# 13 Correspondence analysis and canonical correspondence analysis

In Chapter 12, we discussed PCA and RDA. Both techniques are based on the correlation or covariance coefficient. In this chapter, we introduce correspondence analysis (CA) and canonical correspondence analysis (CCA). We start by giving a historical insight into the techniques community ecologists have used most during the last two decades. This chapter is mainly based on Greenacre (1984), Ter Braak (1985, 1986), Ter Braak and Verdonschot (1995), Legendre and Legendre (1998) and Lepš and Šmilauer (2003).

## 13.1 Gaussian regression and extensions

Little is known about the relationships between abundances of marine ecological species and environmental variables. However, a feature that many species share is their change in abundances related to changes in environmental variables. Figure 13.1-A shows an artificial example of abundances of a particular species along the environmental variable temperature.

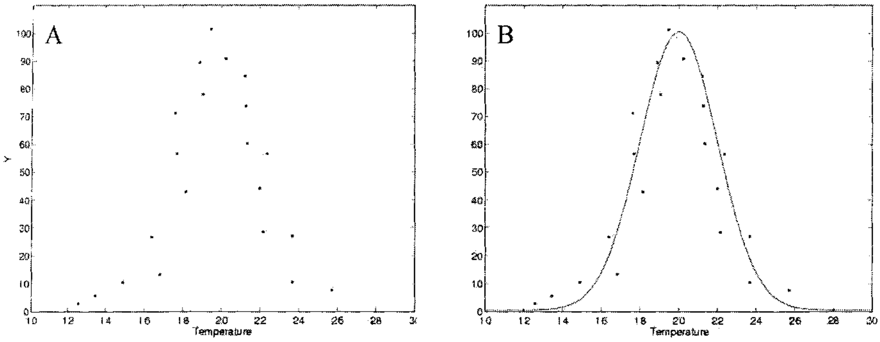

Figure 13.1. A: Observed abundance of a particular species along the environmental variable temperature. B: Fitted Gaussian response curve of a species along environmental variable $X$, with optimum $u = 20$, maximum value $c = 100$ and tolerance $t = 2$.

To model this behaviour, Whitaker (1978), Gauch (1982) and others used the so-called Gaussian response model. This is the simplest model to describe unimodal behaviour. For a particular species this Gaussian response model takes the form:

$$Y_i = ce^{-\frac{(X_i-u)^2}{2t^2}} \tag{13.1}$$

Where $i = 1,..,N$, $Y_i$ is the abundance of the species at site $i$, $N$ is the number of sites, $c$ is the maximum abundance of the species at the optimum $u$, and $t$ is its tolerance (measure of spread). Finally, $X_i$ is the value of environmental variable $X$ at site $i$. In Figure 13.1-B the Gaussian response curve is plotted. Note that equation (13.1) is a *response* function and not a probability density function.

The Gaussian response model is a very simple model, and real life processes in ecology are much more complex. Alternative models are available when many sites (e.g., 100 or more) are monitored. For example, Austin et al. (1994) used so-called $\beta$-functions, which allow for a wide range of asymmetric shaped curves. Because the (original) underlying model for CA and CCA is the Gaussian response model, we restrict ourselves to this model.

Several methods exist to estimate the three parameters $c$, $t$ and $u$ of the Gaussian response model. The easiest option is to use generalised linear modelling (GLM). In order to do so, we need to rewrite equation (13.1) as

$$Y_i = \exp(\ln c - \frac{u^2}{2t^2} + \frac{u}{t^2}x_i - \frac{1}{2t^2}x_i^2) = \exp(b_1 + b_2x_i + b_3x_i^2) \tag{13.2}$$

where $t = 1/\sqrt{-2b_3}$ , $u = -b_2/2b_3$ and $c = \exp(b_1 - b_2^2/4b_3)$. If $Y_i$ represents count data, it is common to assume that the $Y_i$ are independent Poisson distributed. Now GLM (Chapter 6) can be applied to the right most part of equation (13.2). This gives estimates of the parameters $b_1$, $b_2$ and $b_3$. From these estimates, the parameters $c$, $t$ and $u$ can be derived (Ter Braak and Prentice 1988).

### Multiple Gaussian regression

Assume that two environmental variables, say temperature and salinity, are measured at each of the $N$ sites. The Gaussian response model can now be written as

$$Y_i = \exp(b_1 + b_2x_{i1} + b_3x_{i1}^2 + b_4x_{i2} + b_5x_{i2}^2) \tag{13.3}$$

where $x_{i1}$ denotes temperature at site $i$ and $x_{i2}$ the salinity. This model contains five parameters. It is assumed that $x_1$ and $x_2$ do not interact. The bivariate Gaussian response curve is plotted in Figure 13.2. These can be constructed for every species of interest.

If $M$ species and $Q$ environmental variables are observed, and interactions are ignored, $(1 + 2Q)M$ parameters have to be estimated. If for example 10 species and 5 environmental variables are used, you need to estimate 110 parameters.

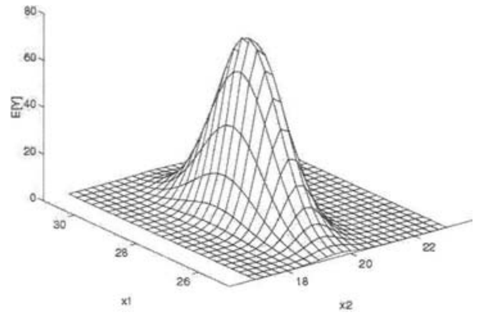

Figure 13.2. Bivariate Gaussian response curve of one species. The $x_1$ and $x_2$ axes are explanatory variables and the $y$-axis shows the expected counts.

### Restricted Gaussian regression

Let $x_{ip}$ be the value of environmental variable $p$ at site $i$, where $p = 1,..,Q$. Instead of using all environmental variables as covariates, we now use a linear combination of them as a single covariate in the Gaussian response model. The model becomes

$$\begin{aligned} Y_i &= ce^{-\frac{(z_i-u)^2}{2t^2}} = \exp(b_1 + b_2 z_i + b_3 z_i^2) \\ z_i &= \sum_{p=1}^{Q} \alpha_p x_{ip} \end{aligned} \qquad (13.4)$$

The model is called restricted Gaussian regression (RGR), and it tries to detect the major environmental gradients underlying the data. The parameters $\alpha_p$, denoted as *canonical coefficients*, are unknown. To interpret the gradient $z_i$, the coefficients $\alpha_p$ can be compared with each other. For this reason, environmental variables are standardised prior to the analysis. Just as in PCA, more gradients (or axes) can be used. This gradient is orthogonal with previous axes. Up to $Q$ gradients can be extracted. The formulae of the RGR model with two or more axes are given in Zuur (1999).

### Geometric interpretation of RGR

The geometric interpretation of restricted Gaussian regression is as follows. In Figure 13.3-A, abundances of three species are plotted against two covariates $x_1$ and $x_2$. A large point indicates a high abundance and a small point low abundance. We now seek a line $z$ that gives the best fit to these points. A potential candidate for this line is drawn in Figure 13.3-A. If abundances of each species are projected perpendicular on $z$ we obtain Figure 13.3-B. Now, the Gaussian response model in equation (13.4) can be fitted for all species, resulting in an overall measure of fit (typically the maximum likelihood). So, now we only need to find the combination of $\alpha_p$'s that gives the best overall measure of fit. Formulated differently, we need to find a gradient $z$ along which projected species abundances are fitted as well as possible by the Gaussian response model in equation (13.4). The mathematical procedure for this is given in Zuur (1999). The number of parameters to be estimated for the model in equation (13.4) is $3M + Q - 1$, where $M$ is the number of species and $Q$ is the number of environmental variables. If for example 10 species and 5 environmental variables are used, one has to estimate 34 parameters for the first gradient. This is considerably less than in Gaussian regression. In general, if $s$ gradients are used, then the Gaussian response model has $M(2s + 1) + Qs - \sum_{j=1}^{s} j$ parameters.

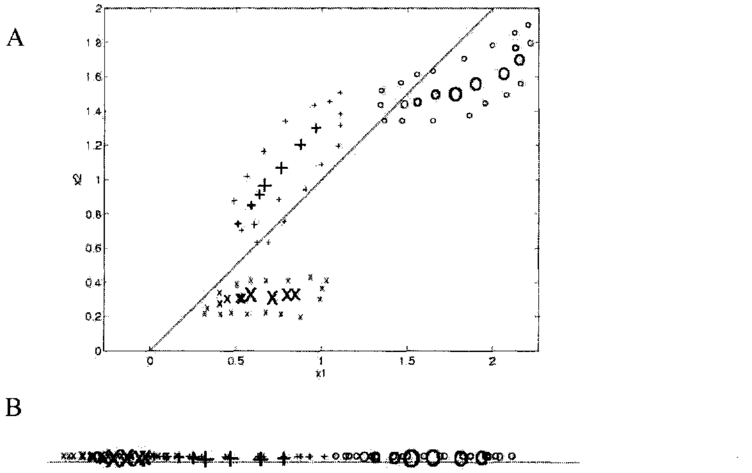

Figure 13.3. A: abundances of three species plotted in the $(x_1,x_2)$ space. The species are denoted by o, x and +, respectively. A thick point indicates a high abundance. The straight line is the gradient $z$. B: Abundances projected on $z$.

### Gaussian ordination

In Gaussian ordination, we do not use measured environmental variables. Instead, we try to estimate a hypothetical gradient. Other names for this hypothetical gradient are latent variable, synthetic variable, or factor variable. This hypothetical gradient is estimated in such a way, that if abundances of species are projected on the gradient, then this gives the best possible fit (measured by the maximum likelihood) by the Gaussian response model. The Gaussian response model now takes the form

$$Y_{ik} = c_k e^{-\frac{(l_i-u_k)^2}{2t_k^2}} \qquad (13.5)$$

where $l_i$ is the value of the latent variable at site $i$, $i = 1,..,N$, and the index $k$ refers to species. Hence in Gaussian ordination we estimate $c_k$, $u_k$, $t_k$ and $l_i$ from the observed abundances $Y_{ik}$. So we have to estimate $N + 3M$ parameters for the first gradient. If 30 sites and 10 species are used, the Gaussian response model contains 60 parameters. Numerical problems arise if more than one latent variable is used (Kooijman 1977).

### Heuristic Solutions

In Ter Braak (1986), the following four assumptions are made:

1. Tolerances of all species along an environmental variable are equal: $t_k = t$ for all k.
2. Maximum values of all species along an environmental variable are equal: $c_k = c$ for all k.
3. The optimum values $u_k$ are equally spaced along the environmental variable, which is long compared with the species tolerance $t_k$.
4. The sites (samples) cover the whole range of occurrence of species along the environmental variable and are equally spaced.

Using these assumptions, restricted Gaussian regression reduces to a method that is computationally fast and produces easily interpretable information on the parameters $u_k$ of the Gaussian response model. This method is called canonical correspondence analysis (CCA). As a result of these assumptions the Gaussian response curves of the species along an environmental variable simplify considerably; see Figure 13.4. This is the so-called *species packing* model. Due to these assumptions, CCA gives only information on the optimum values of species and on the canonical coefficients.

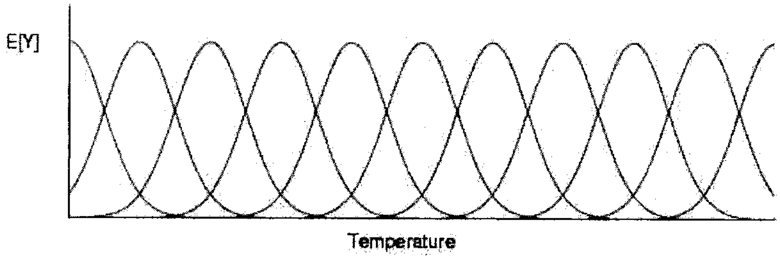

Figure 13.4. Gaussian response curves in the species packing model.

Obviously, the assumptions 1–4 do not hold in practice and they have lead to criticism (Austin and Gaywood 1994). Palmer (1993) showed that CCA is robust against violations of the assumptions. Unfortunately, the simulation studies carried out by Palmer (1993) only concentrate on the estimation of values of $z_i$. As CCA estimates the optimum values and canonical coefficients of the RGR model, one would expect a simulation study that compares these estimated parameters. Zuur (1999) carried out such a simulation study. He looked at what happened if (i) species scores were not evenly spaced along the gradients, and if (ii) species optima and tolerances were not equal for all species. Results indicated that CCA is robust against violations of the assumptions as long as the fourth assumption of Ter Braak (samples cover the whole range of occurrence of species along the environmental variable and are equally spaced) holds.

Using the assumptions 1–4, Ter Braak (1985) showed that Gaussian ordination reduces to a simple iterative algorithm, which gives the same results as correspondence analysis (Greenacre 1984). Based on simulation studies, Ter Braak and Looman (1986) showed that this heuristic solution is robust against violations of the assumptions.

### Historical developments

We started this introduction with the Gaussian response model. Estimating its parameters is basically a regression problem, and this was called (multiple) Gaussian regression. To reduce the number of parameters, we introduced restricted Gaussian regression, which is basically a regression problem with constraints. If no environmental variables have been monitored, Gaussian ordination can be used. This is an *ordination* method. It creates its own latent variables. Finally, we introduced the ordination methods CCA and CA as heuristic solutions for restricted Gaussian regression and Gaussian ordination, respectively. The reason for explaining these techniques in this order (Gaussian regression, restricted regression, Gaussian ordination, CA and CCA) is mainly a logical (in terms of mathematical complexity) one.

Surprisingly, the historical development of these techniques went the other way around. The Gaussian response model has been used by ecologists for many decades. Correspondence analysis was introduced to ecologists by Hill (1973). The method became popular when the software package DECORANA (Hill 1979) was released. CA can probably be considered as the state-of-the-art technique of the 1980s in community ecology. Independently of this, various attempts were made to estimate the parameters of the latent variable model (Equation 13.5) of Gaussian ordination (Kooijman 1977). In 1985, Ter Braak showed that correspondence analysis provides a heuristic approximation of Gaussian ordination if assumptions 1–4 hold. This gave correspondence analysis an ecological rationale. Ter Braak (1986) introduced a restricted form of correspondence analysis, which was called canonical correspondence analysis. CCA is a restricted version of CA in the sense that the axes in CCA are restricted to be linear combinations of environmental variables.

So, the historical development of these techniques went via Gaussian regression, Gaussian ordination, correspondence analysis to canonical correspondence analysis. Ter Braak (1986) argued that CCA is a heuristic approximation of canonical Gaussian ordination, the technique that Zuur (1999) called restricted Gaussian regression.

## 13.2 Three rationales for correspondence analysis

In this section, correspondence analysis (CA) is explained. The reason for this is that CA can be seen as *the* state-of-the-art technique in community ecology in the 1980s, and it forms the basis of canonical correspondence analysis (CCA). We start by presenting three rationales for correspondence analysis, namely a heuristic approximation of Gaussian ordination, reciprocal averaging and weighted principal component analysis. These approaches are computationally equivalent, but they differ in their initial assumptions and interpretations.

### *Rationale 1: Heuristic approximation of Gaussian ordination*

Recall from Section 13.1 that in Gaussian ordination we use the following model

$$E[Y_{ik}] = \mu_{ik} = c_k e^{-\frac{(l_i - u_k)^2}{2t_k^2}}$$

$Y_{ik}$ is the number of species $k$ at site $i$ ($i$ = 1,..,$N$ and $k$ = 1,..,$M$), $l_i$ is the value of the latent variable $l$ at site $i$, with $N$, the total number of sites and $M$ the total number of species. If we assume that $Y_{ik}$ is Poisson distributed with expectation $\mu_{ik}$, the log likelihood function $F$ is given by

$$F(c_1,..,c_M,t_1,...,t_M,u_1,...,u_M,l_1,...,l_N) = \sum_i \sum_k (Y_{ik} \log(\mu_{ik}) - \mu_{ik})$$

Note that $c_k$, $t_k$, $u_k$ and $l_i$ are all estimated from the data. So, the total number of parameters in this model is $3M + N$. We now want to know the values of $u_k$ and $l_i$ that maximise the likelihood. Basic (high school) mathematics dictates calculating the partial derivates of $F$ with respect to $u_k$ and $l_i$, setting them to zero, and solving them. This gives an expression for $u_k$ and $l_i$ that look rather intimidating and we do not show it here. Using the same four assumptions as in Section 13.1, Ter Braak (1985) showed with the help of a simulation study that the partial derivates can be approximated (and simplified) by

$$u_k = \sum_i \frac{Y_{ik}}{Y_{+k}} l_i \qquad \text{and} \qquad l_i = \sum_k \frac{Y_{ik}}{Y_{i+}} u_k$$

where $Y_{i+}$ and $Y_{+k}$ are sums over all species and sites respectively. In matrix notation this becomes

$$\mathbf{u} = \mathbf{D}_c^{-1}\mathbf{Y}'\mathbf{l} \qquad \text{and} \qquad \mathbf{l} = \mathbf{D}_r^{-1}\mathbf{Y}\mathbf{u}$$

where $\mathbf{u}$ = $(u_1, \ldots, u_M)'$, $\mathbf{l}$ =$(l_1, \ldots, l_N)'$ and $\mathbf{Y}$ is a $N$-by-$M$ matrix containing the data. Furthermore, $\mathbf{D}_c$ is an $M$-by-$M$ diagonal matrix with $Y_{+k}$ as $k,k^{th}$ element, and $\mathbf{D}_r$ is a $N$-by-$N$ diagonal matrix with $Y_{i+}$ as $i,i^{th}$ element. The vectors $\mathbf{u}$ and $\mathbf{l}$ are also referred to as species scores, respectively, site scores. In the next two paragraphs, we show that the scores **u** and **l** are the same scores as those obtained by reciprocal averaging, another name for correspondence analysis.

### *Rationale 2: Reciprocal averaging*

Reciprocal averaging (RA) was introduced to ecologists in Hill (1973) and Hill (1974). The aim of RA is to obtain species scores that are weighted averages of site scores and, reciprocally, site scores that are weighted averages of species scores. The algorithm for RA is simple. It starts with arbitrary site scores, and species scores are calculated as the weighted average of the site scores. Then new site scores are calculated as weighted averages of the species scores. This process of reciprocally calculating weighted averages continues until the site scores stabilise. As the range of the weighted averages is smaller than the range of the scores that are used to calculate them, a scaling of site scores is used in each iteration. This prevents the algorithm from drifting into a small range of scores.

Before we present the algorithm for RA, we first need to introduce some more mathematical notation. Let **r** be a $N$-by-1 vector containing the site (row) proportions of **Y**. So, the $i^{th}$ element of **r** is equal to $Y_{i+}/Y_{++}$. Similarly, let **c** be an $M$-by-1 vector containing species (column) proportions. Finally, let $\mathbf{x} = (x_1,\ldots,x_N)'$ and $\mathbf{u} = (u_1,\ldots,u_M)'$. The algorithm for RA has the following form:

1. Start with arbitrary site scores $x_i$.
2. Calculate species scores by $u_k = \sum_i Y_{ik} x_i / Y_{+k}$.
3. Calculate site scores by $x_i = \sum_k Y_{ik} u_k / Y_{i+}$.
4. Standardise the site scores $x_i$ using a weighted mean and standard deviation. The weights are given by site totals.
5. Stop on convergence; else go to step 2.

To obtain further axes, an orthogonalisation procedure can be used between steps 3 and 4 in each iteration (Ter Braak and Prentice 1988). In such a procedure, the site scores are kept uncorrelated with previous axes by a weighted multiple regression of $x_i$ on previous axes. The calculations of the algorithm in the last iteration were

$$\mathbf{u} = \mathbf{D}_c^{-1}\mathbf{Y}'\mathbf{x} \quad \text{and} \quad \mathbf{x} = s^{-1}\mathbf{D}_r^{-1}\mathbf{Y}\mathbf{u} \tag{13.6}$$

The $s$ comes from the standardisation step. So species scores **u** are weighted averages of the site scores **x**, and site scores are *proportional* (due to the $s$) to the weighted averages of species scores. Now suppose that the roles of **u** and **x** interchange (this is called the dual problem). We start with arbitrary species scores, calculate site scores as a weighted average of species scores, and calculate species scores as a weighted average of site scores. The results of the last iteration of the algorithm would be

$$\mathbf{u} = s^{-1}\mathbf{D}_c^{-1}\mathbf{Y}'\mathbf{x} \quad \text{and} \quad \mathbf{x} = \mathbf{D}_r^{-1}\mathbf{Y}\mathbf{u} \tag{13.7}$$

Thus species scores are proportional to weighted averages of site scores and site scores are weighted averages of species scores. For $\alpha = 0$ respectively $\alpha = 1$, equations (13.6) and (13.7) can be summarised by

$$\mathbf{u} = s^{-\alpha}\mathbf{D}_c^{-1}\mathbf{Y}'\mathbf{x} \quad \text{and} \quad \mathbf{x} = s^{1-\alpha}\mathbf{D}_r^{-1}\mathbf{Y}\mathbf{u}$$

### *Rationale 3: Weighted principal component analysis*

Calculations in principal component analysis (PCA) are made in a Euclidean metric. In the early 1970s, the Frenchman Benzecri and co-workers developed a similar method using weights in a Chi-square metric. This method was called 'analyses des correspondence', translated by Hill (1974) into correspondence analysis. A good overview, with applications, of this weighted principal component analysis is given in Greenacre (1984). The biplot can be used in combination with CA (Krzanowski 1988; Gabriel and Odoroff 1990; Greenacre 1993; Gabriel 1995; Ter Braak and Verdonschot 1995; Jongman et al. 1995; Gower and Hand 1996; Legendre and Legendre 1998; Jolliffe 2002).

The starting point in CA is a contingency table. This is a table that gives the counts in the dataset for all combinations of categories of each variable (Krzanowski and Marriott 1994). Most of the theory of CA is based on two-way contingency tables, but extensions to higher dimensions are popular in fields like psychology. Analysing a two-way contingency table, the first question that arises is whether there is a relation between the two variables in the contingency table. This can be investigated by performing a Chi-square test; see also Chapter 10 (Table 10.4) for a worked example. The null hypothesis is the independence of the two variables, and the test statistic (Pearson Chi-square) is

$$\mathrm{X}^2 = \sum\sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \sum_i \sum_k \frac{(Y_{ik} - Y_{i+}Y_{+k}/n)^2}{Y_{i+}Y_{+k}/n}$$

The notation $Y_{i+}$, $Y_{+k}$ and $n$ stand for row total, column total and overall total ($n = Y_{++}$). Using this statistic, you can easily test the null hypothesis. If this hypothesis is rejected, it is valuable to analyse why it is rejected and to see which cells account for the relations between the two variables. In Chapter 10, we calculated the Chi-square statistic for an artificial dataset and determined the contribution of each cell to the test statistic. Define $q_{ik}$ for species $k$ at site $i$ as

$$q_{ik} = \frac{p_{ik} - p_{i+}p_{+k}}{\sqrt{p_{i+}p_{+k}}}$$

where $p_{ik} = Y_{ij}/n$, $n$ is the sum of all species at all sites, $p_{i+}$ is the total abundance at site $i$, and $p_{+k}$ the total abundance for species k. Using basic algebra it can be shown that $n \times q_{ik}^2$ is the contribution of cell $i,k$ to the Chi-square statistic (these are the values in bold font in Table 10.4). The matrix $\mathbf{Q}$, containing the elements $q_{ik}$ for all sites and species, is the starting point in correspondence analysis. Just as in PCA (Chapter 12), the singular value decomposition is used to decompose **Q** into three special matrices:

$$\mathbf{Q} = \mathbf{U}\ \mathbf{L}\ \mathbf{V}' \tag{13.8}$$

The matrices **U** and **V** are orthonormal and **L** contains the square roots of the eigenvalues. Equation (13.1) can be used to provide a low-dimensional approximation of (i) the relationships between sites, (ii) the relationships between species, and (iii) the relationships between sites and species. Just as in PCA, we can choose which aspect to focus on by converting the right part in equation (13.8) into 'something' related to the sites and 'something' to the species. This requires premultiplying **U** and **V** with diagonal matrices containing sites totals and species totals, and we also need to decide where to put the **L**. Technical details can be found in Legendre and Legendre (1998). Just as in PCA, the first few, say two, rows and columns of these matrixes can be plotted in one graph, and these provide a low-dimensional approximation of the information in **Q**.

This process is called scaling, and there are three main choices. We will discuss and illustrate three scaling choices using coverage indices from lowland plant species from 20 sites in Mexico. A full analysis of these data is given in Chapter 32. Here, we only use the 15 most frequently measured families that give us a data matrix of dimension 20-by-15.

In CA, the position of a species represents the optimum value in terms of the Gaussian response model (niche) along the first and second axes. For this reason, most ecological software packages present a species scores as a point or label, and not by a line.

*Scaling 1* is appropriate if one is interested in sites, because distances between sites in the ordination diagram are two-dimensional approximations of their Chi-square distances. The sites are at the centroid of the species. This means that the sites are scattered near the species that occur at those sites. An example is given in Figure 13.5. Site 3 is rather different from the other sites (it has a relatively large Chi-square distance to the other sites). The centroid rule dictates that this site has relatively large values for ci, vo and com.

*Scaling type 2* (or species conditional scaling) is appropriate if one is interested in species because distances between species are two-dimensional approximations of their Chi-square distances. The species are at the centroid of the sites. This means that the species points are close to the sites where they occur. An example is given in Figure 13.6. Grcyn is rather different from the other species as it has a large Chi-square distance to all other species. The centroid rule indicates that grcyn has high values at all sites in the lower left quadrant.

*Scaling type 3* results in a graph in which distances between sites are two-dimensional approximations of their Chi-square distances, and the same holds for the species. But the species and the sites cannot be compared with each other. An example of this scaling is presented in Figure 13.7. The family grcyn is rather different (in terms of the Chi-square distance) from all other families. Site 3 is rather different from the other sites. Distances between sites and species cannot be interpreted. The joint plot of species and sites under this scaling has caused a lot of confusion in the literature. Greenacre (1984) warns not to interpret the joint plot, because it has no formal justification.

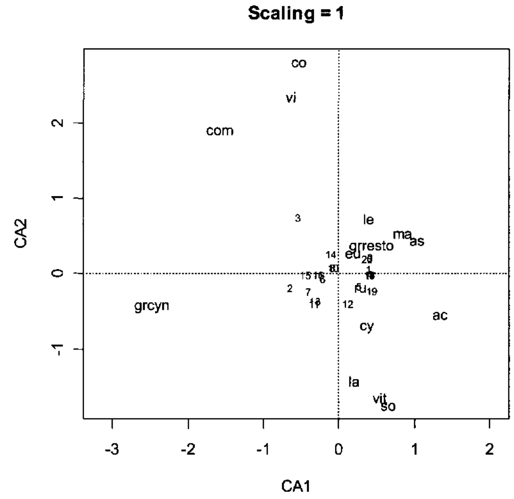

Figure 13.5. Site conditional biplot. Distances between sites are two-dimensional approximations of their Chi-square distances. The first two eigenvalues are 0.13 and 0.06, and the total variation (inertia) is 0.34.

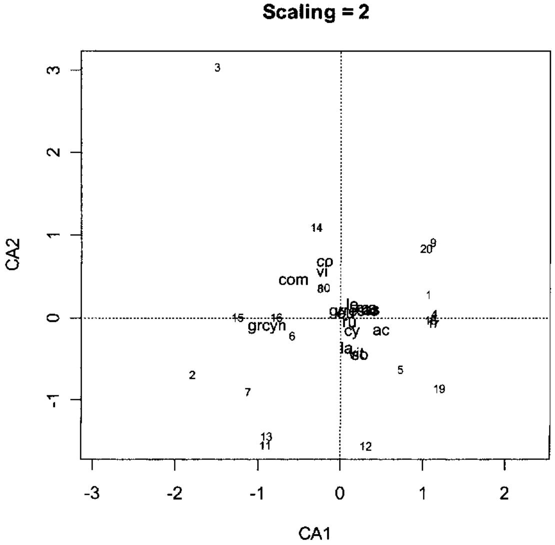

Figure 13.6. Species conditional biplot. The first two eigenvalues are 0.13 and 0.06, and the total variation (inertia) is 0.34.

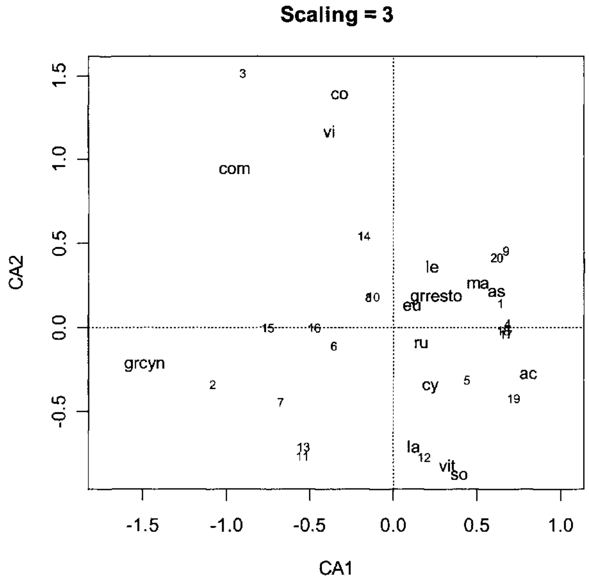

Figure 13.7. Joint plot of species and site scores. Distances between species are two-dimensional approximations of their Chi-square distances. Distances between sites (the numbers) are also two-dimensional approximations of Chi-square distances. Distances between species and sites cannot be interpreted. The first two axes explain 57% of the variation.

The total inertia (or total variance) in CA is defined as the Chi-square statistic of the site-by-species table divided by the total number of observations. Points far away from the origin in each diagram are the most interesting, because these points make a relatively higher contributions to the Chi square statistic than points nearer the origin. So the further away from the origin that a site is plotted, the more different it is from the average site. A numerical example may help. Suppose we have the following (artificial) data:

| | Species 1 | Species 2 | Species 3 | Total |
|---|---|---|---|---|
| Site 1 | 1 | 2 | 3 | 6 |
| Site 2 | 0 | 1 | 2 | 3 |
| Site 3 | 2 | 1 | 0 | 3 |
| Site 4 | 4 | 3 | 3 | 10 |
| Total | 7 | 7 | 8 | 22 |

In scaling 2, we look at species profiles; the abundance of each species is divided by the species total, and the profiles are then compared with the average profile:

| | Species 1 | Species 2 | Species 3 | **Average** |
|---|---|---|---|---|
| Site 1 | 1/7 | 2/7 | 3/7 | **6/22** |
| Site 2 | 0 | 1/7 | 2/7 | **3/22** |
| Site 3 | 2/7 | 1/7 | 0 | **3/22** |
| Site 4 | 4/7 | 3/7 | 3/7 | **10/22** |

In scaling 1, we do it the other way around. Row (site) profiles are calculated, and these are compared with the average profile:

| | Species 1 | Species 2 | Species 3 |
|---|---|---|---|
| Site 1 | 1/6 | 2/6 | 3/6 |
| Site 2 | 0/6 | 1/3 | 2/3 |
| Site 3 | 2/6 | 1/3 | 0 |
| Site 4 | 4/10 | 3/10 | 3/10 |
| **Average** | **7/22** | **7/22** | **8/22** |

Section 10.1 contains an example of how these profiles are compared with each other (Chi-square distance). Just as in PCA, eigenvalues can be used to assess how much variation is explained by each axis. Instead of total variance, the total variation is called inertia.

### Heuristic approximation, CA and RA

Because RA and CA share the same eigenvalue problems, the estimated scores are identical. Furthermore, the species scores **u** and site scores **x** obtained by RA are similar to the scores obtained by the heuristic approximation (rationale 1). This means that we now have three different approaches: Reciprocal averaging, weighted PCA, and the heuristic approximation of Gaussian ordination, which all give the same estimated species and site scores. As well as these three approaches, several other approaches exist, for example, dual scaling (Greenacre 1984).

The difference between RA, CA and the heuristic approximation of Gaussian ordination, besides the estimation procedure, concerns the type of data on which they can be used. RA can analyse any data, as long as the data are non-negative and have the same units. The heuristic approximation of Gaussian ordination assumes that data are Poisson distributed. Correspondence analysis was presented as a method that analyses contingency tables. However, in many textbooks CA is applied to other kinds of datasets; see for example Gauch (1982), Greenacre (1984), Jambu (1991) or Jongman et al. (1995). The use of CA on such datasets can be justified by considering the table as a distribution of a certain amount of *mass* (e.g., weight, length, volume) over the cells. Although it is still interesting to look at row-column interactions, the Chi-square statistic cannot now be used to *test* row-column independence. It merely serves as a measure of association.

In the rest of this chapter, the name *correspondence analysis* will be used for reciprocal averaging, weighted PCA and the heuristic approximation of Gaussian ordination. In the next section, the biplot interpretation is introduced in the context of weighted PCA. Because of the similarities among these approaches, the biplot can be used with all three methods.

## 13.3 From RGR to CCA

Recall from Section 13.1 that the restricted Gaussian response model has the following form:

$$Y_i = ce^{-\frac{(z_i-u)^2}{2t^2}} = \exp(b_1 + b_2 z_i + b_3 z_i^2) \tag{13.9}$$

$$z_i = \sum_{p=1}^{Q} \alpha_p x_{ip}$$

$Y_{ik}$ is the abundance (counts) of species $k$ at site $i$, $x_{ip}$ is the value of the $p^{\text{th}}$ environmental variable at site $i$, $z_i$ is the value of the gradient at site $i$, and $a_p$, $c_k$, $t_k$ and $u_k$ are unknown parameters. Because $Y_{ik}$ are counts, it is common to assume that $Y_{ik}$ is Poisson distributed with expectation $\mu_{ik}$. The log likelihood function is given by

$$F(\mathbf{\alpha},\mathbf{c},\mathbf{t},\mathbf{u}) = \sum\nolimits_i \sum\nolimits_k (Y_{ik} \log(\mu_{ik}) - \mu_{ik})$$

where $\mathbf{\alpha} = (\alpha_1,\ldots,\alpha_Q)$, $\mathbf{c} = (c_1,\ldots,c_M)$, $\mathbf{t} = (t_1,\ldots,t_M)$ and $\mathbf{u} = (u_1,\ldots,u_M)$. An option to obtain parameter estimates for $\mathbf{\alpha}$, $\mathbf{c}$, $\mathbf{t}$ and $\mathbf{u}$ is to formulate the partial differential equations of $F$ with respect to the parameters, set these to zero, and use numerical optimisation routines to solve them. Instead of doing this, Ter Braak (1986) derived partial differential equations of $F$ with respect to $u_k$, $c_k$, $t_k$ and $\alpha_p$. Just as for the Gaussian response model, these partial derivatives do not look friendly. In Section 13.2, four assumptions were used to simplify the partial differential equations resulting in CA. Ter Braak (1986) used the same four assumptions and obtained a considerably easier set of equations for RGR. The resulting technique is called CCA. As a result of assumptions 1–4, we obtain the following set of equations:

$$\mathbf{u} = \mathbf{D}_c\mathbf{Y}'\mathbf{Z} \quad \text{and} \quad \mathbf{Z}_{wa} = \mathbf{D}_r^{-1}\mathbf{Y}\mathbf{u} \quad \text{and} \quad \alpha = (\mathbf{X}'\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_r\mathbf{Z}_{wa}$$

Note that the parameter $\mathbf{\alpha}$ is the weighted least-squares solution of the regression of $\mathbf{Z}_{\text{wa}}$ on $\mathbf{X}$. The scores $\mathbf{u}$ will be referred to as species scores, $\mathbf{Z}_{\text{wa}}$ as site scores that are weighted averages of species scores (weights are given by site totals), and $\mathbf{Z}$ as site scores that are a linear combination of environmental variables. The latter scores are also denoted by $\mathbf{Z}_{\text{env}}$. To calculate the species and site scores, the following algorithm can be used:

1. Start with arbitrary site scores $\mathbf{Z}$.
2. Calculate species scores $\mathbf{u}$, which are weighted averages of site scores, by $\mathbf{u} = \mathbf{D}_c^{-1}\mathbf{Y}'\mathbf{Z}$.
3. Calculate site scores $\mathbf{Z}_{\text{wa}}$, which are weighted averages of species scores, by $\mathbf{Z}_{\text{wa}} = \mathbf{D}_r^{-1}\mathbf{Y}\mathbf{u}$.
4. Use weighted linear regression of the site scores $\mathbf{Z}_{\text{wa}}$ on environmental variables $\mathbf{X}$, and obtain the regression coefficients by $\mathbf{\alpha} = (\mathbf{X}'\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_r\mathbf{Z}_{\text{wa}}$.

5. Obtain new, estimated site scores $\mathbf{Z}$, which are a linear combination of environmental variables, by $\mathbf{Z}_{\text{env}} = \mathbf{X}\, \boldsymbol{\alpha}$ and set $\mathbf{Z}$ equal to these.
6. Standardise the estimated site scores $\mathbf{Z}$.
7. Stop on convergence; else go to step 2.

As the range of the weighted averages is smaller than the range of the scores that are used to calculate them, a scaling of scores $\mathbf{Z}_{\text{env}}$ (= $\mathbf{Z}$) is used in each iteration (step 6). This prevents the algorithm from drifting into a small range of scores and avoids a trivial solution. To obtain a second axis (or further axes), a weighted regression can be carried out in each iteration. In such a procedure, we regress $\mathbf{Z}_{\text{env}}$ on the first axis (or previous axes) and continue to work with the residuals of this regression. See Ter Braak and Prentice (1988) for more details. In a similar way, the linear effects of particular environmental variables can be filtered out. This technique is called partial CCA. It can be useful if one is not interested in the effects of these particular environmental variables.

Note that the algorithm for CCA is similar to the algorithm for reciprocal averaging (alias correspondence analysis). From a mathematical point of view, CCA can be seen as correspondence analysis in which the axes are restricted to be linear combinations of environmental variables. Put simply, CCA is a CA in which the axes are restricted to be linear combinations of explanatory variables.

### *Inertia*

The inertia (or total variance) in CCA is identified the same way as in CA. The eigenvalue of an axis is given by the weighted standard deviation $s$, which is calculated in step 6 of the CCA algorithm. This can be seen by making similar substitutions as in CA. These eigenvalues are also called canonical eigenvalues. The amount of variation that can be explained by all the environmental variables is equal to the sum of all canonical eigenvalues. The amount of variation explained by the first two axes can be expressed as a percentage of the total inertia, and as a percentage of the variance that can be explained by the environmental variables.

### *Canonical coefficients and intraset correlations*

So, how do we know which explanatory variables are important? There are two tools for this: The final regression coefficients $\boldsymbol{\alpha}$, also called canonical coefficients, and the intraset correlations defined as the correlations between environmental variables and axes $\mathbf{Z}_{\text{env}}$. The intraset correlations are also called environmental scores. The environmental variables are standardised, which makes a comparison of the canonical coefficients possible. How to interpret the species scores $\mathbf{u}$, site scores $\mathbf{Z}_{\text{wa}}$ and $\mathbf{Z}_{\text{env}}$, intraset correlations and canonical coefficients is explained in the next section.

## 13.4 Understanding the CCA triplot

The species scores **u**, site scores $\mathbf{Z}_{\text{env}}$ and intraset correlations obtained by CCA are plotted in a figure called a triplot. For the same reason as in CA, species are represented by labels, sites by points or labels, and the explanatory variables by lines. Species points represent the optimum parameter of the Gaussian response model (niche) and can be projected on the axes but also on the explanatory variables showing the optimum value along each of them. Nominal explanatory variables are dealt with in the same way as in RDA. An example of a triplot is presented in Figure 13.8. This triplot was obtained by applying CCA on the Mexican plant data. The same families as in the previous section were used, and we only used four explanatory variables. Families are represented by their name (abbreviated, see Chapter 32), sites by numbers 1–20 and intraset correlations by lines starting at the origin to the point with coordinates given by the intraset correlations of the two axes.

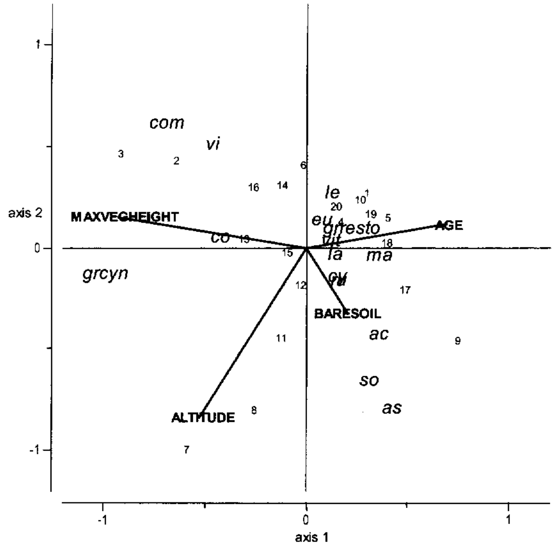

Figure 13.8. Triplot for the Mexican plant data, obtained by canonical correspondence analysis. Species conditional scaling was used.

In a triplot, species scores, site scores and environmental scores are plotted in the same graph and form a series of biplots. These biplots are based on the species scores and site scores, the species scores and intraset correlations (representing the explanatory variables), and the site scores and intraset correlations. CCA produces two sets of site scores; $\mathbf{Z}_{\text{wa}}$ and $\mathbf{Z}_{\text{env}}$. The standard choice is to use the scores $\mathbf{Z}_{\text{env}}$

in the triplot. The motivation for this is that $\mathbf{Z}_{env}$ can be used for two biplots, whereas this is not the case for $\mathbf{Z}_{wa}$.

As with CA, you can choose from various scaling options, such as the species conditional scaling (called scaling 2 in CA) or the site conditional scaling (called scaling 1 in CA). If the interest is on species, then the most sensible choice is species conditional scaling. The same holds for the sites and scaling 1.

Concentrating on the species scores and site scores in the species conditional triplot in Figure 13.8, the interpretation is identical to that in CA. So comparing species (families in this example) scores give information about which species behave similarly. Species close to each other are similar in terms of Chi-square distances, and species relatively far away from the origin, contribute more to the inertia than species close to the origin. The species scores are at the centroids of the sites scores, which allows us to infer relative abundances (just as in CA). Using this interpretation, one can infer from Figure 13.8 that the families so, as, and ac deviate from the average profile at sites 7, 8 and 9. These families are positively related to each other, but negatively to vi and com.

The species scores and intraset correlations (the explanatory variables) can also be compared. Species can be projected perpendicular on the lines showing the species optima. If sites scores are projected perpendicular on the lines, we can infer the values of the environmental variables at those sites. Results indicate that the family grcyn occurs at high values of maximum vegetation height, and low values of age. At sites 2 and 3, maximum vegetation height is large. In fact, we could draw the species packing model (Figure 13.4) along each line.

Recall that the intraset correlations are the correlations between the axes and the original explanatory variables, and the canonical coefficients are the $\alpha$s defining the linear combination of explanatory variables constituting the gradient.

As to the intraset correlations, the tip of a line (representing intraset correlations) can be projected perpendicularly on another line, and the weighted correlation between them is inferred. The tip of a line can also be projected on the axes, and the correlation between the corresponding environmental variable and the axes is inferred. The lines in the triplot indicate that the environmental variables age and maximum vegetation height are negatively correlated. Projecting the lines on the axes shows that the first axis is highly correlated with maximum vegetation height. The second axis is correlated with altitude.

With the canonical coefficients, recall the gradients are linear combinations of environmental variables. The exact form of this linear combination is determined by the canonical coefficients. The canonical coefficients for the first axis are 0.18 (altitude), −0.28 (age), −0.23 (bare soil) and 0.32 (maximum vegetation height). The first axis is mainly determined by maximum vegetation height versus age.

Finally, we discuss the eigenvalues. The total inertia (variation) is 0.33, and the sum of all canonical eigenvalues is 0.15. Hence, all four explanatory variables explain 45% (= 100 × 0.15/0.33) of the variation in the data. The first two eigenvalues are 0.10 and 0.02. Making 82% of the total inertia explained by the first two axes. And 37% (= 82% of 45%) of the variation that can be explained with the environmental variables is explained by the first two axes. Both percentages are relatively high for ecological datasets.

In the distance biplot (scaling 1), distances between sites represent (approximate) Chi-square distances, but distances between species cannot be interpreted.

## 13.5 When to use PCA, CA, RDA or CCA

At this point we introduce two measures of diversity: Alpha and beta diversity. Alpha diversity is the diversity of a site and beta diversity measures the change in species composition from place to place, or along environmental gradients. Examples of these diversity measures are given in Figure 13.9. The total beta diversity is the 'gradient length'. A short gradient has low beta diversity. As explained above, Ter Braak (1986) showed that CA is an approximation of Gaussian ordination and CCA is an approximation of restricted Gaussian regression. This is the ecological rationale of CA and CCA. PCA and RDA analyse linear responses along the gradient, and CA and CCA look at unimodal responses along the gradient. This is summarised in Table 13.1 and described in more detail below:

1. PCA should be used to analyse species data if the relations along the gradients are linear.
2. RDA should be used to analyse linear relationships between species and environmental variables.
3. CA analyses species data and unimodal relations along the gradients.
4. CCA can be used to analyse unimodal relationships between species and environmental variables.
5. PCA or RDA should be used if the beta diversity is small, or if the range of the samples covers only a small part of the gradient.
6. A long gradient has high beta diversity, and this indicates that CA or CCA should be used.

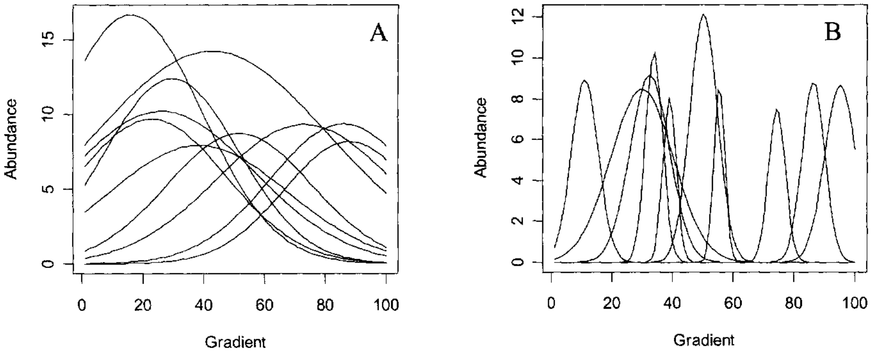

Figure 13.9. Artificial response curves showing high alpha and low beta diversity (A) and low alpha and high beta diversity (B). PCA and RDA should be applied on data in panel A and CA and CCA in panel B.

Table 13.1. Summary of methods. Relationships in PCA and RDA are linear. In RDA and CCA two sets of variables are used, and a cause-effect relationship is assumed.

| | Indirect Gradient Analysis | Direct Gradient Analysis |
|---|---|---|
| Linear model | PCA | RDA |
| Unimodal model | CA | CCA |

## 13.6 Problems with CA and CCA

CA and CCA are useful techniques as long as the data matrix does not contain too many zeros. Figure 13.10 shows what happens if correspondence analysis is applied on the RIKZ data used in Chapter 27. The data contain many observations equal to zero; there are a few species measured at only one site and a few sites where only one species was observed. These more extreme observations for species and sites dominate the first few axes! We also had to remove two sites because no species were observed. Obviously, we can remove such species and sites, but the question is how much data can you afford to remove.

Another potential problem is the arch effect (Chapter 12), which again is due to the many observations equal to zero. Figure 13.11 shows a CCA triplot (species conditional) for the full Mexican plant data. Instead of using averages per pasture, we use all 200 observations. Note that the shape of the site scores may indicate the presence of an arch effect. If this is the case, then there are three options: (i) Argue that the arch shape is a real pattern in the site scores caused by the explanatory variables (risking the referee rejecting the paper because they think that detrended canonical correspondence analysis should have been used), (ii) apply detrended canonical correspondence analysis to bring down both ends of the arch (risking the referee rejecting the paper because they do not agree that detrended correspondence analysis is an appropriate technique), or (iii) applying a special (Chord or Hellinger) transformation followed by an RDA and visualise Chord distances. Detrended correspondence analysis is an artificial way to remove the arch effect by splitting up the axis in segments and detrending the scores in each segment. Obviously, any real pattern will also be removed. Our choice is option (iii); the other two each have 50% chance of getting past a referee. Some books will condemn detrended canonical correspondence analysis (and detrended correspondence analysis), with some software programmers even refusing to add it to their software; yet other books are positive about it. McCune and Grace (2002) say: 'Detrended CA unnecessarily imposes assumptions about the distribution of samples and species in environmental spaces. .... There is no need to use detrended CA'. And Legendre and Legendre (1998) write: 'Present evidence indicates that detrending should be avoided, ...'. Both books advise to use non-metric multidimensional scaling if there is an arch effect.

The case study chapters contain various examples of PCA, RDA, CCA and variance partitioning and expand on the ideas presented in this chapter.

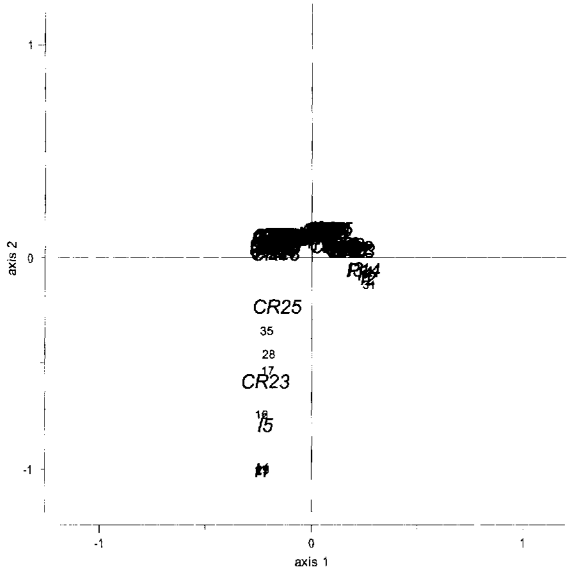

Figure 13.10. Correspondence analysis on the RIKZ data. The species conditional scaling was used. A few species were only measured at one site, and a few sites only had one species (with low values). As a result, the first few axes are dominated by these species and sites.

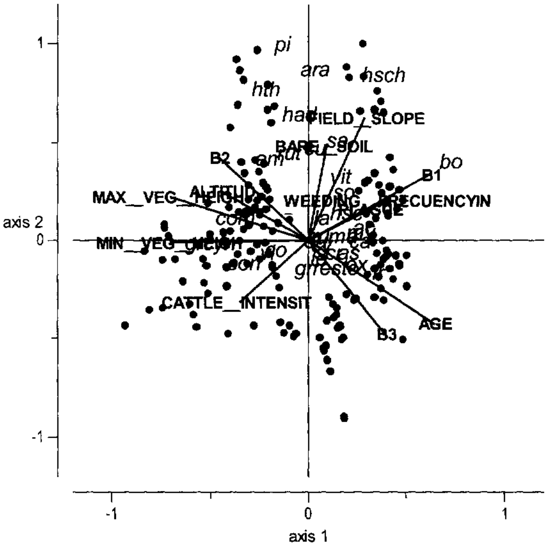

Figure 13.11. CCA applied on the full Mexican plant data. The sites scores (dots) may exhibit the arch effect as they show a U-shape.