Canonical Correspondence Analysis: A New Eigenvector Technique for Multivariate Direct Gradient Analysis

Cajo J. F. Ter Braak

Ecology, Vol. 67, No. 5 (Oct., 1986), 1167-1179.

Stable URL:
http://links.jstor.org/sici?sici=0012-9658%28198610%2967%3A5%3C1167%3ACCAANE%3E2.0.CO%3B2-U

Ecology, 67(5), 1986, pp. 1167–1179
© 1986 by the Ecological Society of America

CANONICAL CORRESPONDENCE ANALYSIS: A NEW EIGENVECTOR TECHNIQUE FOR MULTIVARIATE DIRECT GRADIENT ANALYSIS[1]

Cajo J. F. Ter Braak
TNO Institute of Applied Computer Science, P. O. Box 100, 6700 AC Wageningen, The Netherlands, and Research Institute for Nature Management, Leersum, The Netherlands

Abstract. A new multivariate analysis technique, developed to relate community composition to known variation in the environment, is described. The technique is an extension of correspondence analysis (reciprocal averaging), a popular ordination technique that extracts continuous axes of variation from species occurrence or abundance data. Such ordination axes are typically interpreted with the help of external knowledge and data on environmental variables; this two-step approach (ordination followed by environmental gradient identification) is termed indirect gradient analysis. In the new technique, called canonical correspondence analysis, ordination axes are chosen in the light of known environmental variables by imposing the extra restriction that the axes be linear combinations of environmental variables. In this way community variation can be directly related to environmental variation. The environmental variables may be quantitative or nominal. As many axes can be extracted as there are environmental variables. The method of detrending can be incorporated in the technique to remove arch effects.

(Detrended) canonical correspondence analysis is an efficient ordination technique when species have bell-shaped response curves or surfaces with respect to environmental gradients, and is therefore more appropriate for analyzing data on community composition and environmental variables than canonical correlation analysis. The new technique leads to an ordination diagram in which points represent species and sites, and vectors represent environmental variables. Such a diagram shows the patterns of variation in community composition that can be explained best by the environmental variables and also visualizes approximately the "centers" of the species distributions along each of the environmental variables. Such diagrams effectively summarized relationships between community and environment for data sets on hunting spiders, dyke vegetation, and algae along a pollution gradient.

Key words: biplot; canonical correlation analysis; canonical correspondence analysis; detrended correspondence analysis; Gaussian model; gradient analysis; ordination; reciprocal averaging; regression; species–environment relations; unfolding; weighted averaging.

Introduction

Problems in community ecology often require the inferring of species–environment relationships from community composition data and associated habitat measurements. Typical data for such problems consist of two sets: data on the occurrence or abundance of a number of species at a series of sites, and data on a number of environmental variables measured at the same sites. (A "site" is the basic sampling unit, separated in space or time from other sites, e.g., a quadrat, a woodlot, a light trap, or a plankton sample.) When the data are collected over a sufficient habitat range for species to show nonlinear, nonmonotonic relationships with environmental variables, it is inappropriate to summarize these relationships by correlation coefficients or to analyze the data by techniques that are based on correlation coefficients, such as canonical correlation analysis (Gauch and Wentworth 1976, Gittins 1985). An alternative, two-step approach has become popular: (1) extract from the species data the dominant pattern of variation in community composition by an ordination technique, such as (detrended) correspondence analysis, and (2) attempt to relate this pattern (i.e., the first few ordination axes) to the environmental variables (Gauch 1982*a*). The particular merit of detrended correspondence analysis in this context is that it removes nonlinear dependencies between axes (Hill and Gauch 1980) and has been shown to be an efficient technique to extract one or more ordination axes ("gradients") such that species show unimodal (bell-shaped) response curves or surfaces with respect to these axes (Ter Braak 1985*b*). The axes can be thought of as hypothetical environmental gradients, which are subsequently interpreted in terms of measured environmental variables in the second step of the analysis. This two-step approach is essentially Whittaker's (1967) indirect gradient analysis.

What can be inferred from indirect gradient analysis? If the measured environmental variables relate strongly to the first few ordination axes, they can "account for" (i.e., they are sufficient to predict) the main part of the variation in the species composition. If the environmental variables do not relate strongly to the first few axes, they cannot account for the main part of the variation, but they may still account for some of the remaining variation—which can be substantial. Further, it is nontrivial to detect by indirect gradient anal-

[1] Manuscript received 18 March 1985; revised 12 November 1985; accepted 22 January 1986.

ysis the effects on community composition of a subset of environmental variables in which one is particularly interested (Carleton 1984). These limitations can only be overcome by methods of direct gradient analysis, in which species occurrences are related directly to environmental variables (Gauch 1982*a*). Methods of direct gradient analysis in current use consider essentially one species at a time. Simple methods involve plotting species abundance against a single environmental variable, or isopleths in a space of two environmental variables (Whittaker 1967). More elaborate methods use (generalized linear) regression methods (Austin et al. 1984, Bartlein et al. 1986) and are useful in studying simultaneously the effect of more than one environmental variable. Regression methods allow fitted response surfaces to assume a wide variety of shapes. However, when the number of species is large, separate regression analysis for each species may be impractical. Moreover, separate analyses cannot be combined easily to get an overview of how community composition varies with the environment (in particular, when the number of environmental variables exceeds two or three), and a multivariate method (based on a common response model) is required.

In this paper a multivariate direct gradient analysis technique is developed, whereby a set of species is related directly to a set of environmental variables. The new technique identifies an environmental basis for community ordination by detecting the patterns of variation in community composition that can be explained best by the environmental variables. In the resulting ordination diagram, species and sites are represented by points and environmental variables are represented by arrows. Such a diagram shows the main pattern of variation in community composition as accounted for by the environmental variables, and also shows, in an approximate way, the distributions of the species along each environmental variable. The technique thus combines aspects of regular ordination with aspects of direct gradient analysis. The rationale of the technique is derived from a species packing model wherein species are assumed to have Gaussian (bell-shaped) response surfaces with respect to compound environmental gradients. These gradients are assumed to be linear combinations of the environmental variables. The new technique is called canonical correspondence analysis, because it is a correspondence analysis technique in which the axes are chosen in the light of the environmental variables. Examples demonstrate that canonical correspondence analysis allows a quick appraisal of how community composition varies with the environment.

THEORY

Data and model

Suppose a survey of n sites lists the abundances or occurrences (presence scored as 1, absence as 0) of m species and the values of q environmental variables ($q < n$). Let y_{ik} be the abundance or presence/absence (1/0) of species k ($y_{ik} \geq 0$), and z_{ij} the value of environmental variable j at site i.

The first step in indirect gradient analysis is to summarize the main variation in the species data by ordination. The method of Gaussian ordination (Gauch et al. 1974) does this by constructing an axis such that the species data optimally fit Gaussian response curves along this axis. Then the response model for the species is the bell-shaped function

$$E(y_{ik}) = c_k \exp[{}^1\!/\!_2(x_i - u_k)^2/t_k{}^2], \tag{1}$$

where $E(y_{ik})$ denotes the expected (average) value of y_{ik} at site i that has score x_i on the ordination axis. The parameters for species k are c_k, the maximum of that species' response curve; u_k, the mode or optimum (i.e., the value of x for which the maximum is attained); and t_k, the tolerance, a measure of ecological amplitude. Ter Braak (1985*b*) showed that correspondence analysis approximates the maximum likelihood solution of Gaussian ordination, if the sampling distribution of the species abundances is Poisson, and if:

C1) the species' tolerances are equal ($t_k = t$, $k = 1, \ldots, m$),

C2) the species' maxima are equal ($c_k = c$, $k = 1, \ldots$, m),

C3) the species' optima $\{u_k\}$ are homogeneously distributed over an interval A that is large compared to t,

C4) the site scores $\{x_i\}$ are homogeneously distributed over a large interval B that is contained in A.

(The wording "homogeneously distributed" is used to cover either of two cases, namely (1) that the scores are equispaced, with spacing small compared to t, or (2) that the scores are drawn randomly from a uniform distribution.) Conditions C1–C3 imply a species packing model (Whittaker et al. 1973) with respect to the ordination axis. The species scores resulting from a correspondence analysis actually estimate the optima of the species in this model. Ter Braak (1985*b*) provided a similar rationale for correspondence analysis of presence-absence data. Conditions C1 and C2 are not likely to hold in most natural communities, but the usefulness of correspondence analysis in practice relies on its robustness against violations of these conditions (Hill and Gauch 1980).

The second step of indirect gradient analysis is to relate the ordination axis to the environmental variables, for example graphically, or by calculating correlation coefficients, or by multiple regression (see Montgomery and Peck 1982) of the site scores on the environmental variables

$$x_i = b_0 + \sum_{j=1}^{q} b_j z_{ij}, \tag{2}$$

where b_0 is the intercept and b_j is the regression coefficient for environmental variable j. Note that the species optima u_k and sites scores x_i are estimated from the species data first; the regression coefficients b_j are estimated next, keeping x_i (and u_k) fixed. The species data are thus indirectly related to the environmental variables, via the ordination axis.

The technique proposed in this paper simultaneously estimates the species optima, the regression coefficients and, hence, the site scores by using the model described by Eq. 1, in conjunction with Eq. 2. Simultaneous estimation turns the technique into a direct gradient analysis method. In principle the method of maximum likelihood could be used to obtain the estimates. This analysis could be called Gaussian canonical ordination. It requires excessively heavy computation. The computational task can, however, be alleviated considerably if conditions C1–C4 hold. The reasoning that led from Gaussian ordination to correspondence analysis, now leads to the transition formulae of canonical correspondence analysis (see Appendix):

$$\lambda u_k = \sum_{i=1}^{n} y_{ik} x_i / y_{+k} \tag{3}$$

$$x_i^* = \sum_{k=1}^{m} y_{ik} u_k / y_{i+} \tag{4}$$

$$\boldsymbol{b} = (\mathrm{Z}'\mathrm{R}\mathrm{Z})^{-1}\mathrm{Z}'\mathrm{R}\boldsymbol{x}^* \tag{5}$$

$$\boldsymbol{x} = \mathrm{Z}\boldsymbol{b}, \tag{6}$$

where y_{+k} and y_{i+} are species and site totals, respectively, R is a diagonal $n \times n$ matrix with y_{i+} as the (i, i)-th element; $\mathrm{Z} = \{z_{ij}\}$ is an $n \times (q + 1)$ matrix containing the environmental data and a column of ones; and $\boldsymbol{b}$, $\boldsymbol{x}$ and $\boldsymbol{x}^*$ are column-vectors: $\boldsymbol{b} = (b_0, b_1, \ldots, b_q)'$, $\boldsymbol{x} = (x_1, \ldots, x_n)'$, and $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*)'$. The transition formulae define an eigenvector problem (see Appendix) that is akin to the eigenvector problem posed by canonical correlation analysis, λ in Eq. 3 being the eigenvalue. As in correspondence analysis, the equations have a trivial solution in which all site and species scores are equal and $\lambda = 1$; this trivial solution can either be disregarded or be excluded by requiring that the site scores are centered to zero mean, i.e., $\Sigma_i y_{i+} x_i = 0$.

Algorithm: reciprocal averaging and regression

The transition formulae can be solved by the following iteration algorithm of reciprocal averaging and multiple regression.

S1) Start with arbitrary, but unequal, initial site scores.

S2) Calculate species scores by weighted averaging of the site scores (Eq. 3 with $\lambda = 1$).

S3) Calculate new site scores by weighted averaging of the species scores (Eq. 4).

S4) Obtain regression coefficients by weighted multiple regression of the site scores on the environmental variables (Eq. 5). The weights are the site totals (y_{i+}).

S5) Calculate new site scores by Eq. 6 or, equivalently, Eq. 2. The new site scores are in fact the fitted values of the regression of the previous step.

S6) Center and standardize the site scores such that
$$\Sigma_i y_{i+} x_i = 0 \quad \text{and} \quad \Sigma_i y_{i+} x_i^2 = 1. \tag{7}$$

S7) Stop on convergence, i.e., when the new site scores are sufficiently close to the site scores of the previous iteration; otherwise go to S2.

This procedure is akin to the reciprocal averaging algorithm of correspondence analysis, but steps S4 and S5 are additional. The new technique is a correspondence analysis technique with restrictions (S4 and S5) on the site scores (cf. De Leeuw 1984). The final regression coefficients will be called canonical coefficients, and the multiple correlation coefficient of the final regression will be called the species–environment correlation. The species–environment correlation is a measure of how well the extracted variation in community composition can be explained by the environmental variables and is equal to the correlation between the site scores $\{x_i^*\}$, which are weighted mean species scores (calculated by Eq. 4), and the site scores $\{x_i\}$, which are a linear combination of the environmental variables (calculated by Eq. 2 or Eq. 6). This equality requires the assumption that sites are weighted proportional to y_{i+}, as in steps S4 and S6, and this weighting of sites is assumed in the calculation of means, variances, and correlations throughout the paper.

The standardization of the site scores in S6 is convenient in the algorithm, but it has more meaning ecologically to rescale the solution according to Eq. A.8 of the Appendix, as proposed by Hill (1979). Then, the tolerance of the fitted Gaussian response curves is (on average) about 1 unit, and a species' response curve can be expected to rise and decline over an interval of about 4 units.

More than one dimension and detrending

Second and additional axes can be extracted as in correspondence analysis by adding to the algorithm, after S5, a step that makes the trial site scores uncorrelated with the previous axes. The two-dimensional solution is intended to fit bivariate Gaussian response surfaces to the species data (Ter Braak 1985*b*) but often gives a bad fit because of the arch effect, an approximately quadratic dependence between the scores of the first two axes. This effect crops up whenever a short gradient is dominated by a long gradient (Gauch 1982*a*). The modifications of correspondence analysis that led to detrended correspondence analysis (Hill and Gauch 1980) can also be incorporated in canonical correspondence analysis; the rationale for detrending is the same. Detrending removes the arch effect and im-

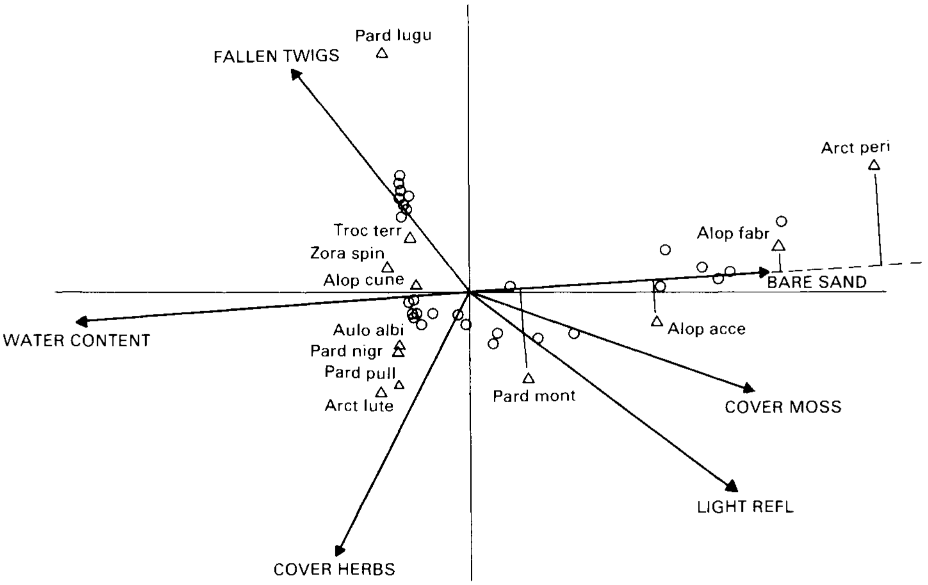

FIG. 1. The distribution of 12 species of hunting spiders caught in pitfall traps in a Dutch dune area. Canonical correspondence analysis (CCA) ordination diagram with pitfall traps (O), hunting spiders (△), and environmental variables (arrows); first axis is horizontal, second axis vertical. Shown also are the projections of the spider points labelled Arct peri, Alop fabr, Alop acce, and Pard mont onto the trajectory of the arrow of bare sand; the order of the projection points indicates the approximate ranking of the centers of the distributions of these spiders along the variable "percentage bare sand," *Arctosa perita* being found in habitats with the highest percentages of bare sand. The spider species are: Alop acce = *Alopecosa accentuata,* Alop cune = *Alopecosa cuneata,* Alop fabr = *Alopecosa fabrilis,* Arct lute = *Arctosa lutetiana,* Arct peri = *Arctosa perita,* Aulo albi = *Aulonia albimana,* Pard lugu = *Pardosa lugubris,* Pard mont = *Pardosa monticola,* Pard nigr = *Pardosa nigriceps,* Pard pull = *Pardosa pullata,* Troc terr = *Trochosa terricola,* Zora spin = *Zora spinimana.* The environmental variables are: Water Content = percentage of soil dry mass, Bare Sand = percentage cover of bare sand, Fallen Twigs = percentage cover of fallen leaves and twigs, Cover Moss = percentage cover of the moss layer, Cover Herbs = percentage cover of the herb layer, and Light Refl = reflection of the soil surface with cloudless sky.

proves the fit to the Gaussian model considerably in simulations where the true site and species scores are homogeneously distributed in a rectangle (the extension to two dimensions of conditions C3 and C4; Ter Braak 1985*b*). Detrending, however, also attempts to impose such a homogeneous distribution of scores on the data where none exists. The computer program CANOCO (Ter Braak 1985*a*) will also perform detrended canonical correspondence analysis. For a comparison of the detrended analysis with the non-detrended analysis, see Tests on Real Data.

Canonical coefficients and intraset correlations

For interpreting the ordination axes one can use the canonical coefficients and the intraset correlations. The canonical coefficients define the ordination axes as linear combinations of the environmental variables through Eq. 2, and the intraset correlations are the correlation coefficients between the environmental variables and these ordination axes. (The term intraset is used here to distinguish these correlations from the interset correlations between the environmental variables and the site scores $\{x_i^*\}$ that are derived from the species data.) For the rest of the analysis it is assumed that the environmental variables have been standardized to zero mean and unit variance prior to the analysis. This standardization removes arbitrariness in the units of measurement of the environmental variables and makes the canonical coefficients comparable to each other, but does not influence other aspects of the analysis.

By looking at the signs and relative magnitudes of the intraset correlations and of the canonical coefficients so standardized, we may infer the relative importance of each environmental variable for predicting the community composition. The canonical coefficients give the same information as the intraset correlations in the special case that the environmental variables are mutually uncorrelated, but may provide rather different information when the environmental variables are correlated with each other, as they usually are in field data. Both a canonical coefficient and an intraset correlation coefficient relate to the rate of change in community composition per unit change in the corresponding environmental variable, but in the former case it is assumed that other environmental variables are being held constant, whereas in the latter case the other environmental variables are assumed to covary with that one environmental variable in the particular way they do in the data set. When the environmental variables are strongly correlated with each other—for example, simply because the number of environmental variables approaches the number of sites—the effects of different environmental variables on community composition cannot be separated out and, consequently, the canonical coefficients are unstable. This is the multicollinearity problem, well known to occur in multiple regression analysis (see Montgomery and Peck 1982). When this problem arises (the program CANOCO [Ter Braak 1985*a*] provides statistics to help detect it) one should abstain from attempts to interpret the canonical coefficients. Fortunately, the intraset correlations do not suffer from this problem and can still be used for interpretation purposes. One can also remove environmental variables from the analysis, keeping at least one variable per set of strongly correlated environmental variables; the eigenvalues and species–environment correlations will usually decrease only slightly. If the eigenvalues and species–environment correlations drop considerably, one has removed too many (or the wrong) variables.

In contrast to canonical correlation analysis, canonical correspondence analysis is not hampered by multicollinearity in the species data; the number of species is therefore allowed to exceed the number of sites.

Ordination diagram

The solution of canonical correspondence analysis can be displayed in an ordination diagram with sites and species represented by points, and environmental variables represented by arrows (see Fig. 1). The species and site points jointly represent the dominant patterns in community composition insofar as these can be explained by the environmental variables, and the species points and the arrows of the environmental variables jointly reflect the species' distributions along each of the environmental variables. For example, when an arrow refers to "water content," the diagram allows us to infer—by rules explained in the following paragraphs—which species largely occur in the wettest sites, which in the driest sites, and which in sites with intermediate moisture values. We shall limit the discussion to two-dimensional diagrams because these are the most convenient to visualize. The rules for construction and interpretation of higher-dimensional ordination diagrams are the same.

For the diagram to represent the approximate community composition at the sites, we must plot species scores and site scores that are weighted mean species scores, as in Hill's (1979) program DECORANA. Because each site point then lies at the centroid of the species points that occur at that site, one may infer from the diagram which species are likely to be present at a particular site. Also, insofar as canonical correspondence analysis is a good approximation to the fitting of Gaussian response surfaces, the species points are approximately the optima of these surfaces; hence the abundance or probability of occurrence of a species decreases with distance from its location in the diagram.

At which values of an environmental variable a species occurred in the data can conveniently be summarized by the weighted average. The weighted average of a species distribution (k) with respect to an environmental variable (j) is defined as the average of the values of that environmental variable at those sites at which that species occurs, the weighting of each site being proportional to species abundance, i.e.,

$$\bar{z}_{kj} = \sum_{i=1}^{n} y_{ik} z_{ij} / y_{+k}. \tag{8}$$

The weighted average indicates the "center" of a species' distribution along an environmental variable (Ter Braak and Looman 1986), and differences in weighted averages between species indicate differences in their distributions along that environmental variable. The ordination diagram of canonical correspondence analysis can be supplemented by arrows for the environmental variables to give a graphical summary of the weighted averages of all species with respect to all environmental variables.

The arrows for the environmental variables must be added in the following way. The position of the head of the arrow for an environmental variable depends on the eigenvalues of the axes and the intraset correlations of that environmental variable with the axes (see Appendix). The coordinate of the head of the arrow on axis s must be $[\lambda_s(1 - \lambda_s)]^{1/2}$ times the intraset correlation of the environmental variable with axis s, where λ_s is the eigenvalue of axis s and it is assumed that the species scores are standardized according to Appendix Eq. A.8, as before. By connecting the origin of the plot (the centroid of the site points) with each of the arrowheads, we obtain the arrows representing the variables (Fig. 1). How to construct such a diagram from a detrended canonical correspondence analysis is described in the Appendix. Only the directions and relative lengths convey information, so one can increase or reduce the lengths of all arrows to fit conveniently in the ordination diagram.

The ordination diagram so constructed allows the following interpretation. Each arrow determines a direction or axis in the diagram, obtained by extending the arrow in both directions (in your mind or on paper). From each species point we must drop a perpendicular to this axis. Fig. 1 shows an example. The arrow for water content has been extended (the axis happens to coincide with the arrow for bare sand) and perpendiculars have been dropped to this axis from four species points. The endpoints indicate the relative positions of the centers of the species distributions along the water content axis or, more precisely, they indicate in an approximate way the relative value of the weighted average of each species with respect to water content. From Fig. 1 we thus infer that *Arctosa perita* has the lowest weighted average with respect to water content (i.e., it largely occurs at the driest sites), *Alopecosa fabrilis* the second lowest value, and so on to *Arctosa lutetiana*, which is inferred to have the highest weight-

TABLE 1. Comparison of the results of ordinations by detrended correspondence analysis (DCA), canonical correspondence analysis (CCA), and detrended canonical correspondence analysis (DCCA) of hunting spider data (see Fig. 1): eigenvalues and species-environment correlation coefficients for the first three axes.

	Axis		
	1	2	3
		Eigenvalues	
DCA	0.58	0.16	0.02
CCA	0.53	0.21	0.06
DCCA	0.53	0.13	0.02
		Correlation coefficients	
DCA	0.96	0.92	0.88
CCA	0.96	0.93	0.64
DCCA	0.97	0.94	0.90

ed average (i.e., to occur largely at the wettest sites). In general, the approximate ranking of the weighted averages for a particular environmental variable can be seen easily from the order of the endpoints of the perpendiculars of the species along the axis for that variable. Further, the weighted averages are approximated in the diagram as deviations from the grand mean of each environmental variable, the grand mean being represented by the origin of the plot. A second useful rule for interpreting the diagram is therefore that the inferred weighted average is higher than average if the endpoint of a species lies on the same side of the origin as the head of an arrow does, and is lower than average if the origin lies between the endpoint and the head of the arrow.

These rules for interpreting the joint plot of species points and environmental arrows are identical to the rules for interpreting a biplot (Gabriel 1971). Biplots have been used so far primarily in connection with principal components analysis (Ter Braak 1983), but a biplot is essentially just a joint plot of two kinds of entities that allows a particular kind of quantitative interpretation (Gabriel 1981, Ter Braak 1983). The joint plot of species and environmental variables is, in fact, a biplot. This biplot provides a weighted least squares approximation of the weighted averages of the species with respect to the environmental variables (see Appendix). The measure of goodness of fit, $100 \times (\lambda_1 + \lambda_2)/(\text{sum of all eigenvalues})$, expresses the percentage variance of the weighted averages accounted for by the two-dimensional diagram. In interpreting percentages of variance accounted for, it must be kept in mind that the goal is not 100%, because part of the total variance is due to noise in the data (cf. Gauch 1982*b*). Even an ordination diagram that explains only a low percentage may be quite informative.

Finally, the length of an arrow representing an environmental variable is equal to the rate of change in the weighted average as inferred from the biplot, and is therefore a measure of how much the species distributions differ along that environmental variable. Important environmental variables therefore tend to be represented by longer arrows than less important environmental variables.

Relation of canonical correspondence analysis with weighted averaging ordination and discriminant analysis

Canonical correspondence analysis generalizes two existing techniques for direct gradient analysis. When a single quantitative environmental variable is considered, it reduces to weighted averaging ordination (Gauch 1982*a*), because x_i in Eq. 1 is then simply the value of this variable at site i, and fitting this model simplifies under condition C4 to weighted averaging (cf. Ter Braak and Looman 1986). With two quantitative environmental variables, the technique represents the same information in a two-dimensional diagram as weighted averaging ordination with respect to these variables, although the variables are not necessarily displayed as orthogonal directions in the ordination diagram. With a single nominal environmental variable, canonical correspondence analysis is a variant of discriminant analysis (canonical variate analysis) that is appropriate to a unimodal response model, and which can be obtained more simply from a correspondence analysis of a two-way table of species by (classes of) the nominal variable (Greenacre 1984: section 7.1). The cells of the table must contain the total abundances of each of the species in each of the classes. In the resulting ordination diagram the classes are represented by points. This equivalence suggests that it can be more natural to represent nominal environmental variables by points instead of arrows. The point for a class of a nominal environmental variable must be located at the centroid (the weighted average) of the sites belonging to that class. Classes consisting of sites with high values for a species will then tend to lie close to that species' point. Gasse and Tekaia (1983) applied this technique to establish a transfer function for estimating paleo-environmental conditions from diatom assemblages.

TABLE 2. Hunting spider abundance data from Fig. 1: canonical coefficients and the intraset correlations of environmental variables with the first two axes of canonical correspondence analysis (CCA). The environmental variables were standardized to unit variance after log-transformation. For a description of variables, see Fig. 1 legend.

Axis	Canonical coefficients		Correlation coefficients	
variable	1	2	1	2
Water Content	−0.51	−0.41	−0.93	−0.08
Bare Sand	0.33	−0.10	0.73	0.06
Fallen Twigs	−0.14	0.37	−0.43	0.78
Cover Moss	0.05	−0.27	0.69	−0.30
Cover Herbs	−0.28	−0.15	−0.32	−0.78
Light Refl	0.27	−0.03	0.64	−0.59

TABLE 3. Hunting spider abundance data, with species (rows) and sites (columns) arranged in order of the scores for the first axis of canonical correspondence analysis (CCA). Site numbers correspond to those of Van der Aart and Smeenk-Enserink (1975: Table 4). The species abundance data have been transformed by taking square roots; the integer part is shown, a blank denoting absence of the species and 9 denoting >80 individuals captured. For this table, the range of each environmental variable was divided into 10 equal-sized classes (denoted by 0–9) after the data were transformed. Abbreviations and a description of the biological system are given in legend of Fig. 1.

	Site numbers																											
	15	19	20	16	17	18	2	8	21	5	6	14	4	7	13	3	1	9	12	25	11	10	28	23	22	27	24	26
Species																												
Arct lute										1	2	1	1	3	1	1												
Pard lugu	2	3	3	2	1	2	1	7	4	1		1	1	1	1	1				1				1				
Zora spin	1	1	1	2	1		3	1	1	4	5	5	5	4	4	1	2			2								
Pard nigr		1		1			3	1		9	5	3	5	9	7	4	3	1	1	2								
Pard pull							6	1	1	8	4	8	9	9	8	6	6	1	2		1							
Aulo albi							5	2		3	2	2	4	4	4	3	2			1	1							
Troc terr	5	4	4	5	4	5	8	5	4	9	7	9	9	9	9	8	7	1	3	4	2	1	1	1	1			1
Alop cune		1	1	1		1	1	3	1	4	2	1	2	2	6	4	3	1	3	1	1							
Pard mont							1	1	1	1	3	3	2	5	4	5	7	5	9	3	9	4	2	2	1	1	1	
Alop acce										1			1	1	1	3	5	1	4	3	3	1	3	4	2	5	3	1
Alop fabr															1	1				3	1	1	3	3	4	3	4	2
Arct peri																							1	2	1	2	2	4
Environmental variable																												
Water Content	9	7	8	8	9	8	8	6	7	8	9	8	6	8	9	6	5	5	5	3	4	4	0	0	1	0	2	0
Bare Sand	0	0	0	0	0	0	0	0	0	0	5	0	0	0	3	0	0	0	0	7	0	8	7	6	7	5	7	9
Cover Moss	1	3	1	1	1	0	2	2	1	0	5	4	5	1	1	5	7	9	8	2	9	7	8	9	9	8	9	4
Light Refl	1	0	0	0	2	2	3	1	0	5	1	2	6	5	7	8	8	7	8	5	8	8	8	9	8	8	9	9
Fallen Twigs	9	9	9	9	9	9	3	9	9	0	7	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0
Cover Herbs	5	2	0	0	5	5	9	6	2	9	6	9	9	9	9	9	9	6	8	8	7	5	6	6	0	6	5	2

TESTS ON REAL DATA

Hunting spider data

The first data set, taken from Van der Aart and Smeenk-Enserink (1975), concerns the distributions of 12 species of hunting spiders (Fig. 1) in a Dutch dune area, in relation to environmental data. The species data are the numbers of individuals of each species caught in pitfall traps over a period of 60 wk. Twenty-six environmental variables were measured at 28 of the pitfall traps. This number of variables is too large to sort out their independent effects on community composition. Eighteen variables were removed on a priori grounds, and two more variables were removed because they were strongly correlated with one of the remaining six variables (Fig. 1). The species data were transformed by taking square roots to down-weight high abundances; the environmental data were transformed by taking logarithms, as in the original paper.

The ordinations by detrended correspondence analysis (DCA), canonical correspondence analysis (CCA), and detrended canonical correspondence analysis (DCCA) are very similar for these data. The first eigenvalue of CCA is only slightly lower than the first eigenvalue of DCA, and the species–environment correlations of the first three axes are all high (Table 1). Apparently the measured environmental variables are sufficient to explain the major variation among the spider catches. From Table 2 we infer that the first axis is a moisture gradient, on which the drier sites have a high percentage of bare sand or of moss. The correlations of the second axis show a contrast between sites with a high cover of leaves and twigs and sites with a well-developed herb and moss layer.

From the species and site points in the CCA ordination diagram (Fig. 1) we infer, for example, that *Arctosa perita* and *Alopecosa fabrilis* reached their maximum abundance in the six pitfall traps represented on the right-hand side of the diagram, that *Pardosa monticola* had maximum abundance in the pitfall traps shown in the middle, and that *Pardosa lugubris* was most abundant in the cluster of pitfall traps represented in the top-left of the diagram. These inferences from the diagram largely agree with the data (cf. Table 3).

The arrows for environmental variables in Fig. 1 account, in conjunction with the species points, for 87% of the variance in the weighted averages of the 12 spiders with respect to the six environmental variables, the sum of all eigenvalues being 0.85. For example, projecting the spider points on the axis of percentage bare sand shows that *Arctosa perita* and *Alopecosa fabrilis* were mainly found in habitats with the highest percentages of bare sand, *Alopecosa accentuata* and *Pardosa monticola* in habitats with intermediate bare sand percentages, and the species on the left-hand side of the diagram in habitats with the lowest percentages of bare sand. For *Ar. perita, Al. fabrilis, Al. accentuata,* and *P. monticola,* the same ranking applies with respect to the cover of the moss layer. The ranking is more or less the reverse with respect to soil water content. *Arctosa lutetiana, Pardosa pullata, Pardosa nigriceps, Aulonia albimana,* and *Pardosa monticola* occurred in

TABLE 4. Comparison of the results of ordinations by detrended correspondence analysis (DCA), canonical correspondence analysis (CCA), and detrended canonical correspondence analysis (DCCA) of dyke vegetation data (see Fig. 2): eigenvalues and species-environment correlation coefficients for the first four axes.

	Axis			
	1	2	3	4
		Eigenvalues		
DCA	0.34	0.25	0.22	0.19
CCA	0.20	0.13	0.12	0.07
DCCA	0.20	0.12	0.09	0.05
		Correlation coefficients		
DCA	0.52	0.40	0.58	0.22
CCA	0.82	0.81	0.80	0.77
DCCA	0.83	0.81	0.76	0.66

habitats with a well-developed herb layer. *Pardosa lugubris* occupies an aberrant position in the diagram, being the single spider species that occurred mainly in habitats with a high cover of fallen leaves and twigs (i.e., in woods). *Trochosa terricola, Zora spinimana,* and *Alopecosa cuneata* occupy an intermediate position between the woody and grassier sites. Van der Aart and Smeenk-Enserink (1975) gave a similar description, but the CCA ordination diagram tells the main story at a glance. The DCCA ordination diagram provided essentially the same information. The main structure in the data is also clear from Table 3, where species and sites are reordered according to their scores on the first CCA axis. The species data show a diagonal band; soil water content decreases along the first axis, whereas percentage bare sand, cover of moss, and light reflection increase along this axis.

Dyke vegetation

De Lange (1972) studied the occurrences of macrophytes in dykes in the Netherlands in relation to electrical conductivity, phosphate and chloride concentration in the water, and soil type (clay, peaty soil, sand). A total of 125 fresh water dykes (conductivity <126 mS/m) were selected, with in total 133 plant species. Conductivity data were transformed by taking logarithms, because of a skewed distribution, and chloride concentration was transformed to chloride ratio (the share of chloride ions in the electrical conductivity; G. Van Wirdum, *personal communication*). The nominal variable "soil type" (with three classes) was dealt with, as in multiple regression (see Montgomery and Peck 1982: chapter 6), by defining two dummy environmental variables "peat" and "sand." (The variable "peat" takes the value 1 when a dyke has soil type "peat" and the value 0 otherwise. The variable "sand" is defined analogously. A dyke in clay thus scores the value 0 on each of the two variables. The canonical coefficient of "peat" then measures the difference in expected site scores between peaty and clay soils. Other choices of dummy variables could have been used equivalently, e.g., "clay" and "sand.")

Table 4 shows that the environmental variables are poorly related to the first four species axes of DCA. But by choosing the axes in the light of the environmental variables, by applying CCA or DCCA, the species–environment correlations increase considerably. The interpretation of the axes is unambiguous (Table 5): the first axis is defined by conductivity and phosphate, the second by the chloride ratio and soil type; the soil types further differentiate on the third and fourth axes. CCA and DCCA do not differ much for this data set. On the CCA ordination diagram (Fig. 2) the dykes are not displayed because the diagram would have been too crowded; the undisplayed dykes all lie in the open center region of Fig. 2. Fig. 2 accounts for 56% of the variance and shows that the weighted averages of the species with respect to conductivity and phosphate result in similar rankings; this similarity cannot be explained by the correlation between these variables in the data set, because this correlation is only 0.44. In contrast, the ranking with respect to chloride ratio is different. The soil types are also represented by arrows (Fig. 2). Species whose distribution is the most restricted to peaty soils lie somewhat to the top-left-hand corner of the diagram. Analogously, species with a distribution mainly on clay tend to lie somewhat to the bottom-right-hand corner of the diagram.

The eigenvalues (Table 4) show that the extracted gradients are quite short (cf. Gauch and Stone 1979). The scores (optima) of most species therefore lie outside the center region where the sites lie, and the probability of occurrence of such species simply increases

TABLE 5. Dyke vegetation data from Fig. 2: canonical coefficients and intraset correlations, as in Table 2. For a description of variables see Fig. 2 legend.

Axis	Canonical coefficients				Correlation coefficients			
variable	1	2	3	4	1	2	3	4
EC	0.27	0.03	−0.02	0.10	0.83	0.17	−0.25	0.20
Phosphate	0.30	0.01	0.16	−0.15	0.86	−0.08	0.30	−0.21
Chloride Ratio	0.01	0.30	−0.09	0.09	0.14	0.86	−0.30	0.29
Clay	0	0	0	0	0.27	−0.21	−0.89	−0.31
Peat*	−0.09	0.44	0.78	−0.03	−0.38	0.49	0.72	−0.17
Sand*	0.01	−0.30	0.58	0.99	0.13	−0.40	0.40	0.78

* Not standardized to unit variance.

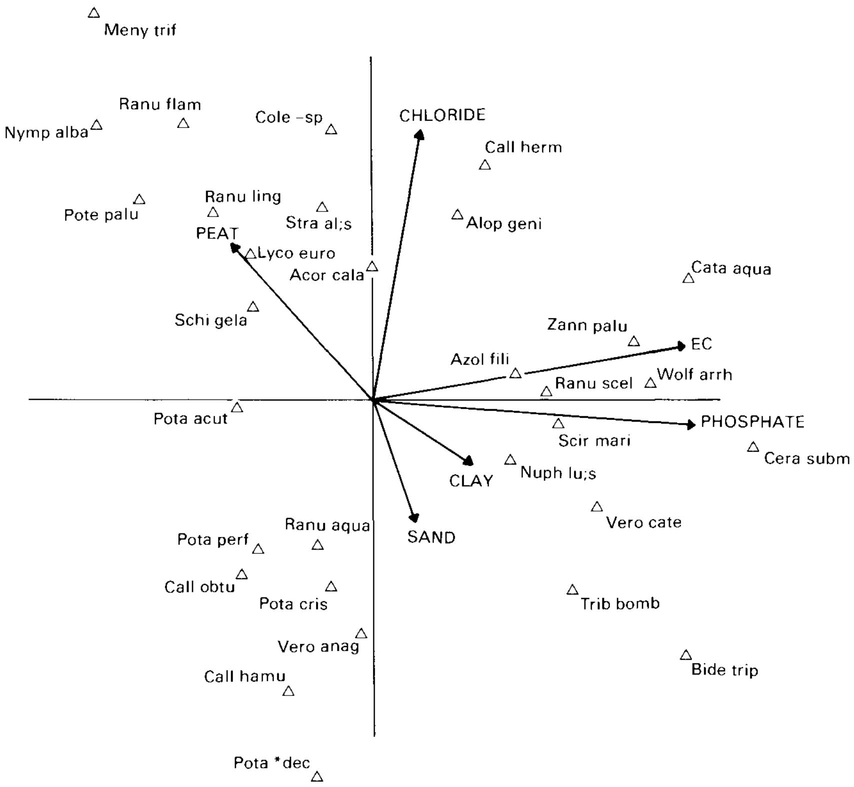

FIG. 2. Dyke vegetation data: CCA ordination diagram with plant species (△) and environmental variables (arrows); first axis is horizontal, second axis vertical. Species with positions near the center and some other species elsewhere are not shown because the diagram would have become too crowded. The plant species shown are: Acor cala = *Acorus calamus*, Alop geni = *Alopecurus geniculatus*, Azol fili = *Azolla filiculoides*, Bide trip = *Bidens tripartita*, Call hamu = *Callitriche hamulata*, Call herm = *Callitriche hermophroditica*, Call obtu = *Callitriche obtusangula*, Cata aqua = *Catabrosa aquatica*, Cera subm = *Ceratophyllum submersum*, Cole -sp = *Coleochaete* sp., Lyco euro = *Lycopus europaeus*, Meny trif = *Menyanthes trifoliata*, Nuph lu;s = *Nuphar lutea* (submerged form), Nymp alba = *Nymphaea alba*, Pota acut = *Potamogeton acutifolius*, Pota cris = *Potamogeton crispus*, Pota *dec = *Potamogeton decipiens*, Pota perf = *Potamogeton perfoliatus*, Pote palu = *Potentilla palustris*, Ranu aqua = *Ranunculus aquatilis s.l.*, Ranu flam = *Ranunculus flammula*, Ranu ling = *Ranunculus lingua*, Ranu scel = *Ranunculus sceleratus*, Schi gela = *Schizochlamys gelatinosa*, Scir mari = *Scirpus maritimus*, Stra al;s = *Stratiotes aloides* (submerged form), Trib bomb = *Tribonema bombycinum*, Vero anag = *Veronica anagallis-aquatica*, Vero cate = *Veronica catenata*, Wolf arrh = *Wolffia arrhiza*, Zann palu = *Zannichellia palustris*. The environmental variables are: EC = electrical conductivity, Phosphate = orthophosphate concentration, Chloride ratio = share of chloride ions in the electrical conductivity, and Clay, Peat, Sand (=type of soil surrounding the dyke).

or decreases monotonically along the gradients actually sampled, instead of being unimodal as required (see Theory). Condition C4 is clearly violated in this data set; nevertheless CCA worked well.

Algae along a pollution gradient

Fricke and Steubing (1984) sampled 25 sites in rivulets near the Ederstausee (Western Germany), recorded the abundances of 34 algae on a scale from 0 to 5, and measured seven environmental variables (Fig. 3), six of which (all but °D) were transformed by taking logarithms in our analysis because of skewed distributions. The first axis of DCA and that of CCA nearly coincided (Table 6), being a clear pollution gradient: positive correlations with ammonium, phosphate, biological oxygen demand (BOD5), and electrical conductivity, and a negative correlation with oxygen (Table 7). Although the ordination diagram of CCA (Fig.

TABLE 6. Comparison of the results of ordinations by detrended correspondence analysis (DCA), canonical correspondence analysis (CCA), and detrended canonical correspondence analysis (DCCA) of data on algae along a pollution gradient: eigenvalues and species–environment correlation coefficients for the first three axes.

	Axis		
	1	2	3
		Eigenvalues	
DCA	0.70	0.17	0.09
CCA	0.67	0.14	0.10
DCCA	0.67	0.08	0.05
		Correlation coefficients	
DCA	0.97	0.50	0.67
CCA	0.98	0.72	0.89
DCCA	0.98	0.80	0.79

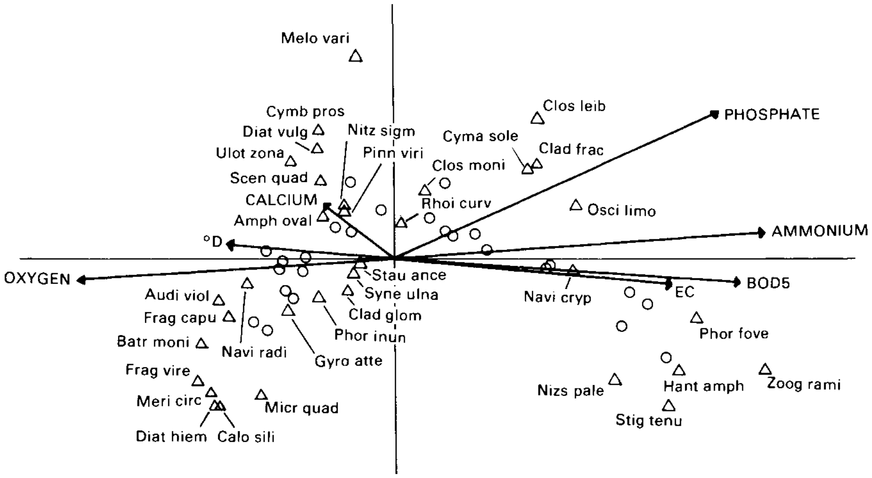

FIG. 3. Algae along a pollution gradient: CCA ordination diagram with algae (△), sites (○), and environmental variables (arrows); first axis is horizontal, second axis vertical. The algae are: Amph oval = *Amphora ovalis*, Audi viol = *Audionella violacea*, Batr moni = *Batrachospermum moniliforme*, Calo sili = *Caloneis silicula*, Clad frac = *Cladophora fracta*, Clad glom = *Cladophora glomerata*, Clos moni = *Closterium moniliferum*, Clos leib = *Closterium leibneinii*, Cyma sole = *Cymatopleura solea*, Cymb pros = *Cymbella prostata*, Diat hiem = *Diatoma hiemale mesodon*, Diat vulg = *Diatoma vulgare*, Frag capu = *Fragilaria capucina*, Frag vire = *Fragilaria virescens*, Gyro atte = *Gyrosigma attenuatum*, Hant amph = *Hantzschia amphioxis*, Melo vari = *Melosira varians*, Meri circ = *Meridion circulare*, Micr quad = *Microspora quadrata*, Navi cryp = *Navicula cryptocephala*, Navi radi = *Navicula radiosa*, Nizs pale = *Nizschia palea*, Nitz sigm = *Nitzschia sigmoidea*, Osci limo = *Oscillatoria limosa*, Phor fove = *Phormidium foveolarum*, Phor inun = *Phormidium inundatum*, Pinn viri = *Pinnularia viridis*, Rhoi curv = *Rhoicophenia curvata*, Scen quad = *Scenedesmus quadricauda*, Stau ance = *Stauroneis anceps*, Stig tenu = *Stigeoclonium tenue*, Syne ulna = *Synedra ulna*, Ulot zona = *Ulotrix zonata*, Zoog rami = *Zoogloea ramigera*. The environmental variables are: Oxygen = oxygen concentration, BOD5 = biological oxygen demand, Ammonium = ammonium concentration, Phosphate = orthophosphate concentration, Calcium = calcium concentration, °D = German standard measure for the total concentration of calcium and magnesium, and EC = electrical conductivity.

3) explains most of the variance (73%), the diagram is unsatisfactory because of the arch effect (Gauch 1982*a*). The detrending in DCCA largely removes this effect (Fig. 4) and shows that the variation in species composition on the second axis is small ($\lambda_2 = 0.08$). This variation has surprisingly high correlation with the environmental variables (Table 6). The canonical coefficients of the second axis (Table 8) suggest that this minor component of the variation is related to the ratio of ammonium to phosphate.

In this example the interpretations of the CCA diagram and the DCCA diagram (Figs. 3 and 4) are not very different, but in more complicated data sets the difference can be large. As in regular ordination, detrending is a method to prevent the second axis from being obscured by dependence on the first.

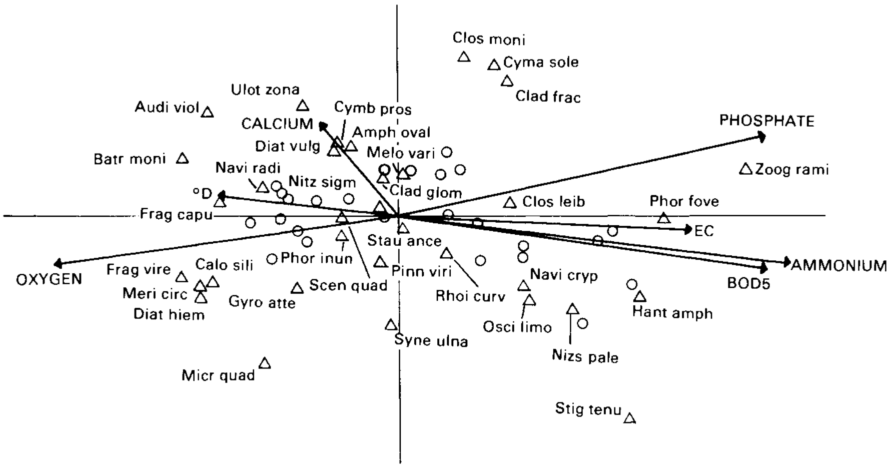

FIG. 4. Algae along a pollution gradient: DCCA ordination diagram. For an explanation of symbols see Fig. 3 legend.

TABLE 7. Data on algae along a pollution gradient, from Fig. 3: canonical coefficients and intraset correlations, as in Table 2. For a description of variables see Fig. 3 legend.

Axis	Canonical coefficients		Correlation coefficients	
variable	1	2	1	2
Oxygen	−0.47	0.20	−0.81	−0.06
BOD5	0.06	−0.11	0.88	−0.08
Ammonium	0.80	−0.07	0.94	0.09
Phosphate	−0.04	0.64	0.83	0.51
Calcium	−0.25	0.28	−0.19	0.19
°D	−0.07	−0.10	−0.44	0.05
EC	0.28	−0.27	0.71	−0.09

DISCUSSION

Canonical correspondence analysis provides an integrated description of species–environment relationships by assuming a response model that is common to all species, and the existence of a single set of underlying environmental gradients to which all the species respond. The same strong assumption is implicit in all ordination techniques. Canonical correspondence analysis has the advantage over other techniques in that it focuses on the relations between species and measured environmental variables and so provides an automated interpretation of the ordination axes.

Canonical correspondence analysis derives theoretical strength from its relation to maximum likelihood Gaussian canonical ordination under conditions C1–C4 and furthermore seems extremely robust in practice when these assumptions do not hold. The vital assumption is that the response surfaces of the species are unimodal, the Gaussian (bell-shaped) response model being the example for which the method's performance is particularly good. For the simpler case where species–environment relationships are monotone, the results can still be expected to be adequate in a qualitative sense (see Tests on Real Data: Dyke Vegetation). The method would not work if a large number of species were distributed in a more complex way, e.g., bimodally; the restriction to a unimodal model is necessary for practical solubility, but as Hill (1977) points out, a good choice of environmental variable should minimize the number of species with more complex distributions. Some care, however, is required with the interpretation of the ordination diagram when the additional assumptions (C1–C4) do not hold. Species in the center of the ordination diagram may then have their optima there, but may alternatively be unrelated to the axes. Which possibility is most likely can be decided upon by tabular rearrangement of the species data with respect to each axis, as is done in Table 3 for the first axis. Further work still needs to be done on the statistical significance of eigenvalues, species–environment correlations, and canonical coefficients.

As in correspondence analysis, any kind of transformation of the species abundance data may influence the results. When the abundance data have a very skewed distribution, it is recommended to transform them by taking square roots or logarithms. In this way we prevent a few high abundance values from unduly influencing the analysis. Because the compound environmental gradients constructed by canonical correspondence analysis are required to be linear combinations of environmental variables, nonlinear transformation of environmental variables can also be considered if there is some reason to do so. Prior knowledge about the possible impact of the environmental variables on community composition may suggest particular nonlinear transformations and particular nonlinear combinations, i.e., environmental scalars in the sense of Loucks (1962) and Austin et al. (1984). The use of environmental scalars can also circumvent the multicollinearity problem described in Theory: Canonical Coefficients. In contrast to the ordination techniques in common use, canonical correspondence analysis allows one to incorporate existing knowledge about species–environment relationships into the analysis and thus potentially is a more powerful tool to advance this knowledge.

Canonical correspondence analysis can be used fruitfully in combination with (detrended) correspondence analysis, as in the examples described. When the solutions do not differ much, we infer that the measured environmental variables can account for the main variation in the species data. When the solutions do differ, we infer either that the environmental variables account for less conspicuous directions of variation in the species data (when the correlations between species and environment axes are high) or that they cannot account for any of the variation (when the correlations are small). These possibilities considerably extend the analytical power of ordination by allowing comparison of results from indirect and direct gradient analysis techniques that have a common theoretical basis. Direct and indirect gradient analysis can also be combined in a single analysis to answer such questions as "Does the known environmental variation account for all the community variation, or is there a substantial residual variation?" Suppose we believe two environmental variables govern the species composition in a

TABLE 8. Data on algae along a pollution gradient, from Fig. 3: canonical coefficients and intraset correlations in DCCA. For a description of variables see Fig. 3 legend.

Axis	Canonical coefficients		Correlation coefficients	
variable	1	2	1	2
Oxygen	−0.37	0.05	−0.81	0.04
BOD5	0.07	0.21	0.88	−0.40
Ammonium	0.65	−0.60	0.95	−0.47
Phosphate	0.10	0.50	0.86	0.06
Calcium	−0.22	0.23	−0.19	0.37
°D	−0.06	−0.07	−0.43	0.18
EC	0.22	−0.17	0.70	−0.22

region. We may choose two ordination axes in the light of these variables, then extract further axes as in detrended correspondence analysis by reciprocal averaging and detrending with respect to all previous axes. The lengths of the extra axes measure the residual variation. The program CANOCO (Ter Braak 1985*a*) has an option to do such combined analyses. The same option allows analysis of nested data (subplots within plots, e.g., yearly vegetation records from several permanent plots, or bird records from woodlots in several regions). The first axes can be chosen to represent variation between plots, so that the further axes represent variation between subplots. Swaine and Greig-Smith (1980) used a variant of principal components analysis in this way to obtain an ordination of within-plot vegetation change in permanent plots; canonical correspondence analysis could be used for the same purpose but is not hampered by the unwarranted assumption of a linear relationship between species abundance and environment.

Acknowledgments

I would like to thank Dr. I. C. Prentice for inspiring discussions. The comments of the reviewers helped to clarify a number of practical aspects of the technique.

Literature Cited

Austin, M. P., R. B. Cunningham, and P. M. Fleming. 1984. New approaches to direct gradient analysis using environmental scalars and statistical curve-fitting procedures. Vegetatio **55**:11–27.

Bartlein, P. J., I. C. Prentice, and T. Webb, III. 1986. Climatic response surfaces from pollen data for some eastern North American taxa. Journal of Biogeography **13**:35–57.

Carleton, T. J. 1984. Residual ordination analysis: a method for exploring vegetation-environment relationships. Ecology **65**:469–477.

De Lange, L. 1972. An ecological study of ditch vegetation in the Netherlands. Dissertation. University of Amsterdam, Amsterdam, The Netherlands.

De Leeuw, J. 1984. The GIFI system of nonlinear multivariate analysis. Pages 415–424 *in* E. Diday, M. Jambu, L. Lebart, J. Pagès, and R. Tomassone, editors. Data analysis and informatics III. North-Holland Publishing, Amsterdam, The Netherlands.

Fricke, G., and L. Steubing. 1984. Die Verbreitung von Makrophyten und Mikrophyten in Hartwasser-Zuflüsse des Ederstausees. Archiv für Hydrobiologie **101**:361–372.

Gabriel, K. R. 1971. The biplot graphic display of matrices with application to principal component analysis. Biometrika **58**:453–467.

———. 1981. Biplot display of multivariate matrices for inspection of data and diagnosis. Pages 147–173 *in* V. Barnett, editor. Interpreting multivariate data. J. Wiley and Sons, New York, New York, USA.

Gasse, F., and F. Tekaia. 1983. Transfer functions for estimating paleoecological conditions (pH) from East African diatoms. Hydrobiologia **103**:85–90.

Gauch, H. G. 1982*a*. Multivariate analysis in community ecology. Cambridge University Press, Cambridge, England.

———. 1982*b*. Noise reduction by eigenvector ordinations. Ecology **63**:1643–1649.

Gauch, H. G., G. B. Chase, and R. H. Whittaker. 1974. Ordination of vegetation samples by Gaussian species distributions. Ecology **55**:1382–1390.

Gauch, H. G., and E. L. Stone. 1979. Vegetation and soil pattern in a mesophytic forest at Ithaca, New York. American Midland Naturalist **102**:332–345.

Gauch, H. G., and T. R. Wentworth. 1976. Canonical correlation analysis as an ordination technique. Vegetatio **33**: 17–22

Gittins, R. 1985. Canonical analysis. A review with applications in ecology. Springer-Verlag, Berlin, Germany.

Greenacre, M. J. 1984. Theory and applications of correspondence analysis. Academic Press, London, England.

Hill, M. O. 1977. Use of simple discriminant functions to classify quantitative phytosociological data. Pages 597–613 *in* E. Diday, L. Lebart, J. P. Pagès, and R. Tomassone, editors. Data analysis and informatics, I. Institut de Recherche d'Informatique et d'Automatique, Le Chesnay Cedex, France.

———. 1979. DECORANA: A FORTRAN program for detrended correspondence analysis and reciprocal averaging. Section of Ecology and Systematics, Cornell University, Ithaca, New York, USA.

Hill, M. O., and H. G. Gauch. 1980. Detrended correspondence analysis, an improved ordination technique. Vegetatio **42**:47–58.

Loucks, O. L. 1962. Ordinating forest communities by means of environmental scalars and phytosociological indices. Ecological Monographs **32**:137–166.

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. Multivariate analysis. Academic Press, London, England.

Montgomery, D. C., and E. A. Peck. 1982. Introduction to linear regession analysis. J. Wiley and Sons, New York, New York, USA.

Swaine, M. D., and P. Greig-Smith. 1980. An application of principal components analysis to vegetation change in permanent plots. Journal of Ecology **68**:33–41.

Ter Braak, C. J. F. 1983. Principal components biplots and alpha and beta diversity. Ecology **64**:454–462.

———. 1985*a*. CANOCO: A FORTRAN program for canonical correspondence analysis and detrended correspondence analysis. IWIS-TNO, Wageningen, The Netherlands.

———. 1985*b*. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics **41**:859–873.

Ter Braak, C. J. F., and C. W. N. Looman. In press 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio **65**:3–11.

Van der Aart, P. J. M., and N. Smeek-Enserink. 1975. Correlations between distributions of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. Netherlands Journal of Zoology **25**:1–45.

Whittaker, R. H. 1967. Gradient analysis of vegetation. Biological Reviews of the Cambridge Philosophical Society **42**:207–264.

Whittaker, R. H., S. A. Levin, and R. B. Root. 1973. Niche, habitat and ecotope. American Naturalist **107**:321–338.

APPENDIX

Here canonical correspondence analysis is shown to be (1) an approximation to Gaussian canonical ordination, (2) an eigenvector technique akin to canonical correlation analysis, and (3) a method for weighted least squares approximation of weighted averages of species with respect to environmental variables. For an explanation of the notation, see Theory.

The model of Gaussian canonical ordination is Eq. 1 in conjunction with Eq. 2 (see Theory). It is assumed that the species data are Poisson-distributed counts with $E(y_{ik}) = \mu_{ik}$ and that the species tolerances are all equal to 1. Then the maximum likelihood equations for u_k and b_j are, after some rearrangement, respectively:

$$u_k = \sum_i y_{ik}x_i/y_{+k} - \left[\sum_i (x_i - u_k)\mu_{ik}/y_{+k}\right] \tag{A.1}$$

$$\sum_i z_{ij}\left[\sum_k y_{ik}(x_i - u_k)\right] = \sum_i \left[\sum_k (x_i - u_k)\mu_{ik}\right]z_{ij}. \tag{A.2}$$

Under conditions C1–C4 and Eq. 7, we may use the approximations

$$\sum_k (x_i - u_k)\mu_{ik} \approx 0 \tag{A.3}$$

$$\sum_i (x_i - u_k)\mu_{ik} \approx -\lambda^* u_k y_{+k} \tag{A.4}$$

because μ_{ik} is symmetric about x_i and about u_k; the proportionality constant λ^* comes in because the species' curves are the more truncated the more their optima lie towards or beyond the edge of the sampling interval (Ter Braak 1985*b*). The transition formulae Eqs. 3–6 now follow from Eqs. A.1 and A.2 by using Approximations A.3 and A.4 and the equation $\lambda = 1 - \lambda^*$.

Starting from Eq. 5 we substitute for $\boldsymbol{x}^*$ (Eq. 4), u_k (Eq. 3), and finally x_i (Eq. 6) and obtain

$$(\mathrm{S}_{21}\mathrm{S}_{11}^{-1}\mathrm{S}_{12} - \lambda\ \mathrm{S}_{22})\boldsymbol{b} = 0, \tag{A.5}$$

where $\mathrm{S}_{21} = \mathrm{Z}'\mathrm{Y}$, $\mathrm{S}_{12} = \mathrm{Y}'\mathrm{Z}$, $\mathrm{S}_{11} = \mathrm{diag}\,(y_{+1}, y_{+2}, \ldots, y_{+m})$, $\mathrm{S}_{22} = \mathrm{Z}'\mathrm{RZ}$ and $\mathrm{Y} = \{y_{ik}\}$. Similarly, successive substitutions in Eq. 3 lead to

$$(\mathrm{S}_{12}\mathrm{S}_{22}^{-1}\mathrm{S}_{21} - \lambda\ \mathrm{S}_{11})\boldsymbol{u} = 0, \tag{A.6}$$

where $\boldsymbol{u} = (u_1, \ldots, u_m)'$. Apart from the particular definitions of the matrices in Eqs. A.5 and A.6, these equations are the eigenvector equations of canonical correlation analysis, and the eigenvalue λ lies between 0 and 1 (Gittins 1985). The eigenvectors are all uncorrelated; using subscripts r and s for different axes we obtain that $\boldsymbol{u}_r'\mathrm{S}_{11}\boldsymbol{u}_s = 0$, $\boldsymbol{b}_r'\mathrm{S}_{22}\boldsymbol{b}_s = 0$ and $\boldsymbol{x}_r'\mathrm{R}\boldsymbol{x}_s = 0$. Algorithms based on Eq. A.5 or Eq. A.6 will in general be more efficient than the algorithm developed in Theory.

The first axis of canonical correspondence analysis does not maximize the species–environment correlation, i.e., the correlation between $\boldsymbol{x}$ and $\boldsymbol{x}^*$. I have also developed an eigenvector technique that maximizes the species–environment correlation. This technique requires that the number of species is smaller than the number of sites. This requirement is often a nuisance in ecological research. As we have seen, the rationale for canonical correspondence analysis is different: it is, under conditions C1–C4, almost a maximum likelihood technique.

The weighted averages of the species with respect to the environmental variables in Eq. 8 are, in matrix notation, $\mathrm{W} = \mathrm{S}_{11}^{-1}\mathrm{Y}'\mathrm{Z} = \mathrm{S}_{11}^{-1}\mathrm{S}_{12}$, where $\mathrm{W} = \{\tilde{z}_{kj}\}$. We want a least squares approximation of W in an ordination diagram. However, when a species total is low, the weighted average is imprecise (cf. Ter Braak and Looman 1986), so that it is not worthwhile to approximate that species' weighted averages very accurately in the diagram. This consideration suggests giving the species weights that are proportional to the species totals contained in S_{11}. The result would still depend on the scale of measurement of the environmental variables. To make the method scale-invariant we use S_{22}^{-1} as weights for the environmental variables. The desired weighted least squares approximation of W follows now from the singular value decomposition (see for example Greenacre 1984: Appendix A).

$$\mathrm{S}_{11}^{1/2}\mathrm{W}\mathrm{S}_{22}^{-1/2} = \mathrm{S}_{11}^{-1/2}\mathrm{S}_{12}\mathrm{S}_{22}^{-1/2} = \mathrm{P}\Lambda^{1/2}\mathrm{Q}', \tag{A.7}$$

where P and Q are orthonormal $m \times q$ and $q \times q$ matrices (respectively) and $\Lambda = \mathrm{diag}\,(\lambda_1, \ldots, \lambda_q)$. For convenience of notation it is assumed here that $q \leq m$. This singular value decomposition is just another way to solve Eqs. A.5 and A.6 (see Mardia et al. 1979: chapter 10). With Hill's (1979) scaling of site and species scores, namely

$$\sum_{i,k} y_{ik}(x_i - u_k)^2 = y_{++}, \tag{A.8}$$

the coordinates of the species points are the first two columns of the matrix

$$\mathrm{U} = y_{++}^{1/2}\mathrm{S}_{11}^{-1/2}\mathrm{P}(\mathrm{I} - \Lambda)^{-1/2}, \tag{A.9}$$

and the coordinates of the points for the environmental variables are the first two columns of the matrix

$$\mathrm{B}_e = y_{++}^{-1/2}\ \mathrm{S}_{22}^{1/2}\mathrm{Q}(\mathrm{I} - \Lambda)^{1/2}\Lambda^{1/2} = y_{++}^{-1}\mathrm{Z}'\mathrm{RX}(\mathrm{I} - \Lambda), \tag{A.10}$$

where the second equality follows after some algebra, with X the matrix whose s^{th} column is $\boldsymbol{x}_s$. In this scaling $\mathrm{U}'\mathrm{S}_{11}\mathrm{U} = y_{++}(\mathrm{I} - \Lambda)^{-1}$ and $\mathrm{X}'\mathrm{RX} = y_{++}\Lambda(\mathrm{I} - \Lambda)^{-1}$. It is easy to verify using Eqs. A.7, A.9, and A.10 that $\mathrm{W} = \mathrm{UB}_e'$. Therefore the points for species and environmental variables form a biplot (Gabriel 1971) in the sense that inner products approximate the elements of the matrix W, leading to a two-dimensional approximation W_2, say. A measure of goodness of fit is $(\lambda_1 + \lambda_2)$/(sum of all eigenvalues), which is equal to trace $(\mathrm{S}_{11}\mathrm{W}_2\mathrm{S}_{22}^{-1}\mathrm{W}_2')$/trace $(\mathrm{S}_{11}\mathrm{W}\ \mathrm{S}_{22}^{-1}\mathrm{W}')$ and is, loosely speaking, the percentage variance in the weighted averages accounted for by the biplot. When the environmental variables are scaled to zero mean and unit variance (using y_{i+} as site weights), we obtain from Eq. A.10 that the coordinate of the point for environmental variable j on axis s must be $[\lambda_s(1 - \lambda_s)]^{1/2}$ times the correlation coefficient of the environmental variable with the site scores $\boldsymbol{x}_s$. In detrended canonical correspondence analysis the coordinates of the points for the environmental variables are obtained from a multivariate regression of W on the first two columns of U, U_2 say:

$$\mathrm{B}_e = \mathrm{W}'\mathrm{S}_{11}\mathrm{U}_2(\mathrm{U}_2'\mathrm{S}_{11}\mathrm{U}_2)^{-1} = \mathrm{Z}'\mathrm{RX}(\mathrm{U}_2'\mathrm{S}_{11}\mathrm{U}_2)^{-1}, \tag{A.11}$$

which reduces to Eq. A.10 in canonical correspondence analysis.