2_{\mathbf{Y}_{\mathrm{res}|\mathbf{W}}|\mathbf{X}_{\mathrm{res}|\mathbf{W}}}$, is the proportion of explained variation with respect to the total variation in $\mathbf{Y}$ residualized on the matrix of covariables **W**. Although more rarely used than the semipartial R^2, it is computed as the R^2 of the simple RDA of $\mathbf{Y}_{\mathrm{res}|\mathbf{W}}$ by $\mathbf{X}_{\mathrm{res}|\mathbf{W}}$:

$$R^2_{\mathbf{Y}_{\mathrm{res}|\mathbf{W}}|\mathbf{X}_{\mathrm{res}|\mathbf{W}}} = \frac{\mathrm{SS}(\mathbf{Y}_{\mathrm{fit}})}{\mathrm{SS}(\mathbf{Y}_{\mathrm{res}|\mathbf{W}})} \qquad \textbf{(11.24)}$$

8 — Tests of significance in partial RDA

Permutation test

Tests of significance in partial RDA, using the F-statistic described in eq. 11.22, involve either permutation of the raw data, unrestricted permutation of the residuals of the reduced model (a method proposed by Freedman & Lane, 1983), or unrestricted permutation of the residuals of the full model (a method proposed by ter Braak, 1990, 1992). These methods are described in Anderson & Legendre (1999) for multiple linear regression, which is RDA with a single response variable.

Permute raw data

• In permutation of the raw data (method = "direct" in VEGAN's ***permutest.cca()***), the rows of **Y** are permuted at random to produce the matrix of permuted response data **Y***. This permutation method is used in simple RDA. It can also be used in partial RDA when the covariables do not contain outlying values, e.g. when they represent experimental factors (Subsection 11.1.10, point 4).

Permute residuals of reduced model

• In permutation of the residuals of the reduced model (method = "reduced" in VEGAN's ***permutest.cca()***), one computes the matrix of fitted values $\mathbf{Fit_{Y|W}}$ and the matrix of residuals $\mathbf{Res_{Y|W}}$ of the multivariate regression of **Y** on the matrix of covariables **W**. The rows of $\mathbf{Res_{Y|W}}$ are permuted, producing matrix $\mathbf{Res^*_{Y|W}}$. The matrix of permuted response data, **Y***, is obtained by adding $\mathbf{Fit_{Y|W}}$ (unpermuted) to $\mathbf{Res^*_{Y|W}}$.

Permute residuals of full model

• In permutation of the residuals of the full model (method = "full" in VEGAN's ***permutest.cca()***), one computes the matrix of fitted values $\mathbf{Fit_{Y|XW}}$ and the matrix of residuals $\mathbf{Res_{Y|XW}}$ of the multivariate regression of **Y** on the matrix obtained by concatenation of **X** and **W** by columns into a single matrix. The rows of $\mathbf{Res_{Y|XW}}$ are permuted, producing matrix $\mathbf{Res^*_{Y|XW}}$. The matrix of permuted response data, **Y***, is obtained by adding $\mathbf{Fit_{Y|XW}}$ (unpermuted) to $\mathbf{Res^*_{Y|XW}}$.

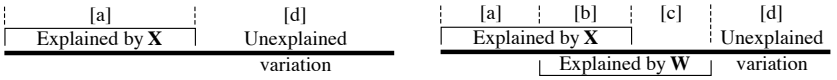
Permutation of the residuals of the reduced and full models were found by Anderson and Legendre (1999) to produce equivalent results. Permutation of the raw data should not be used in partial RDA when the covariables contain outliers. It can, however, be used when partial RDA is used as a form of 2-way MANOVA (Subsection 11.1.10, point 4): in tests of individual factors or the interaction, matrix **W** contains variables coding for the factors or the interaction, and these variables do not have outlier values.

Restricted permutation

Besides these methods, one can also permute the rows of **Y** in a way imposed by the logic of the problem at hand. The most important methods of restricted permutation are: permutation within the levels of a factor or block which is used as a covariable in the study, loop permutation along a time series, and toroidal permutation of the points on a geographic surface (Lotwick & Silverman, 1982).

Methods of permutation of raw data or residuals are compared in Table 11.6 in terms of the permuted portions of variation, in the presence or absence of covariables **W**. *Without covariables*, permutation of raw data involves fraction [a + d] of variation partitioning (Subsection 10.3.5) whereas permutation of residuals of the full model involves [d]. No residual can be computed under a reduced model in the absence of covariables; the method becomes a permutation of raw data. *With covariables*, permutation of residuals may only involve the residuals of the reduced model of the covariables (fraction [a + d]), or the residuals of the full model of the explanatory variables and covariables (fraction [d]). Permutation of the raw data may result in unstable (often inflated) type I error when the covariable contains outliers. This does not occur, however, when using restricted permutations of raw data within groups of a qualitative covariable, which produces an exact test.

Table 11.6 Tests of statistical significance in canonical analysis. Comparison of the methods of permutation of raw data or residuals in terms of the permuted fractions of variation, in the presence or absence of a matrix of covariables **W**. Fractions of variation are noted as in Fig. 10.10: [a] is the variation of matrix **Y** explained by **X** alone, [c] the variation explained by **W** alone, [b] the variation explained jointly by **X** and **W**, and [d] the residual variation.



	Without covariables	With matrix **W** of covariables
Permute raw data	Permute [a + d]	Permute [a + b + c + d]
Permute residuals:		
• reduced model	Equivalent to permuting raw data	Permute [a + d]
• full model	Permute [d]	Permute [d]

9 — Numerical example, partial RDA

Partial RDA provides an answer to the question: what is the partial contribution of one set of explanatory variables when controlling for the effect of another set?

Example 1. — Consider the data in Table 11.3. In that data table, the species can be analysed with respect to *substrate types* while controlling for the effect of *depth*, which is correlated with substrate types. The semipartial R^2 of the analysis is 0.73271; the partial effect of substrate types is highly significant (p = 0.001 after 999 random permutations of the residuals of the reduced model). The two canonical axes produce the triplot shown in Fig. 11.5a.

Example 2. — The converse analysis of the partial effect of *depth* on the distributions of species across the sites while controlling for *substrate types* is also interesting. The semipartial R^2 of this analysis is 0.08274. This is a much weaker effect than that of substrate types, but the partial effect of depth remains significant (p = 0.002 after 999 random permutations of the residuals of the reduced model). A single canonical axis (abscissa of the triplot, Fig. 11.5b) is produced, with the explanatory variable *depth* pointing to the right. Since there is no second canonical axis available, the first axis of the PCA of the residual variation is used as the ordinate of the diagram. This axis separates the coral sites 5, 7 and 9 from the other sites.

These two effects will be considered jointly within the framework of variation partitioning in Subsection 11.1.11 below.

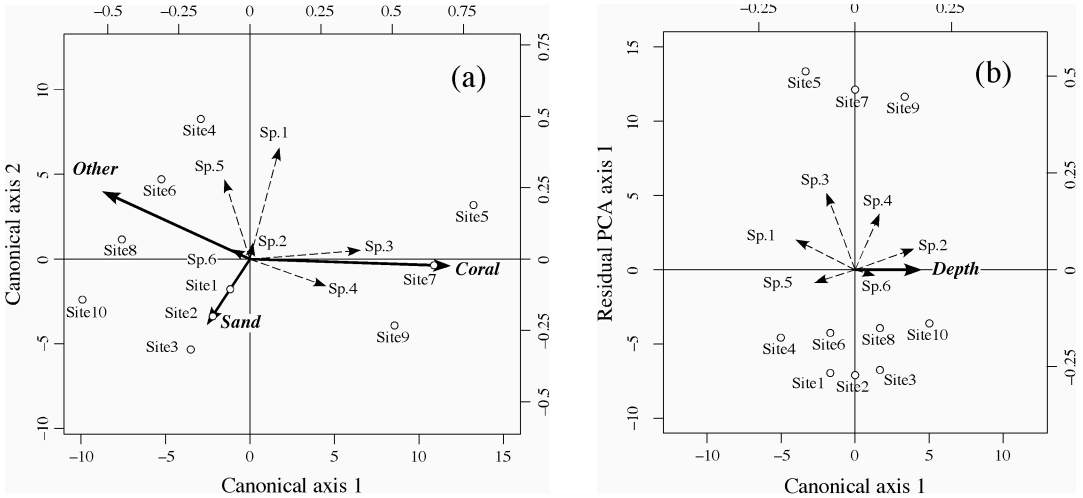


Figure 11.5 Partial RDA triplots of the data in Table 11.3. (a) The explanatory matrix **X** is substrate types, the covariable **W** is depth; (b) the explanatory matrix **X** is depth, the covariable **W** is substrate types. The sites are represented by open circles, the species by dashed arrows, and the explanatory variables in **X** by bold arrows.

10 — Some applications of partial RDA

Partial canonical analysis can be used to investigate a variety of problems. Here are some examples. In most of these applications, CCA (Section 11.2) can be used instead of RDA when **Y** contains frequency data and one wants the analysis to preserve the chi-square distance instead of the Euclidean distance.

Control for effect of **W**

1. *Control for well-known linear effects*. — Consider the case where **W** contains variables whose effects on **Y** are well understood. One wants to control for these well-known effects when analysing the effect of a set of variables **X** on **Y**. For example, one may want to control for the well-known co-variation between phytoplankton assemblages and salinity in a river estuary when analysing the linear effect of nutrient concentrations on phytoplankton. Partial RDA should be used in that case.

Partial effect of a variable

2. *Isolate the effect of a single explanatory variable or factor.* — After conducting a standard RDA as in Subsections 11.1.4, one may want to isolate the partial effect of a single explanatory variable, as in the two examples presented in Subsection 11.1.9 (example 1: a factor with 3 levels; example 2: a quantitative variable). Using all the other explanatory variables as covariates produces a single canonical axis that represents the partial effect of a single quantitative explanatory variable on **Y**. The

corresponding canonical eigenvalue divided by the total variance of **Y** quantifies the partial fraction of the variation of **Y** that is accounted for by that variable (semipartial R^2). The effect of a factor with more than two levels can be isolated by the same method, but then more than one canonical axis are produced because a factor with k levels produces $(k - 1)$ canonical axes.

Related samples

3. *Analysis of related samples*. — Ecological sampling often results in related samples (Box 1.1), where each observation at a site shares some properties with observations at other sites. This is the case, for instance, when sampling different lakes at several depths, the same in all lakes, to study the variation in zooplankton composition. A large portion of the variation in community composition may be associated with the different depths, possibly more than among lakes. Partial RDA offers a way to take this source of variation into account in the analysis of the species-environment relationships. Depth can be coded as a factor or a series of dummy variables, or else as Helmert or polynomial contrasts (Subsection 1.5.7). (Ecologists usually do not hypothesize that zooplankton composition is linearly related to depth, so the covariable depth structuring the sample should be treated as a multi-level factor instead of a quantitative variable.) Including the coding variables in the analysis as a matrix of covariables **W** will effectively control for the effect of the structuring variable. The semipartial R^2 will correctly estimate the partial effect of the environmental variables included in the analysis while controlling for the effect of the structuring variables. Carrying out the analysis for one factor (lakes in this example) while controlling for the effect of the other (depths), and then the opposite, is a form of two-way analysis-of-variance without replication.

Related samples are also obtained when sampling a single lake at different dates and at several depths, or a set of lakes at different dates, the same for all lakes. One may wish to control for the effect of the sampling dates in an analysis of the effect of depths, or lakes, on species composition, bacterial production, or other response variables of interest. As in the previous paragraph, this can be done by using the variable(s) describing the sampling dates as covariable(s) in the analysis. Dates may be represented by dummy variables, or by a quantitative variable whose effect on **Y** is assumed to be linear, or by a sine transformation of the "day of year" (also called "ordinal date", and often "Julian date"[*]), etc. The analysis will effectively control for the effect of dates (days, weeks, years, …) if they only affect the means of the response variables and nothing else. If there is an interaction between sampling dates and the other environmental or spatial variables included in the analysis, the effect of dates cannot be controlled through this simple approach. In the presence of an interaction, the interaction terms must remain in the analysis for the model to be valid (see

[*] The "day of year", also called "ordinal date", is a calendar date starting on 1st January and ranging between 001 and 366. The "Julian day" is used in the "Julian date" system of time measurement, mostly by the astronomy community, where the interval of time is stated in days and fractions of a day since 1st January 4713 BC Greenwich noon. The use of "Julian date" to refer to the day of year, although technically incorrect, is widespread in ecology and other natural sciences. Readers may check the entries "Julian day" and "ordinal date" on Wikipedia.

paragraph 4 below). How to test the space-time interaction in the absence of replication is described in Subsection 14.5.1.

In the same way, one can control for the effect of the sampling locations. Sampling locations may be represented by dummy variables, or by a trend-surface polynomial (Chapter 13) or a set of spatial eigenfunctions (Chapter 14) derived from the geographic coordinates of the sites. The caveat of the previous paragraph concerning interactions applies here as well.

MANOVA by RDA

4. *MANOVA by RDA*. — Partial canonical analysis may be used, instead of MANOVA, to analyse a multivariate response data matrix **Y** in cross-factor experimental designs, including tests of significance for the main effects and the interaction term. For a single experimental factor, the analysis can be conducted using simple RDA or CCA. For two or more factors and their interactions, partial RDA or CCA must be used.

In MANOVA by RDA involving two or more crossed factors, the factors and their interactions are coded by Helmert contrasts (Subsection 1.5.7). The interaction between factors A and B, for example, is represented by a series of variables obtained by the Hadamard product of each Helmert variables coding for factor A by each Helmert variable coding for factor B. Three partial RDAs are necessary to conduct an analysis involving two crossed factors:

- Test the interaction through a RDA of **Y** with the interaction variables in the explanatory matrix **X** and the Helmert variables coding for factors A and B together in the matrix of covariables **W**. If the interaction is significant, analyse separately the effect of factor A in each class of factor B, and conversely the effect of factor B in each class of factor A, because a significant interaction indicates that the effects of factor A on **Y** depend on the levels of factor B, and conversely. If the interaction is not significant, proceed with the next two steps.

- Test the effect of factor A on **Y** through a RDA of **Y** with the variables coding for A in **X** in the presence of a matrix of covariables **W** containing all variables coding for B and the interaction.

- Test the effect of factor B on **Y** through a RDA of **Y** with the variables coding for B in **X** in the presence of a matrix of covariables **W** containing all variables coding for A and the interaction.

The results of the three tests of significance can be assembled in a MANOVA table.

The condition of homogeneity of the variance-covariance matrices applies to this form of MANOVA, as it does to regular MANOVA. It can be tested by the function ***betadisper()*** of the VEGAN package, which implements the testing method described by Anderson (2006). A fully worked out example of MANOVA by RDA, including a test of homogeneity of the multivariate dispersion, is given in Section 6.3.2.8 of Borcard *et al*. (2011).

As stated in Subsection 11.1.5, when **Y** is a matrix of species presence-absence or abundance data, one can either transform **Y** prior to MANOVA by RDA using the transformations described in Section 7.7 (transformation-based RDA, tb-RDA) to force the partial RDA to preserve the distance that is implicit in the transformation, or use partial CCA to preserve the chi-square distance among sites. Else, one can use the distance-based RDA method (db-RDA, Subsection 11.1.5) to preserve some other distance function appropriate for community data.

Principal response curves

5. *Principal response curves (PRC).* — *Principal response curves* is another form of MANOVA; it was developed by van den Brink and co-authors (1998, 1999, 2003, 2009) to analyse the results of experiments conducted over time, that involved multivariate response data (e.g. community composition data). PRC is a special case of RDA with a single factor for treatments and a single factor representing the time series of repeated observations. The method studies the changes in the multivariate (e.g. species) response variables associated with the treatments over time. In this type of analysis, one is interested in displaying the values of the coefficients (contrasts against the control level) computed for the first RDA axis representing the effects of treatment along time. Significance of the canonical relationship and of the first axis can be tested when there is replication in the experimental design. This is an omnibus test: H_0 corresponds to 'no treatment effect'. H_1 includes all functional forms that the treatment effects can take, i.e. main effect and/or interaction. No effect at all produces coinciding treatment lines in the plot. One can also test separately the effect of the main factor (treatment) and, when there is replication, the treatment-time interaction. A significant interaction indicates that the treatment levels had different effects on the response data at different times; it is displayed as non-parallel or crossing treatment lines in the plot.

Partial PCA

6. *Partial PCA.* — Partial principal component analysis is the PCA of a response data table **Y** residualized on a set of explanatory variables. This method allows researchers to examine the multivariate structure of the data after removing the effect of the **X** variables on **Y**, which may already be well understood, by computing residuals. Note that the results of a partial PCA differ from those of a partial RDA.

The three tables represented at the bottom of Fig. 11.2 illustrate how partial PCA is carried out: the residuals of **Y** by **X** are computed, followed by a PCA of the matrix of residuals. Alternatively, since RDA of **Y** by **Y** is a PCA of **Y**, as shown in Subsection 11.1.3, partial PCA can be obtained by computing a partial RDA of **Y** by **Y** with **X** as covariables. In R, the regression function ***lm()*** can be used to easily obtain a matrix of residuals: res = ***residuals(lm(***$\mathbf{Y} \sim \mathbf{X}$***))***. With the data of Table 11.3 for instance, one could examine the residual structure, after controlling for depth and substrate, by plotting a PCA biplot of the non-canonical axes shown in Table 11.4. In spatial analysis, one could detrend the data by computing the regression residuals of **Y** on the geographic coordinates of the sites before computing a PCA.

Selection of explanatory variables

7. *Selection of explanatory variables.* — Different selection methods are available in canonical analysis, as well as in multiple regression (Subsection 10.3.3): backward,

forward, and stepwise. Function ***ordistep()*** in VEGAN offers all three methods of selection. In *forward selection*, the significance of the partial *F*-statistics associated with all candidate variables is tested using permutations, and the explanatory variable that has the most significant partial effect is selected if its p-value satisfies the "p-to-enter" significance level; in case of equality, the variable that has the lowest value of the Akaike Information Criterion (*AIC*, eq. 10.22)* is selected for inclusion in the model. The *backward* option sequentially drops variables from the model using the same criteria of significance (the highest p-value is compared to a "p-to-exclude" significance level) and *AIC* in case of equality (the variable whose removal produces the model with the lowest *AIC* value is excluded). The *stepwise* option tries to eliminate variables from the model (*backward*) after each *forward* step. In this function, "best" refers to the most significant variable.

Functions ***ordiR2step()*** of VEGAN and ***forward.sel()*** of PACKFOR offer the forward method. In these functions, the basic algorithm, developed by ter Braak (1990), is the same as in CANOCO: considering the variables already selected, the explanatory variable with the highest partial R^2 is selected if the additional contribution of that variable is significant (permutation test) at a pre-selected significance level. In these functions, "best" refers to the variable that explains the largest portion of the remaining unexplained variance of **Y**. These two functions offer the option of applying a second stopping criterion proposed by Blanchet *et al*. (2008b): the selection stops either when the tested variable has a p-value higher than the pre-selected significance level or when the adjusted R^2 of the full model, before any selection, is exceeded.

Before applying these variable selection methods, one should look at the collinearity among the variables in **X** by computing variance inflation factors (VIF, eq. 10.17), and remove variables as needed to reduce collinearity. Borcard *et al*. (2011) present examples of forward selection prior to RDA.

11 — Variation partitioning by RDA

Variation partitioning, described in Subsection 10.3.5 for univariate response data, was originally developed for the analysis of multivariate response tables (Borcard *et al*., 1992; Borcard & Legendre, 1994). It is especially useful for partitioning the variation of community composition data with respect to two or more sets of explanatory variables.

Ecological application 11.1a

The method is illustrated here using fish assemblage data (27 species) from 29 sampling sites along the Doubs River in eastern France. The calculations reported in this application were done

* The *AIC* criterion is not meant to identify the "true" model (which is only known in simulation studies) among several alternative models, but to find the best predictive model for new observations.

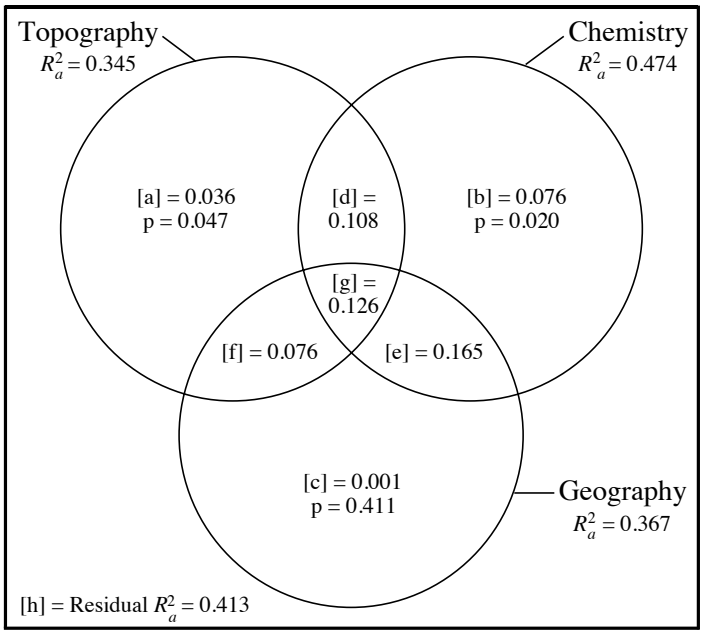


Figure 11.6 Venn diagram illustrating the results of variation partitioning of the Doubs River fish assemblage data (29 sites) among three sets of explanatory variables: Topography, Chemistry and Geography. The fractions of variation are identified by letters [a] to [h]. The value next to each identifier is the adjusted R^2 (R^2_a). The circles, drawn by the plotting function ***plot.varpart()***, are of equal sizes despite differences in the corresponding R^2_a. Circle sizes and shapes can be modified using a graphics editor prior to the publication of the partitioning results.

by RDA, whereas they involved multiple regression in Subsection 10.3.5. The partitioning example reported here uses the species and environmental data collected by Verneaux (1973), which are available in the R package ADE4. The data were reanalysed for variation partitioning by Borcard *et al*. (2011, Section 6.3.2.7); here as in that book, site 8, where no fish were caught, was removed from the original 30-site species and environment data tables.

Partitioning involved three data sets: Topography (variables: altitude, slope, water flow), Chemistry (variables: pH, hardness, concentrations of phosphate, nitrate, ammonia, dissolved O_2, and biological oxygen demand), and Geography (variable: linear distance from the source along the course of the river). The partitioning results, obtained from function ***varpart()*** of the VEGAN package, are illustrated by a Venn diagram (Fig. 11.6); the decomposition into fractions [a] to [h] was done from the adjusted R^2 values (R^2_a) calculated by RDAs involving 1, 2, and all 3 explanatory data tables, as in Subsection 10.3.5. The first finding was that each of the three data sets explained approximately the same fraction of the spatial variation in the fish assemblage along the river (R^2_a = 0.345, 0.474, and 0.367, respectively). A great deal of the variation was shared among two or all three sets of explanatory variables. There was a small but significant portion of the fish variation explained by Topography that was not shared with the

two other data sets (fraction [a]: R^2_a = 0.036, p = 0.047), and likewise for Chemistry (fraction [b]: R^2_a = 0.076, p = 0.020). However, all the fish variation among sites explained by Geography was also explained by one of the other two explanatory data tables, or by both, leaving no significant portion of variation explained solely by Geography (R^2_a = 0.001, p = 0.411). Whereas the three explanatory data sets explained jointly 58.7% of the species variation, it was the Topography and Chemistry data that were the most informative and complementary, adding to the model portions of variation that were not explained by Geography alone. Results of a partitioning involving soil mite assemblages by four explanatory data sets are presented in Borcard *et al*. (2011, Section 7.4.2.5).

In Chapter 14, which describes multiscale spatial analysis, variation partitioning is used to partition the variation of data **Y** between two components, environmental (**X**) and spatial (**W**). Two ecological applications (14.1a and 14.1b) involving variation partitioning by partial canonical analysis are presented.

Ecological application 11.1b

In a classical study of spider community ecology, Aart & Smeenk-Enserink (1975) used canonical correlation analysis (CCorA, Section 11.4) to analyse a portion of the hunting spider data collected in pitfall traps at 100 sites in the Bierlap dune valley of the Netherlands. The paper related the species to environmental descriptors obtained at 28 of the 100 sites. The authors used canonical correlation analysis, a symmetric method of canonical analysis that was available in computer packages in the 1970s, to describe the influence of environmental conditions on the spider assemblages; their objective was to test the hypothesis of an asymmetric relationship between species and environmental conditions. The present example will show original results that we obtained by RDA, which is a more appropriate method to study and test asymmetric relationships. An additional advantage is that RDA can be carried out on unstandardized species data, thus preserving the original variances of the individual species in the analysis (Subsection 11.1.5), whereas the species data are always standardized in CCorA (Subsection 11.4.1). The Aart & Smeenk-Enserink spider data have been reanalysed, after recoding, by ter Braak (1986)* using CCA. The same data (recoded by ter Braak, 1986) were also analysed by De'ath (2002) using multivariate regression tree analysis (MRT, Ecological application 8.11).

At the 28 sites included in the canonical correlation analysis of Aart & Smeenk-Enserink (1975), the community composition data were the abundances of 12 hunting spider species normalized by logarithmic transformation, log(y + 1). Among the 27 environmental descriptors characterizing the light, vegetation, and soil that had been observed, only the 15 that were linearly correlated with the species variables were used by these authors for their canonical correlation analysis in order to ensure linearity of the relationships between the two sets of descriptors.†

* Warning to users: the 28 sites for which environmental data are provided in the Aart & Smeenk-Enserink (1975) paper are presented in a different order in Table 3 of ter Braak (1986).

† The spider (28 sites, 12 species) and environmental data (28 sites, 15 variables) used in this Application are available on the Web page http://numericalecology.com/data.

For the present application, the 15 environmental variables selected by Aart & Smeenk-Enserink (1975) were used as matrix **X** to insure comparability of the present results with theirs. The adjusted R^2 (R^2_a) of the RDA provided a criterion to select the best transformation for the species data: after computing RDA of the spider data, transformed in different ways, with the 15 environmental variables, R^2_a was higher for the log-transformed species data than for any of the other transformations of Section 7.7; so the log-transformed data were used in the RDA. Forward selection (with the stopping criterion $p \leq 0.05$) was carried out among the 15 environmental variables (Subsection 11.1.10, point 7). A parsimonious model containing six environmental variables was selected, which provided the same explanation as the full set of explanatory variables: $R^2_a = 0.761$ for the full set of 15 environmental variables, $R^2_a = 0.768$ for the subset of six variables, i.e. water content of the soil, illuminance under cloudless sky, ground cover by leaves and twigs, cover by the herb layer, cover by *Calamagrostis epigejos* (a grass, family Poaceae), and cover by the tree layer.

A search for species associations was carried out using concordance analysis, described in Subsection 8.9.2. The first statistically significant association comprised three species: *Alopecosa accentuata*, *Alopecosa fabrilis* and *Arctosa perita*; a fourth species, *Pardosa monticola*, was weakly associated with this group. The second significant association contained seven species: *Alopecosa cuneata*, *Arctosa lutetiana*, *Aulonia albimana*, *Pardosa nigriceps*, *Pardosa pullata*, *Trochosa terricola* and *Zora spinimana*. The species *Pardosa lugubris* formed a single-species group.

The RDA triplot (Fig. 11.7) shows the relationships between the species and the environmental variables. The species belonging to association 1 (upper ellipse) were found in greater abundances at very dry and more intensely lit sites. Those belonging to association 2 (right ellipse) were found at sites with higher soil humidity and higher cover by herbs and by *Calamagrostis epigejos*. The single-species association *Pardosa lugubris* exhibited preference for shaded sites with higher soil humidity and higher cover by trees and by leaves and twigs.

The total species variation, which is a measure of beta diversity (Section 6.5), was partitioned between the physical (soil, light) and vegetation influences using variation partitioning (Fig. 11.8). Fractions [a] and [c] were both statistically significant (tested by partial RDA), but fraction [c], which depicted the fraction of beta diversity explained exclusively by vegetation ($R^2_a = 0.43$), was much larger than fraction [a], which corresponded to the variation explained only by the physical environment ($R^2_a = 0.07$). Most of the explanation ([b] = 0.27) provided by the physical variables was shared with the vegetation variables.

11.2 Canonical correspondence analysis (CCA)

Canonical correspondence analysis is a canonical asymmetric ordination method developed by ter Braak (1986, 1987a, 1987c). First implemented in the program CANOCO (ter Braak, 1988b, 1988c, 1990; ter Braak & Smilauer, 1998), it is now available in several computer packages and R functions for community ecology. It is the canonical form of correspondence analysis. Any data table that could be subjected to correspondence analysis (CA, Section 9.2) is a suitable *response matrix* **Y** for CCA; this is the case, in particular, for species presence-absence or abundance data (Subsection 9.2.4).

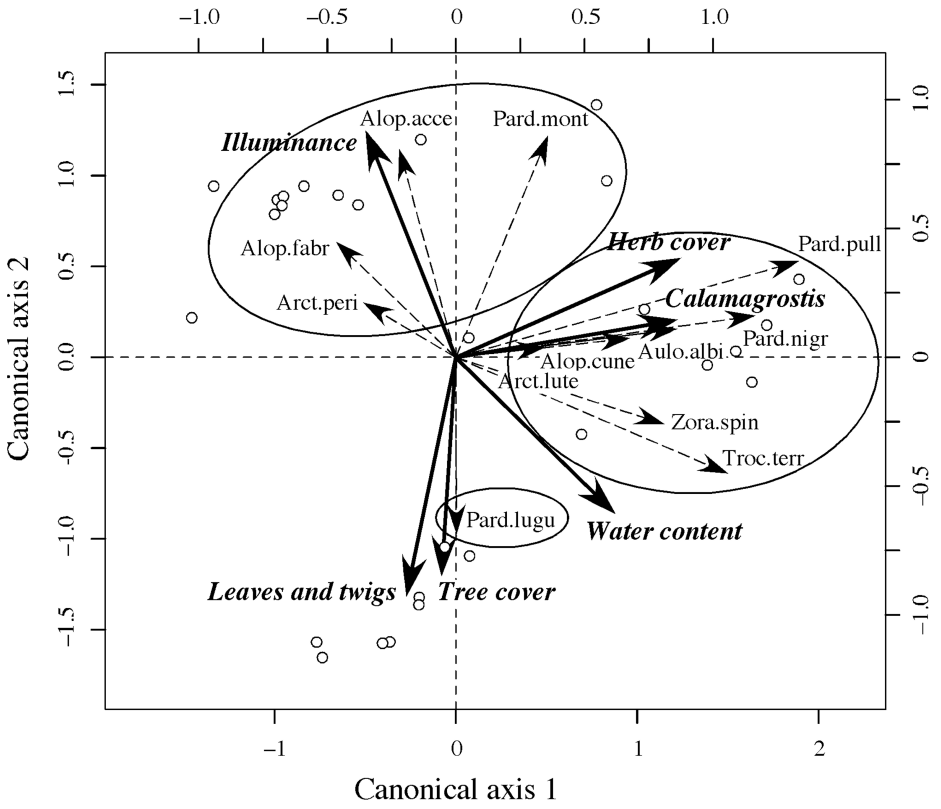


Figure 11.7 RDA triplot relating the spider species (dashed arrows) to the selected environmental variables (full-line arrows). Scaling type 2 was used to emphasize the covariances among the species. Small open circles represent the 28 sites; site names were not printed to keep the diagram simple. The species associations are indicated by ellipses. Association 1: *Alopecosa accentuata* (abbreviation: Alop.acce), *Alopecosa fabrilis* (Alop.fabr), *Arctosa perita* (Arct.peri) and *Pardosa monticola* (Pard.mont, weakly associated with this group). Association 2: *Alopecosa cuneata* (Alop.cune), *Arctosa lutetiana* (Arct.lute), *Aulonia albimana* (Aulo.albi), *Pardosa nigriceps* (Pard.nigr), *Pardosa pullata* (Pard.pull), *Trochosa terricola* (Troc.terr) and *Zora spinimana* (Zora.spin). Single-species group: *Pardosa lugubris* (Pard.lugu).

1 — The algebra of canonical correspondence analysis

The mathematics of CCA is derived from that of RDA. The first difference is that matrix $\overline{\mathbf{Q}}$ is used instead of $\mathbf{Y}$ as the response matrix in the calculations, as it was the case in correspondence analysis (Section 9.2). The second difference is that a diagonal matrix of row weights, $\mathbf{D}(p_{i+})$, is used in the regression portion of the calculation. For

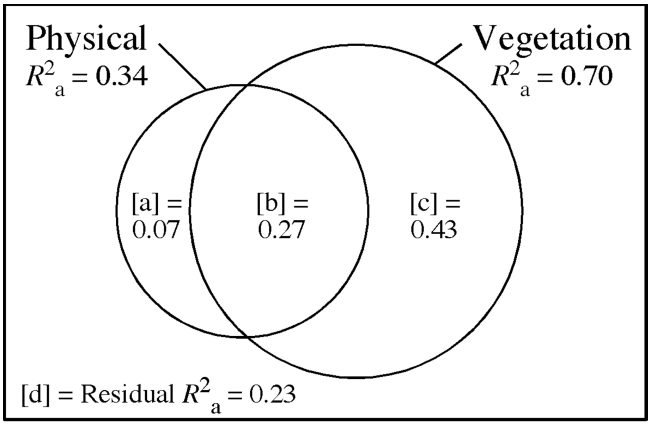


Figure 11.8 Venn diagram partitioning the total spider species variation (rectangle) between physical (water content of the soil, illuminance under cloudless sky) and vegetation influences (ground cover by leaves and twigs, cover by the herb layer, cover by *Calamagrostis epigejos*, and cover by the tree layer). The fraction identifiers [a], [b] and [c], are as in Fig. 10.10. The fractions are expressed as R^2_a, as in Fig. 11.6. Circle sizes are approximate.

each row of $\mathbf{Y}$, f_{i+} is the sum of the values in row i, and p_{i+} is f_{i+} divided by the grand total, f_{++}, of the frequencies in $\mathbf{Y}$.

Inflated data matrix

To obtain a CCA, the regression portion of the calculation is modified, in eq. 11.25 (below), in such a way as to emulate a RDA carried out on *inflated data matrices* $\mathbf{Y}_{infl}$ and $\mathbf{X}_{infl}$ constructed as follows. $\mathbf{Y}$ $(n \times p)$ contains frequency data, such as species presences or abundances of p species observed at n sites, and $\mathbf{X}$ $(n \times m)$ contains explanatory, e.g. environmental, variables. The presence of *a single individual* in $\mathbf{Y}$ produces a new row in $\mathbf{Y}_{infl}$, so that there are as many rows in $\mathbf{Y}_{infl}$ as there are individual organisms, or presences, in $\mathbf{Y}$. The number of rows of $\mathbf{Y}_{infl}$ is thus f_{++}. $\mathbf{Y}_{infl}$ still has p columns for the p species, but a single individual is present in each row. In $\mathbf{X}_{infl}$, the row vectors of explanatory data are duplicated as many times as needed to make every individual organism (i.e. every species presence) in $\mathbf{Y}_{infl}$ face, in $\mathbf{X}_{infl}$, a copy of the appropriate vector of explanatory data. Compute $\overline{\mathbf{Q}}_{infl}$ from $\mathbf{Y}_{infl}$ using eq. 9.24. CCA is the RDA of $\overline{\mathbf{Q}}_{infl}$ by $\mathbf{X}_{infl}$: the eigenvalues* and matrix of eigenvectors are the same.

* The eigenvalues of RDA of $\overline{\mathbf{Q}}_{infl}$ by $\mathbf{X}_{infl}$ computed on the covariance matrix of $\hat{\mathbf{Y}}$, instead of the cross-product matrix, are smaller than those of CCA by a multiplicative factor $(f_{++} - 1)$.

In computer programs, it is possible to use matrices $\overline{\mathbf{Q}}$ and $\mathbf{X}$ for the calculations instead of $\overline{\mathbf{Q}}_{infl}$ and $\mathbf{X}_{infl}$, which would be cumbersome to compute when $\mathbf{Y}$ has a large sum f_{++}. The modified algorithm is the following:

• The dependent data matrix is not $\mathbf{Y}$ centred by columns as in RDA. In this algorithm, CCA uses matrix $\overline{\mathbf{Q}}$ of the contributions to chi-square, also used in correspondence analysis, as the response matrix. $\overline{\mathbf{Q}}$ is derived from matrix $\mathbf{Y}$ through eq. 9.24.

• Matrix $\mathbf{X}$ is standardized to $\mathbf{X}_{stand}$ using weights $\mathbf{D}(f_{i+})$. To achieve this, the inflated matrix $\mathbf{X}_{infl}$ is constructed as described above; it contains f_{++} rows. Then the mean and standard deviation of each column of $\mathbf{X}_{infl}$ are computed and used to standardize the explanatory variables in $\mathbf{X}$. For the standard deviations (eq. 4.5), the maximum likelihood estimator of the variance is used instead of the usual unbiased estimator (eq. 4.3); in other words, the sum of squared deviations from the mean of the variables in $\mathbf{X}_{infl}$ is divided by the number of rows of that matrix (which is equal to f_{++}), instead of the number of rows minus 1.

• To obtain the regression coefficients, weighted multiple regression is used instead of conventional multiple regression. The row weights, written in diagonal matrix $\mathbf{D}(p_{i+})^{1/2}$ (Subsection 9.2.1), are applied to matrix $\mathbf{X}$ everywhere it occurs in the multivariate regression equation, which becomes:

$$\mathbf{B} = [\mathbf{X}_{stand}' \, \mathbf{D}(p_{i+}) \, \mathbf{X}_{stand}]^{-1} \, \mathbf{X}_{stand}' \, \mathbf{D}(p_{i+})^{1/2} \, \overline{\mathbf{Q}}$$

and

$$\hat{\mathbf{Y}} = \mathbf{D}(p_{i+})^{1/2} \, \mathbf{X}_{stand} \, \mathbf{B}$$

The equation for computing $\hat{\mathbf{Y}}$ is then:

$$\hat{\mathbf{Y}} = \mathbf{D}(p_{i+})^{1/2} \, \mathbf{X}_{stand} \, [\mathbf{X}_{stand}' \, \mathbf{D}(p_{i+}) \, \mathbf{X}_{stand}]^{-1} \, \mathbf{X}_{stand}' \, \mathbf{D}(p_{i+})^{1/2} \, \overline{\mathbf{Q}} \qquad \textbf{(11.25)}$$

The matrix of residuals is computed as $\overline{\mathbf{Q}}_{\mathbf{res}} = \overline{\mathbf{Q}} - \hat{\mathbf{Y}}$. This is the equivalent, for CCA, of equation $\mathbf{Y}_{\mathbf{res}} = \mathbf{Y} - \hat{\mathbf{Y}}$ found in Fig. 11.2 for RDA.

• Eigenvalue decomposition (eqs. 11.15 and 11.16) is carried out on matrix $\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}}$ which, in this case, is simply the matrix of sums of squares and cross products, without division by the number of degrees of freedom — as in correspondence analysis:

$$\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} \qquad \textbf{(11.26)}$$

One can show that $\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}}$ (eq. 11.26) is equal to $\mathbf{S}_{\overline{\mathbf{Q}}\mathbf{X}}\mathbf{S}_{\mathbf{XX}}^{-1}\mathbf{S}'_{\overline{\mathbf{Q}}\mathbf{X}}$ if the covariance matrices $\mathbf{S}_{\overline{\mathbf{Q}}\mathbf{X}}$ and $\mathbf{S}_{\mathbf{XX}}$ are computed as follows, with weights on $\mathbf{X}$ given by matrix $\mathbf{D}(p_{i+})^{1/2}$:

$$\mathbf{S}_{\overline{\mathbf{Q}}\mathbf{X}} = \overline{\mathbf{Q}}' \, \mathbf{D}(p_{i+})^{1/2} \, \mathbf{X} \quad \text{and} \quad \mathbf{S}_{\mathbf{XX}} = \mathbf{X}' \, \mathbf{D}(p_{i+}) \, \mathbf{X}$$

without division by degrees of freedom.

In the modified algorithm, CCA is the eigen-decomposition of $\mathbf{S}_{\hat{\mathbf{Y}}'\hat{\mathbf{Y}}}$ (eq. 11.26). It produces matrices $\mathbf{\Lambda}$ of eigenvalues and $\mathbf{U}$ of eigenvectors. Canonical correspondence analysis is thus a weighted form of redundancy analysis, applied to response matrix $\overline{\mathbf{Q}}$. The solution approximates the chi-square distances among the rows (objects) of the dependent data matrix, subject to the constraint that the canonical ordination vectors be maximally related to weighted linear combinations of the explanatory variables. The method is well suited to analyse the relationships between species presence/absence or abundance data matrices and tables of environmental variables. The number of canonical and non-canonical axes expected from the analysis are shown in Table 11.1. Tests of significance for the total canonical variation and for individual canonical axes are carried out in the same way in CCA as described for RDA in Subsections 11.1.2 and 11.1.8.

- The normalized matrix $\hat{\mathbf{U}}$ is obtained using eq. 9.30:

$$\hat{\mathbf{U}} = \overline{\mathbf{Q}}\,\mathbf{U}\mathbf{\Lambda}^{-1/2}$$

In CCA, matrix $\hat{\mathbf{U}}$ defined here does not contain the loadings of the rows of $\hat{\mathbf{Y}}$ on the canonical axes. It contains instead the loadings of the rows of $\overline{\mathbf{Q}}$ on the ordination axes, as in CA. It will be used to find the site scores (matrices $\mathbf{F}$ and $\hat{\mathbf{V}}$) in the space of the original variables $\mathbf{Y}$. The site scores in the space of the fitted values $\hat{\mathbf{Y}}$ will be found using $\mathbf{U}$ instead of $\hat{\mathbf{U}}$.

Scalings in CCA

- Matrix $\mathbf{V}$ of species scores (for scaling type 1) and matrix $\hat{\mathbf{V}}$ of site scores (for scaling type 2) are obtained from $\mathbf{U}$ and $\hat{\mathbf{U}}$ using the transformations described for correspondence analysis (Subsection 9.2.1):

eq. 9.33 (species scores, scaling 1): $\mathbf{V} = \mathbf{D}(p_{+j})^{-1/2}\mathbf{U}$

and eq. 9.34 (site scores, scaling 2): $\hat{\mathbf{V}} = \mathbf{D}(p_{i+})^{-1/2}\hat{\mathbf{U}}$

or combining eqs. 9.30 and 9.34: $\hat{\mathbf{V}} = \mathbf{D}(p_{i+})^{-1/2}\overline{\mathbf{Q}}\,\mathbf{U}\mathbf{\Lambda}^{-1/2}$

Scalings 1 and 2 are the same as in correspondence analysis (Subsection 9.2.1). Matrices $\mathbf{F}$ (site scores for scaling type 1) and $\hat{\mathbf{F}}$ (species scores for scaling type 2) are found using eqs. 9.35a and 9.36a:

$$\mathbf{F} = \hat{\mathbf{V}}\mathbf{\Lambda}^{1/2} \quad \text{and} \quad \hat{\mathbf{F}} = \mathbf{V}\mathbf{\Lambda}^{1/2}$$

Equations 9.35b and 9.36b cannot be used here to find $\mathbf{F}$ and $\hat{\mathbf{F}}$ because the eigenanalysis has been conducted on a covariance matrix (eq. 11.26) computed from the matrix of fitted values $\hat{\mathbf{Y}}$ (eq. 11.25) and not from $\mathbf{Q}$ defined in Subsection 9.2.1.

As mentioned in Subsection 9.2.1 about correspondence analysis, scaling type 3, which is called "symmetric scaling" in program CANOCO, is a compromise between

scalings 1 and 2. This scaling does not preserve the chi-square distances among the species or among the site scores. It is obtained by drawing together matrices $\hat{\mathbf{V}}\,\mathbf{\Lambda}^{1/4}$ (or $\mathbf{F}\mathbf{\Lambda}^{-1/4}$) for sites and $\mathbf{V}\mathbf{\Lambda}^{1/4}$ (or $\hat{\mathbf{F}}\,\mathbf{\Lambda}^{-1/4}$) for species.

The site scores that are linear combinations of the environmental variables, corresponding to eq. 11.18 of RDA, are found from $\hat{\mathbf{Y}}$ using the following equations:

For scaling type 1: $$\mathbf{Z}_1 = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{Y}}\,\mathbf{U} \qquad \textbf{(11.27)}$$

For scaling type 2: $$\mathbf{Z}_2 = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{Y}}\,\mathbf{U}\mathbf{\Lambda}^{-1/2} \qquad \textbf{(11.28)}$$

For scaling type 3: $$\mathbf{Z}_3 = \mathbf{D}(p_{i+})^{-1/2}\,\hat{\mathbf{Y}}\,\mathbf{U}\mathbf{\Lambda}^{-1/4} \qquad \textbf{(11.29)}$$

Before computing the biplot scores, matrix $\mathbf{Z}_1$ (or $\mathbf{Z}_2$ or $\mathbf{Z}_3$: identical results) must be standardized to $\mathbf{Z}_{stand}$ using the procedure described for the standardization of $\mathbf{X}$: generate the inflated matrix $\mathbf{Z}_{1\,infl}$, compute the vectors of column means and standard deviations (in the computation of the variances, divide the sums of squares by f_{++} instead of $(f_{++} - 1)$) for $\mathbf{Z}_{1\,infl}$, and use these vectors to standardize $\mathbf{Z}_1$. Applying this concept, computational shortcuts can be used to obtain matrix $\mathbf{Z}_{stand}$ without actually generating matrix $\mathbf{Z}_{1\,infl}$. The matrices of biplot scores (**BS**) for the explanatory variables can now be computed using $\mathbf{X}_{stand}$, $\mathbf{Z}_{stand}$, and the diagonal matrix of row weights $\mathbf{D}(p_{i+})$:

For scaling type 1: $$\mathbf{BS}_1 = \mathbf{X}_{stand}'\,\mathbf{D}(p_{i+})\,\mathbf{Z}_{stand}\,\mathbf{\Lambda}^{1/2} \qquad \textbf{(11.30)}$$

For scaling type 2: $$\mathbf{BS}_2 = \mathbf{X}_{stand}'\,\mathbf{D}(p_{i+})\,\mathbf{Z}_{stand} \qquad \textbf{(11.31)}$$

For scaling type 3: $$\mathbf{BS}_3 = \mathbf{X}_{stand}'\,\mathbf{D}(p_{i+})\,\mathbf{Z}_{stand}\,\mathbf{\Lambda}^{1/4} \qquad \textbf{(11.32)}$$

For scaling type 1, triplots are drawn using matrix $\mathbf{V}$ for the species, either $\mathbf{Z}_1$ or $\mathbf{F}$ for the sites, and $\mathbf{BS}_1$ for the explanatory variables. For scaling type 2, matrix $\hat{\mathbf{F}}$ is used for the species, either $\mathbf{Z}_2$ or $\hat{\mathbf{V}}$ for the sites, and $\mathbf{BS}_2$ for the explanatory variables. For scaling type 3, matrix $\mathbf{V}\mathbf{\Lambda}^{1/4}$ is used for the species, either $\mathbf{Z}_3$ or $\hat{\mathbf{V}}\,\mathbf{\Lambda}^{1/4}$ for the sites, and $\mathbf{BS}_3$ for the explanatory variables. The construction and interpretation of CCA triplots is discussed in more detail in ter Braak & Verdonschot (1995).

• Residuals can be analysed by applying eigenvalue decomposition (eq. 11.15) to matrix $\bar{\mathbf{Q}}_{\mathbf{res}}$, producing a matrix of eigenvalues $\mathbf{\Lambda}$ and a matrix of eigenvectors $\mathbf{U}$. Matrix $\hat{\mathbf{U}}$ is obtained using eq. 9.30: $\hat{\mathbf{U}} = \bar{\mathbf{Q}}\,\mathbf{U}\mathbf{\Lambda}^{-1/2}$. Species and site scores are obtained for scaling types 1 and 2 (eqs. 9.33, 9.34, 9.35a and 9.36a) using the matrices of row and column sums $\mathbf{D}(p_{i+})^{-1/2}$ and $\mathbf{D}(p_{+j})^{-1/2}$ of the original matrix $\mathbf{Y}$.

CCA can be computed following the algorithm described in the present subsection*. One may also use the iterative algorithm proposed by ter Braak (1986, 1987a) and implemented in the program CANOCO. The latter algorithm, which has historical significance, is described in Table 11.6 of Legendre & Legendre (1998).

Partial CCA Developed by ter Braak (1988a), partial CCA is computed essentially like partial RDA (Subsection 11.1.6), after residualizing $\mathbf{Q}$ and $\mathbf{X}$ on the covariables $\mathbf{W}$. The weights $\mathbf{D}(p_{i+})^{1/2}$ are used in the computation of these residuals.

CCA can be used for variation partitioning (Subsections 10.3.5 and 11.1.11). The difficulty with CCA resides in the calculation of the adjusted R^2, which is necessary to obtain unbiased estimates of the fractions of explained variation. A method to compute the adjusted R^2 in CCA, involving a permutation procedure, was described by Peres-Neto *et al*. (2006). In Supplements to their paper, Peres-Neto *et al*. (2006) provided a MATLAB package and an executable program to conduct variation partitioning in CCA. At the time this paragraph is written, however, that method has not been incorporated into any major package for community ecology, with the consequence that variation partitioning is not yet generally available for CCA.

2 — Numerical example

Table 11.3 will now be used to illustrate the computation and interpretation of CCA. The 9 species were used in matrix **Y**. Matrix **X** comprised the four columns shown in the right-hand portion of Table 11.3. CCA results are presented in Table 11.7 and Fig. 11.9; the CANOCO program and the CCA functions in R* provide more output tables than presented here. There was a possibility of 3 canonical and 8 non-canonical axes. It turned out that the last 2 non-canonical axes had zero variance; they are consequently not displayed. An overall test of significance showed that the canonical relationship between matrices **X** and **Y** was very highly significant ($p = 0.001$ after 999 permutations of residuals under a full model; Subsection 11.1.8). The canonical axes explained 47%, 24% and 10% of the response table's inertia, respectively. They were all significant ($p < 0.05$) and displayed strong row-weighted species-environment correlations ($r = 0.998, 0.940$, and 0.883, respectively).

Scaling type 2 (Subsection 11.2.1) was used, in this example, to emphasize the relationships among species. As a result, the species (matrix $\hat{\mathbf{F}}$) are at the centroids of the sites (matrix $\hat{\mathbf{V}}$) in Fig. 11.9a, and distances among species approximate their chi-square distances. Species 3 and 4 characterize the sites with coral substrate, whereas species 5 and 6 indicate the sites with "other substrate". Species 1 and 2, which occupy an intermediate position between the sites with coral and other substrate, are not well represented in the biplot of canonical axes I and II; axis III is needed to adequately represent the variance of these species. Among the ubiquitous species 7 to 9, two are well represented in the subspace of canonical axes I and II; they fall near the middle of the area encompassing the three types of substrate. The sites are not perfectly ordered along the depth vector; the site ordering along this variable mainly reflects differences in species composition between the shallow sandy sites (1, 2 and 3) and the other sites.

* Function CCA.R was written to demonstrate the CCA algorithm described in this subsection. It produces results identical to those of CANOCO 4.x. The function is available on the Web page http://numericalecology.com/rcode.

Table 11.7 Results of canonical correspondence analysis of the data in Table 11.3 (selected output). Matrix **Y**: species 1 to 9; **X**: depth and 3 substrate classes. Non-canonical axes VIII and IX not shown.

	Canonical axes			Non-canonical axes			
	I	II	III	IV	V	VI	VII
Eigenvalues (their sum is equal to the total inertia in matrix $\overline{\mathbf{Q}}$ of species data = 0.78417)							
	0.36614	0.18689	0.07885	0.08229	0.03513	0.02333	0.00990
Fraction of the total variance in $\overline{\mathbf{Q}}$							
	0.46691	0.23833	0.10055	0.10494	0.04481	0.02975	0.01263
Cumulative fraction of total inertia in $\overline{\mathbf{Q}}$ accounted for by axes 1 to k							
	0.46691	0.70524	0.80579	0.91072	0.95553	0.98527	0.99791
Eigenvectors ("species scores", scaling 2): matrices $\hat{\mathbf{F}}$ for the canonical and non-canonical portions (eq. 9.36a)							
Species 1	–0.11035	–0.28240	–0.20303	0.00192	0.08223	0.08573	–0.01220
Species 2	–0.14136	–0.30350	0.39544	0.14127	0.02689	0.14325	0.04303
Species 3	1.01552	–0.09583	–0.19826	0.10480	–0.13003	0.02441	0.04647
Species 4	1.03621	–0.10962	0.22098	–0.22364	0.24375	–0.02591	–0.05341
Species 5	–1.05372	–0.53718	–0.43808	–0.22348	0.32395	0.12464	–0.11928
Species 6	–0.99856	–0.57396	0.67992	0.38996	–0.29908	0.32845	0.21216
Species 7	–0.25525	0.17817	–0.20413	–0.43340	–0.07071	–0.18817	0.12691
Species 8	–0.14656	0.85736	–0.01525	–0.05276	–0.35448	–0.04168	–0.19901
Species 9	–0.41371	0.70795	0.21570	0.69031	0.14843	–0.33425	–0.00629
Site scores ("sample scores", scaling 2): matrices $\hat{\mathbf{V}}$ for the canonical and the non-canonical portions (eq. 9.34)							
Site 1	–0.71059	3.08167	0.21965	1.24529	1.07293	–0.50625	0.24413
Site 2	–0.58477	3.00669	–0.94745	–2.69965	–2.13682	0.81353	0.47153
Site 3	–0.76274	3.15258	2.13925	3.11628	2.30660	–0.69894	–1.39063
Site 4	–1.11231	–1.07151	–1.87528	–0.66637	1.10154	1.43517	–1.10620
Site 5	0.97912	0.06032	–0.69628	0.61265	–0.98301	0.31567	0.57411
Site 6	–1.04323	–0.45943	–0.63980	–0.28716	0.57393	–1.44981	1.70167
Site 7	0.95449	0.08470	0.13251	0.42143	0.11155	–0.39424	–0.67396
Site 8	–0.94727	0.10837	0.52611	0.00565	–1.26273	–1.06565	–1.46326
Site 9	1.14808	–0.49045	0.47835	–1.17016	1.00599	0.07350	0.08605
Site 10	–1.03291	–1.03505	2.74692	1.28084	–0.36299	1.98648	1.05356
Correlations of environmental variables with site scores							
Depth	0.18608	–0.60189	0.65814				
Coral	0.99233	–0.09189	–0.04614				
Sand	–0.21281	0.91759	0.03765				
Other subs.	–0.87958	–0.44413	0.02466				
Correlations of environmental variables with fitted site scores (for biplot, scaling 2)							
Depth	0.18636	–0.64026	0.74521				
Coral	0.99384	–0.09775	–0.05225				
Sand	–0.21313	0.97609	0.04263				
Other subs.	–0.88092	–0.47245	0.02792				
Centroids of sites with code "1" for the BINARY environmental variables, scaling 2							
Coral	1.02265	–0.10059	–0.05376				
Sand	–0.66932	3.06532	0.13387				
Other subs.	–1.03049	–0.55267	0.03266				

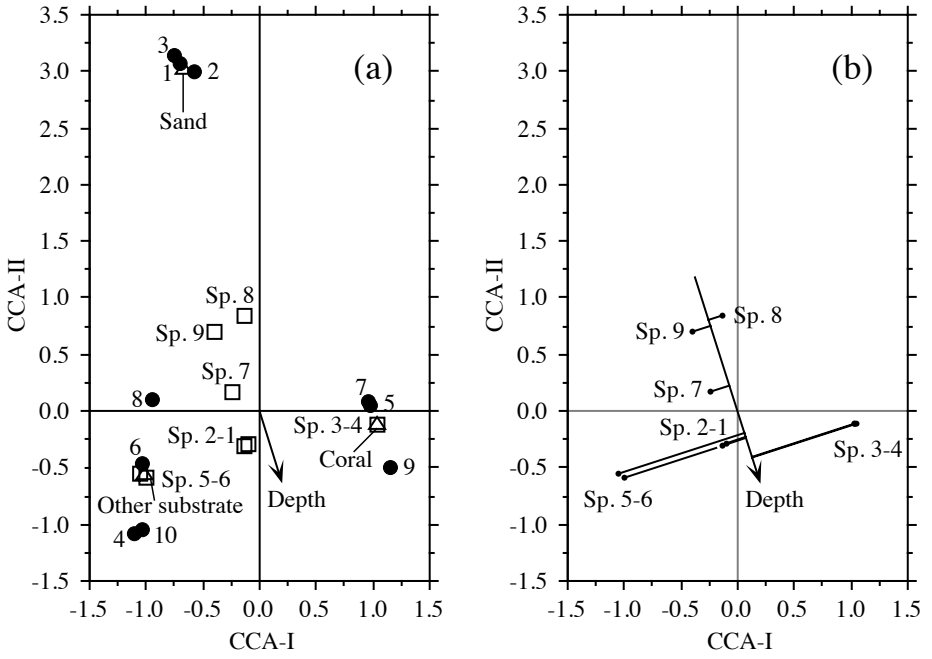


Figure 11.9 CCA ordination triplot (scaling type 2) of the artificial data in Table 11.3; the numerical results of the analysis are in Table 11.7. (a) Triplot representing the species (squares), sites (dots, with site identifiers that also correspond to water depths in m), and environmental variables (full arrow for depth, triangles for the three binary substrate variables). (b) Ranking of the species along the quantitative environmental variable (depth) is inferred by projecting the species at right angle onto the arrow representing that variable.

Figure 11.9b shows how to infer the ranking of species along a quantitative environmental variable. Depth is used in this example. The graphical method simply consists in projecting (at right angle) the species onto the arrow representing that variable. This gives an approximation of the weighted averages of the species with respect to environmental variables. Ecologists like to interpret this ranking as representing the niche optima for the species under consideration. It is important to realize that three rather strong assumptions are made when attempting such an interpretation:

- that the various species have unimodal distributions along the environmental variable of interest (Subsection 9.2.4);

- that the species distributions are under environmental control (Whittaker, 1956; Bray & Curtis, 1957), so that the mode of each species is at its optimum along each environmental variable; and

- that the environmental gradient under study is long enough to allow each species to go from some less-than-optimum low frequency to its high-frequency optimum, and back to some past-optimum low frequency.

In the data of the present example (Table 11.3), only species 1, 3 and 5 were constructed to approximately correspond to these criteria. Species 7, which may also look like it has a unimodal distribution, has actually been constructed using a pseudo-random number generator; so its optimum along depth is fortuitous.

To investigate the similarities among sites or the relationships among species after controlling for the linear effects of depth and type of substrate, one could draw ordination biplots of the *non-canonical axes* in Table 11.7. These axes correspond to a correspondence analysis of the table of regression residuals, as shown in Fig. 11.2.

Ecological application 11.2a

Ecological application 9.2b described the spatial distribution of chaetodontid fish assemblages (butterflyfishes) around a tropical island, using correspondence analysis. This application is continued here. Cadoret *et al*. (1995) next described the relationships between the fish species (quantitative relevés) and some environmental variables, using canonical correspondence analysis. The environmental variables were: the type of environment (qualitative descriptor: bay, lagoon, or outer slope of the reef on the ocean side), geomorphology (qualitative: reef flat, crest, and reef wall of the fringing reefs of bays; fringing reef, shallow, barrier reef, and outer slope for transect sites), depth (quantitative: from 0.5 to 35 m), and exposure to swell (qualitative: low, high, or sites located in bays).

The ordination of sampling sites by CCA is virtually identical to that in Fig. 9.14; this indicates that the first two CA axes are closely related to the environmental variables. The canonical axes account together for 35% of the variation in the species data ($p = 0.001$ after 999 permutations). The description of the ordination of sites presented in Ecological application 9.2b may be compared to Fig. 11.10. This figure shows which types of environment are similar in their chaetodontid species composition and which species are associated with the various types of environment. It indicates that the reef flats of the fringing reefs in bays are similar in species composition to the fringing reefs in the lagoon; likewise, the crests of the fringing reefs in bays are similar to the barrier reefs in the lagoon. The species composition along the reef walls in bays and that on the outer slopes differ, however, from all the other types of environment. The authors discuss the ecology of the most important chaetodontid species in their paper.

Ecological application 11.2b

Canonical correspondence analysis is widely used in palaeoecology, together with regression and calibration, to infer past ecological conditions (climatic, limnological, etc.) from palaeo-assemblages of species. The first 10 years of that literature (1986-1996) was summarized in a bibliography assembled by Birks *et al*. (1998), under the headings *limnology*, *palaeoecology*, *palaeolimnology*, etc. Several applications of CCA are described in a chapter by Legendre & Birks (2012) in a book about numerical methods in palaeoecology edited by Birks *et al*. (2012).

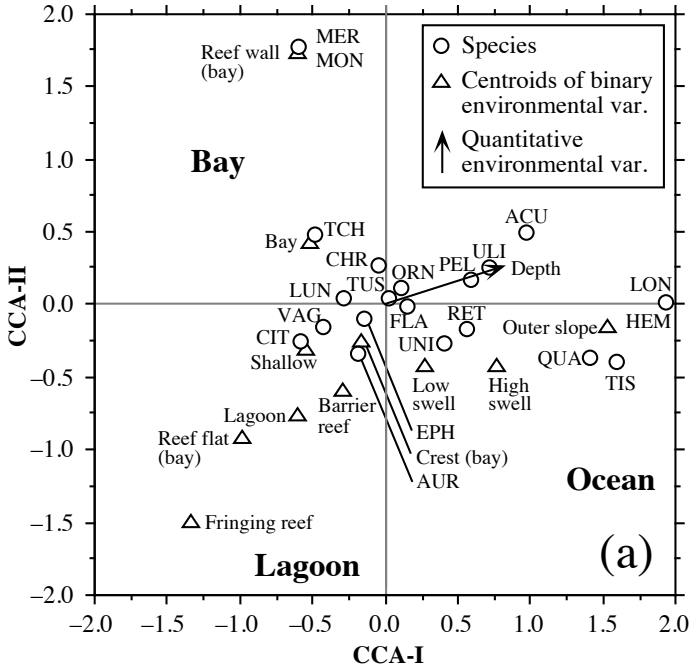


Figure 11.10 (a) CCA ordination diagram: presence/absence of 21 Chaetodontid fish species at 42 sampling sites around Moorea Island, French Polynesia, related to environmental variables. The species (names abbreviated to 3 letters) are represented by circles for readability of the diagram. Axis I: 14.6% of the variation ($p = 0.001$ after 999 permutations); axis II: 7.4% ($p = 0.010$). Redrawn from the original data of Cadoret *et al*. (1995).

One of the classical papers on the subject was written by Birks *et al*. (1990a). Palaeolimnological reconstruction involves two main steps: modelling from a *training data set*, followed by the construction of forecasting models that are then applied to the palaeo-data. In this paper, diatoms were used to reconstruct past water chemistry. The training data set consisted of diatom assemblages comprising 287 species, from present-day surface samples from 138 lakes in England, Norway, Scotland, Sweden, and Wales. Data were also available on pH, conductivity, Ca, Mg, K, SO_4, Cl, alkalinity, total Al, and DOC. Data from more lakes were available for subsets of these variables. CCA was used to relate species composition to water chemistry. The first two canonical eigenvalues were significant and displayed strong species-environment correlations ($r = 0.95$ and 0.84, respectively). The first axis expressed a significant diatom gradient which was strongly and positively correlated with alkalinity and its close correlates, Ca and pH, and negatively but less strongly correlated with total Al; the second axis corresponded to a significant gradient strongly correlated with DOC. This result indicated that pH (or alkalinity), Al, and DOC were potentially reconstructible from the fossil diatom assemblages.

The fossil data set contained 101 slices of a sediment core from a small lake, the Round Loch of Glenhead, in Galloway, southwestern Scotland. The data series covered the past 10000 years. The fossil data (292 diatom taxa) were included in the CCA as passive objects (called *supplementary objects* in Subsection 9.1.9) and positioned in the ordination provided by canonical axes I and II. All fossil objects were well-fitted in that space (they had low squared residual distances), indicating that the pattern of variation in diatom composition can be linked to the modern chemical variables.

Reconstruction of past surface-water chemistry involved two steps. First, the training set was used to model, by regression, the responses of modern diatoms to the chemical variables of interest (one variable at a time). Secondly, the modelled responses were used to infer past chemistry from the composition of fossil diatom assemblages; this phase is called *calibration* (ter Braak, 1987b; ter Braak & Prentice, 1988). Extensive simulations led Birks *et al*. (1990b) to prefer weighted averaging (WA) over maximum likelihood (ML) regression and calibration. Consider pH in lakes, for example. *WA regression* simply consists in applying eq. 9.39 to estimate the pH optimum of each taxon of the training set as the weighted average of all the pH values for lakes in which this taxon occurs, weighted by the taxon's relative abundance. *WA calibration* consists in applying eq. 9.38 to estimate the pH of each lake as the weighted average of the pH optima of all the taxa present. Taxa with a narrow pH tolerance or amplitude may, if required, be given greater weight in WA regression and calibration than taxa with a wide pH tolerance (Birks *et al*., 1990b).

Application of eqs. 9.39 and 9.38 to the data resulted in shrinkage of the range of pH scores. Shrinkage occurred for the same reason as in the TWWA algorithm for correspondence analysis; in step 6.4 of that algorithm (Table 9.8), the eigenvalue was actually estimated from the amount of shrinkage incurred by the site scores after each iteration through eqs. 9.39 and 9.38 (steps 3 and 4). Deshrinking may be done in at least two ways; the relative merits of the two methods are discussed by Birks *et al*. (1990b).

• Deshrinking by classical regression proceeds in two steps. (1) The pH values inferred by WA regression and calibration ($\hat{x}_i$) are regressed on the observed values x_i for the training set, using the linear regression model $\hat{x}_i = b_0 + b_1 x_i + \varepsilon_i$. (2) The parameters of that model are then used to deshrink the $\hat{x}_i$ values, using the equation: final $\hat{x}_i = (\hat{x}_i - b_0)/b_1$. This method was used to deshrink the inferred pH values.

• Another way of deshrinking, advocated by ter Braak & van Dam (1989) for palaeolimnological data, is to use "inverse regression" of x_i on $\hat{x}_i$ (ter Braak, 1987b). Inverse regression was used to deshrink the inferred Al and DOC values.

Training sets containing different numbers of lakes were used to infer pH, total Al, and DOC. Past values of these variables were then reconstructed from the palaeo-assemblages of diatoms, using the pH optima estimated above (eq. 9.39) for the various diatom species, followed by deshrinking. Reconstructed values were plotted against depth and time, together with error estimates obtained by bootstrapping. The past history of the Round Loch of Glenhead over the past 10000 years is discussed in the paper.

This approach involving CCA, WA regression, and WA calibration, is now widely used in palaeolimnology to reconstruct, for example, surface-water temperatures from fossil chironomid assemblages, as well as lake salinity, lake water phosphorus concentrations, or surface water chlorophyll *a* concentrations from fossil diatom assemblages. The WA regression and WA calibration method was further improved by ter Braak & Juggins (1993). ter Braak (1995) made a theoretical comparison of reconstruction methods. For a recent presentation, see the chapter by Birks (2010) in a book edited by Smol & Stoermer (2010). How to carry out the calculations was

described by ter Braak & Juggins (1993) and Line *et al*. (1994). R functions for palaeoenvironmental reconstruction are available in package RIOJA (Juggins, 2009).

A little-known application of CCA is worth mentioning here. Consider a qualitative environmental variable and a table of species presence-absence or abundance data. How can one "quantify" the qualitative states, i.e. give them values along a quantitative scale that would be related in some optimal way to the species data? CCA provides an easy answer to this problem. The species data form matrix **Y**; the qualitative variable, which may be coded as a factor or recoded as a set of dummy variables, is placed in matrix **X**. Compute CCA and take the fitted site scores (or "site scores that are linear combinations of environmental variables"): they provide a quantitative rescaling of the qualitative variable, maximizing the weighted linear correlation between the dummy variables and matrix $\mathbf{\bar{Q}}$. In the same way, RDA may be used to rescale a qualitative variable (factor) with respect to a table of quantitative variables of the objects if linear relationships can be assumed.

McCune (1997) warns users of CCA against inclusion of noisy or irrelevant explanatory variables in the analysis: they may lead to misleading interpretations.

11.3 Linear discriminant analysis (LDA)

A situation that often occurs is to start with an already known grouping of the objects, considered to form a qualitative response variable **y** in this type of analysis, and try to determine to what extent a set of quantitative descriptors, which are the explanatory variables **X**, can actually explain this grouping. In this type of analysis, the grouping is known at the start of the analysis. It may be the result of a cluster analysis computed from a *different* data set, or reflect an ecological hypothesis to be tested. The problem no longer consists in delineating groups, as in cluster analysis, but in interpreting them.

Linear discriminant analysis is a method of linear modelling, like the analysis of variance, multiple linear regression, redundancy analysis, and canonical correlation analysis. It proceeds in two steps. (1) First, one tests for differences in the predictor variables (**X**) among the predefined groups using Wilks' lambda (eq. 11.42). This part of the analysis is identical to the overall test performed in MANOVA. (2) If the test supports the alternative hypothesis of significant differences among groups in the **X** variables, the analysis proceeds to find the linear combinations (called *discriminant functions* or *identification functions*) of the **X** variables that best discriminate among the groups.

Like one-way analysis of variance, discriminant analysis considers a single classification criterion (i.e. division of the objects into groups) and allows one to test whether the explanatory variables can discriminate among the groups. Testing for differences among group means, in discriminant analysis, is identical to ANOVA for a single explanatory variable and to MANOVA for multiple variables (**X**).