Aquatic Sciences 57/3, 1995

1015–1621/030255-35 $ 1.50 + 0.20/0

Canonical correspondence analysis and related multivariate methods in aquatic ecology

Cajo J. F. ter Braak[1,2], and Piet F. M. Verdonschot[2]

[1] DLO Agricultural Mathematics Groups, Box 100, NL-6700 AC Wageningen, the Netherlands
[2] DLO Institute for Forestry and Nature Research, Box 23, NL-6700 AC Wageningen, the Netherlands

Key words: Multivariate response data, compositional data, unimodal model, community ecology, partial least squares.

ABSTRACT

Canonical correspondence analysis (CCA) is a multivariate method to elucidate the relationships between biological assemblages of species and their environment. The method is designed to extract synthetic environmental gradients from ecological data-sets. The gradients are the basis for succinctly describing and visualizing the differential habitat preferences (niches) of taxa *via* an ordination diagram. Linear multivariate methods for relating two set of variables, such as two-block Partial Least Squares (PLS2), canonical correlation analysis and redundancy analysis, are less suited for this purpose because habitat preferences are often unimodal functions of habitat variables. After pointing out the key assumptions underlying CCA, the paper focuses on the interpretation of CCA ordination diagrams. Subsequently, some advanced uses, such as ranking environmental variables in importance and the statistical testing of effects are illustrated on a typical macroinvertebrate data-set. The paper closes with comparisons with correspondence analysis, discriminant analysis, PLS2 and co-inertia analysis. In an appendix a new method, named CCA-PLS, is proposed that combines the strong features of CCA and PLS2.

Introduction

People wish to know how human activity influences the fascinating diversity of biological communities. Yet, this very diversity creates problems for the statistical analysis of ecological observations: it implies a large number of species and a large inherent variability. A set of community samples and associated environmental measurements (e.g. water chemistry variables) typically yields an enormous amount of noisy data which is difficult to interpret. Multivariate methods provide a means to structure the data by separating systematic variation from noise (Gauch, 1982). Two important aspects distinguish ecological data from other noisy multivariate data. First, most species occur only in a subset of the samples; the data have therefore the character of incidence data (1/0 indicating presence/absence) even if abundance is measured quantitatively (e.g. number of individuals or biomass of

each species present). Second, relationships between species and environmental variables are generally nonlinear, and what is worse, even non-monotonic. Because of Shelford's law of tolerance (Odum, 1971) and the associated idea of niche-space partitioning (Whittaker et al., 1973), species abundance or probability of occurrence is often a unimodal function of the environmental variables. These two aspects make traditional linear-based multivariate methods unsuitable. In contrast, canonical correspondence analysis takes advantage of these aspects (ter Braak, 1986, 1987a, b; Chessel et al., 1988).

Historically, canonical correspondence analysis builds on the method of weighted averaging of indicator species proposed by the early great ecologists such as Gause (1930), Ellenberg (1948) and Whittaker (1948: in Gauch, 1982), and widely used in biological water-quality assessment (Pantle and Buck, 1955; von Tümpling, 1966; Sládecek, 1986; Zelinkan and Marvan, 1961; Descy, 1979). It extends weighted averaging to the simultaneous analysis of many species and many environmental variables. Canonical correspondence analysis also builds on the ordination method of reciprocal averaging, alias correspondence analysis (Hill, 1973, 1974; Hill and Gauch, 1980). It adds to correspondence analysis the statistical methodology of regression. The method thus provides a general framework for estimation and statistical testing of the effects of environmental variables and other explanatory variables on biological communities, even if the effects are hidden by other large sources of variation. In summary, canonical correspondence analysis is a method that can help aquatic ecologists unravel how a multitude of species simultaneously respond to external factors, such as environmental variables, pollutants and management regime, using data either from observational studies or from designed experiments.

Canonical correspondence analysis (CCA) and related methodology has found wide-spread use in aquatic sciences. The bibliography by Birks et al. (1994) lists under the subject headings limnology, marine biology, and palaeolimnology 86, 25 and 49 papers, respectively. Organisms studied are (with number of entries between brackets) diatoms (62), other algae (18), aquatic invertebrates (17), chrysophytes (11), fish (11), phytoplankton (6), zooplankton (4), oligochaetes (3) and foraminifera (2). The most frequent use is to identify environmental gradients in ecological data-sets (Barker, 1994), in particular, which environmental variables are important in the determination of the community composition. Recent examples include Jones, Juggins and Ellis-Evans (1993), Grantham and Hann (1994) and Malmqvist and Maki (1994). In palaeolimnology, CCA is frequently used as a preliminary analysis for determining whether particular variables influence the present-day communities sufficiently to warrant palaeo-reconstruction from fossil assemblage data (Walker et al., 1991; Cumming, Smol and Birks, 1992; Anderson, Rippey and Gibson, 1992; Fritz, Juggins and Batterbee, 1993; Charles and Smol, 1994). Although CCA can be used for palaeo-reconstruction (Stevenson et al., 1989), for example by adding fossil assemblage data to a CCA ordination diagram of the modern data (Birks, Juggins and Line, 1990a), more specialized methods are available (ter Braak and van Dam, 1989; Birks et al., 1990b; Anderson, 1993; Line, ter Braak and Birks, 1994; ter Braak and Juggins, 1993; ter Braak, 1995a). CCA is also a means of studying seasonal and spatial variation in communities (Snoejis and Prentice, 1989; Bakker, Herman and Vink, 1990; Anderson, Korsman and Renberg,

1994) and of assessing to what extent this variation can be explained by associated environmental variation (Soetaert et al., 1994; Kautsky and van der Maarel, 1990). The variance can be fully decomposed into seasonal, spatial, environmental and random components (Borcard et al., 1992; Økland and Eilertsen, 1994). Copp (1992) and Reilly and Fiedler (1994) used CCA for niche analysis. This use of CCA has an early pre-cursor in the form of Green's (1971, 1974) multi-group discriminant analysis for quantifying the multivariate Hutchinsonion niche of species. CCA has also been used in a number of impact studies (van Nes and Smit, 1993; Snoeijs, 1989; Gower et al., 1994) and for testing hypotheses about the effect of particular water chemistry variables on community composition (Walker et al., 1991; Kingston et al., 1992). Verdonschot (1989) used CCA for biological water-quality assessment and related management problems. CCA can also be used for analyzing community data from experiments (Sundbäck and Snoeijs, 1991; Fairchild and Sherman, 1993; Verdonschot and ter Braak, 1994). In some advanced uses of CCA, it is a powerful alternative for the multivariate analysis of variance (MANOVA; Verdonschot and ter Braak, 1994; van Wijngaarden et al., 1995), for example in the analysis of data from Before-After-Control-Impact studies (Green, 1979; Stewart-Oaten, Murdoch and Parker, 1986; Carpenter, Frost and Heisey, 1989), both with (Verdonschot and ter Braak, 1994) and without replication of the impacted site (Underwood, 1992).

As an introduction to CCA, this paper summarizes how CCA identifies major environmental gradients in ecological data-sets and how the analysis can focus on the effect of particular environmental variables by partialling out nuisance variation (partial CCA). This part of the paper follows ter Braak (1987a). Subsequently, an attempt is made to single out the key assumptions underlying CCA by comparing various derivations of CCA. Since 1987, the standard ordination diagram of CCA has undergone some changes that aid interpretation. This is the first paper to fully discuss the new standard (ter Braak, 1990), which follows and extends proposals by Chessel et al. (1987), Lebreton et al. (1988a, b) and Greenacre (1993). Thereafter, the paper introduces the method of ranking environmental variables in importance by forward selection. The theory is exemplified using macroinvertebrate data from two man-made tributaries of a Dutch lowland stream (Higler and Verdonschot, in prep.). Additional insight into CCA is provided by contrasting it with other multivariate methods, such as discriminant analysis, correspondence analysis, two-block PLS and co-inertia analysis. CCA inherits many of its unimodal properties from its close relationship to discriminant analysis. If more and more environmental variables are added to CCA, the method becomes increasingly similar to correspondence analysis, paradoxically a method that was designed to work without environmental data! The problem of many environmental variables also plays a major role in the comparisons of CCA with two-block PLS and co-inertia analysis. In the appendix, a new method is proposed that attempts to combine many strong points of these three methods. The discussion attempts to delimit the role of CCA in the aquatic sciences.

Example data

We analyze macro-fauna data from two man-made tributaries in the upstream part of the Hierden stream, a well-studied lowland stream on the Veluwe, the Netherlands (Higler and Repko, 1981). The discharge area of the stream is situated in fluvio-glacial deposits and the main source of water is diffuse ground-water seepage. The aim of the study was to compare the macrofauna in the tributaries which are similar in morphology, but different in nutrient load as a result of differences in land use in the drainage area. As a proxy for nutrient load, electrical conductivity (EC) is used. The distribution of the macrofauna in the tributaries will be related to hydraulic, physical and chemical variables.

The two tributaries L (Leuvenum stream) and U (Uddel stream) were sampled from source to mouth in the Hierden stream at 21 and 19 different locations, respectively. Sampling took place in five different months (Table 1) in the period October 1983 – August 1984. Each location (henceforth called site) was sampled once. In each of the months, as many upstream as downstream sites were sampled and nearly as many L- as U-sites. At each site, material from the upper layer of the sediment and the vegetation, if present, was collected over a 1 metre stretch along the stream and over the total width of the stream. Simultaneously, material was taken from the sediment for grain-size and organic-matter content analysis. The vegetation, shading and current velocity were recorded. The electrical conductivity (EC) was averaged across 3–4 measurements taken in the period January–June 1984. The environmental variables recorded are listed in Table 1. Three variables are ordinal, but are treated quantitatively in CCA with the codings 0, 1, 2, 3, and so on. Apart from the sampling month (five classes), there are two qualitative variables that classify the bank vegetation and substrate in four and three classes, leading to seven binary class variables. The classes of substrate are not mutually exclusive; if the substrate is heterogeneous, more than one class was ticked. The sample distribution of each quantitative variable was inspected for outliers and strong asymmetry, but eventually the data were left unmodified. In the laboratory, the animals in the sample material were sorted alive, identified and counted. The number per taxon was logarithmically transformed so as to downweight large numbers. The problem of taking the logarithm of zero was circumvented by adding 1 to each number before transformation. In total, 197 taxa were identified. The abundance table (**Y** in Fig. 1) thus contains 197×40 non-negative values; 84% are, however, zero. The number of taxa per site varies between 9 and 68. The number of occurrences per taxon varies between 1 and 35. Many taxa occur only a few times and could have been deleted without much influence on the analysis.

Theory of canonical correspondence analysis (CCA)

Ecological derivation: niche separation and CCA

In this section canonical correspondence analysis is introduced: the method operates on (field) data on occurrences or abundances (e.g. counts of individuals) of species and data on environmental variables at sites (Fig. 1), and extracts from the measured environmental variables synthetic gradients (ordination axes) that maximize the niche separation among species.

Table 1. Example data: quantitative and qualitative environmental variables (a) and qualitative covariables (b) recorded at 40 sites along two tributaries from the Hierden stream (sd: standard deviation, min: minimum, max: maximum)

a. Environmental variables

Quantitative variables	mean	sd	min	max
Source distance (m)	953	513	25	1730
Stream width (m)	1.0	0.23	0.6	1.6
Stream depth (m)	0.12	0.09	0.05	0.45
Mean current velocity (m/s)	23.7	11.0	5.0	45.0
Electrical conductivity, EC (μS/m)	358	201	120	690
Discharge (m^3/s)	85.3	49.0	5	175
Cover percentage of:				
algae	11.4	26.7	0	100
submerged vegetation	10.7	22.2	0	90
emergent vegetation	9.5	26.5	0	100
bank vegetation	3.8	8.4	0	40
Total cover percentage of vegetation	36.0	40.5	0	100
Soil grain size[1]	2.5	0.6	1	3
Coverage of substrate[2]	1.7	1.4	0	7
Shading[3]	1.0	0.9	1	3
Qualitative variables	**frequency**			
Bank vegetation				
Grassy	7			
Hanging weedy	21			
High weedy	5			
Shrubs	7			
Substrate				
Coarse detritus	17			
Fine detritus	31			
Silt	3			

b. Covariables	frequency
Month of sampling:	
October 1983	6
January 1984	8
April 1984	9
June 1984	9
August 1984	8

[1] Levels and coding: coarse sand/gravel (1), coarse sand/fine sand (2), fine sand (3)

[2] Levels and coding: none (0), local thin layer (1), spread thick layer (2), thin layer < 2 cm (3), mixed soil-substrate layer (4), less thick layer 2–5 cm (5), thick layer 5 cm (6), very thick layer 10 cm (7)

[3] Levels and coding: none (0), low (1), average (2), high (3)

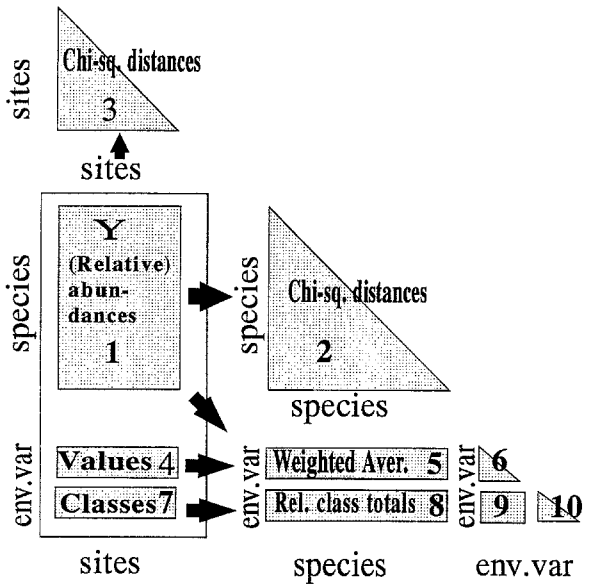

Figure 1. Data-tables in an ecological study on species-environment relations. Primary data are the sub-table 1 of abundance values of species and the sub-tables 4 and 7 of values and class labels of quantitative and qualitative environmental variables (env. var), respectively. The primary data are input for canonical correspondence analysis (CCA). The other sub-tables contain derived (secondary) data, as the arrows indicate, named after the (dis)similarity coefficient they contain. The coefficients shown in the figure are optimal when the species-environment relations are unimodal. The CCA ordination diagram represents these sub-tables, with emphasis on sub-tables 5 (weighted averages of species with respect to quantitative environmental variables), 8 (totals of species in classes of qualitative environmental variables) and 1 (with fitted, as opposed to observed, abundance values of species). The sub-tables 6, 9 and 10 contain correlations among quantitative environmental variables, means of the quantitative environmental variables in each of the classes of the qualitative variables and chi-square distances among the classes, respectively. See also Table 2 (Chi-sq. = Chi-square; Aver. = Averages; Rel. = Relative)

The occurrence or abundance of a species along an environmental gradient often follows Shelford's Law of Tolerance (Shelford, 1911; Odum, 1971): each species thrives best at a particular value (its optimum) and cannot survive when the value is either too low or too high. Each species' occurrence is thus confined to a limited range, its niche. The fundamental niche of a species is determined by physiological processes and cannot normally be observed in the real world because species coexist in communities. What can be observed is the realised niche as modified by competition among species and other intra-community processes. It is the realised niche that is of interest in applied ecology. Species tend to separate their niches, partly so as to minimize competition. If the separation is strong, successive species replacements occur along the environmental gradient. The composition of biotic communities thus changes along environmental gradients according to unimodal functions (Fig. 2). Of course, some species may prefer extreme environmental conditions or their optima may fall outside the environmental region

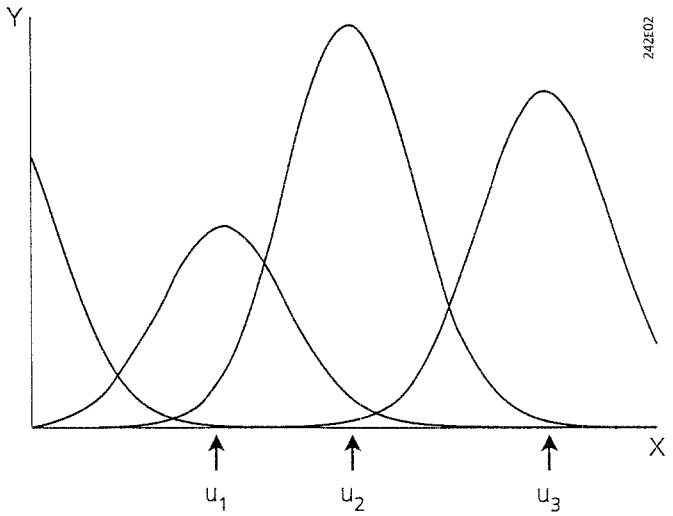

Figure 2. Unimodal curves for the expected abundance or response (y) of four species against an environmental gradient or variable (x). The optima, estimated by weighted averages, (u_k) $[k = 1, 2, 3]$, of three species are indicated. The curve for the species on the left is truncated and therefore appears monotonic instead of unimodal; its optimum is outside the sampled interval, but its weighted average is inside. The curves drawn are symmetric, but this is no strict requirement for CCA

actually sampled in a particular study, so that their observed response function is not unimodal, but monotonic decreasing or increasing (Fig. 2). Hutchinson (1968) extended the niche concept to p-dimensions. Each species thus occurs in a characteristic, limited range of the multi-dimensional habitat space; and within this range, each species tends to be most abundant around a specific environmental optimum (Green, 1971). Of course, not all measurable features are equally important and some features may perhaps be combined into a synthetic environmental gradient so as to enhance the niche separtion along that gradient. Canonical correspondence analysis is the method that extracts the "best" synthetic gradients from field data on biological communities and environmental features: it forms a linear combination of environmental variables that maximally separates the niches of the species. Niche separation is hereby expressed as the weighted variance of species centroids on a standardized gradient, the species centroid being the (weighted) average of the gradient values of the sites at which the species occurs (Boxes 1 and 2). The species centroid, or weighted average, is an estimate of the species' optimum if the response curve of the species is symmetric as in Fig. 2. The first synthetic gradient is termed the first ordination axis. The achieved maximum amount of niche separation is given by the eigenvalue of the ordination axis (Box 2). The mathematics involved is given in ter Braak (1987a) and Jongman et al. (1995). Subsequent ordination axes are also linear combinations of the environmental variables that maximally separate the niches, but subject to the constraint that they are uncorrelated with the axis or axes extracted previously. In principle, as many ordination axes can be extracted as there are environmental variables, but because the amount

Box 1. The weighted average and standard deviation of a species.

The weighted average or niche-centroid (u_k) of species k with respect to any gradient x (environmental variable, synthetic gradient or ordination axis) is defined as the weighted average of the gradient values of the sites at which the species occurs, i.e.

$$u_k = \sum_{i=1}^{n} \frac{y_{ik}}{y_{+k}} x_i \qquad (1)$$

with y_{ik} the abundance (0/1, count, biomass or other nonnegative value) of species k in site i ($i=1,\ldots,n$; $k=1,\ldots,m$), x_i the value of gradient x at site i and the subscript "+" replacing an index denoting the sum over the index, hence y_{i+} is the abundance total across species in site i.

The weighted standard deviation of a species (also termed its tolerance, a measure of niche-breadth) is

$$t_k = \sqrt{\sum_{i=1}^{n} \frac{y_{ik}}{y_{+k}} (x_i - u_k)^2}. \qquad (2)$$

The weighted standard deviation gives a good impression of the range of x-values over which a species occurs, but underestimates the true tolerance or true niche-breadth. An extreme case is that $t_k = 0$ if a species occurs only once. For a fair statistical comparison of niche breadth, the bias must be removed. This can be achieved (as in Hill, 1979: p. 28), by division in (2) by $y_{+k}(1-1/N_2)$ instead of by y_{+k} with N_2 the effective number of occurrences of species k,

$$N_2 = \left\{ \sum_{i=1}^{n} \left(\frac{y_{ik}}{y_{+k}} \right)^2 \right\}^{-1}. \qquad (3)$$

Intuitively, if a species occurs at three sites with abundances 1000, 1, and 1, then its u_k is effectively determined by the x-value of the site where the abundance is 1000, so that $t_k \approx 0$. The effective number of occurrence is close to 1 (instead of being 3) and the N_2-adjusted tolerance is correspondingly large. For incidence data, the N_2-adjusted tolerance is precisely the sample (instead of: population) weighted standard deviation (Carnes and Slade, 1982: 892).

of niche separation (the eigenvalue) decreases with increasing axis number, it is often sufficient to inspect only the first few axes. The computer program CANOCO (ter Braak, 1987–1990) extracts only the first four ordination axes per run.

CCA adds the full power of regression methodology to ordination. This comes about because CCA uses, as linear regression does, linear combinations of environmental (explanatory) variables to explain optimally the species (response) variables. The unusual features of CCA are that the measure of fit is unconventional (weighted variance of species centroids) and that the data of many species are explained simultaneously. The consequences for the statistical analysis are discussed later on.

Covariables: partial CCA

The example macrofauna data were sampled in five different months. It is therefore likely that there is considerable seasonal variation in the biological assemblage and

Box 2. Definition of CCA by maximum niche separation.

For a standardized gradient x, i.e. a gradient for which

$$\sum_{i=1}^{n} \frac{y_{i+}}{y_{++}} x_i = 0 \ \wedge \ \sum_{i=1}^{n} \frac{y_{i+}}{y_{++}} x_i^2 = 1 \tag{4}$$

the weighted variance of species centroids $\{u_k\}$ $(k=1 \ldots m)$ of equation (1) is defined by

$$\lambda = \sum_{k=1}^{m} \frac{y_{+k}}{y_{++}} u_k^2. \tag{5}$$

Now let x be a synthetic gradient, i.e. a linear combination of environmental variables

$$x_i = \sum_{j=1}^{p} c_j z_{ij} \tag{6}$$

with z_{ij} the value of environmental variable j $(j=1,\ldots,p)$ in site i and c_j its coefficient or weight (not necessarily positive). Then, CCA is the method that chooses the optimal weights $\{c_j\}$, i.e. the weights that result in a gradient x for which the weighted variance of the species scores (5) is maximum. Mathematically, the synthetic gradient x can be obtained by solving an eigenvalue problem; x is the first eigenvector $\mathbf{x}_1$ with eigenvalue the maximum λ (ter Braak, 1987a). The optimized weights are termed canonical coefficients. Each subsequent eigenvector $\mathbf{x}_s=(x_{1s}, \ldots, x_{ns})'$ $(s>1)$ maximizes (5) subject to constraint (6) and the extra constraint that it is uncorrelated with previous eigenvectors, i.e. $\sum_i y_{i+} x_{it} x_{is}=0$ $(t<s)$.

the environment. This seasonal variation was not the prime research question, however, and should therefore not enter the synthetic gradients. This can be achieved by a partial canonical correspondence analysis (partial CCA: ter Braak, 1988a) with the five class variables representing sampling months as covariables. A partial CCA amounts to a normal CCA, but with the extra requirement that each synthetic gradient must be uncorrelated with the covariables. This requirement takes the same form as that for a second or later axis in CCA, namely that it must be uncorrelated with previously extracted synthetic gradients. The covariables thus take the role of extra gradients that are already extracted. Partial CCA is effective if the sets of covariables and environmental variables are uncorrelated or show a moderate correlation. Our example data were collected according to a reasonably balanced sampling design, leading to very moderate correlations between seasonal variation and other time-independent variation. The CCA example presented later on is a CCA adjusted for seasonal variation, i.e. a partial CCA.

Assumptions and alternative derivations

This section contains some advanced material; readers may wish to skip to the closing paragraph on first reading. Originally, CCA was derived as an approximation to maximum likelihood Gaussian ordination with linear external constraints (ter Braak, 1986, 1988a). In this derivation, strong assumptions were used that are

Box 3. CCA as a form of redundancy analysis.

Sabatier et al. (1989) and Lebreton et al. (1991) showed that CCA is a weighted form of principal components analysis with respect to instrumental variables $\{z_{ij}\}$ (Rao, 1964), alias redundancy analysis, alias least-squares reduced-rank regression (ter Braak and Looman, 1994). In particular, the first ordination axis of CCA minimizes

$$L = \sum_{i,k} y_{i+}y_{+k} \left\{ \frac{y_{ik}y_{++}}{y_{i+}y_{+k}} - 1 - u_k x_i \right\}^2 \tag{7}$$

subject to constraint (6). On inserting (6) in (7), CCA is seen to be a regression method that minimizes

$$L = \sum_{i,k} y_{i+}y_{+k} \left\{ \frac{y_{ik}y_{++}}{y_{i+}y_{+k}} - 1 - \sum_{j=1}^{p} b_{jk} z_{ij} \right\}^2 \tag{8}$$

subject to the constraints

$$b_{jk} = u_k c_j \tag{9}$$

The matrix of regression coefficients $\{b_{jk}\}$ is thus required to be of rank 1. With r ordination axes (rank r),

$$L = \sum_{i,k} y_{i+}y_{+k} \left\{ \frac{y_{ik}y_{++}}{y_{i+}y_{+k}} - 1 - \sum_{s=1}^{r} u_{ks} x_{is} \right\}^2 \tag{10}$$

is minimized. Equivalently, (8) is minimized subject to the constraints

$$b_{jk} = u_{k1}c_{j1} + \ldots + u_{kr}c_{jr} \tag{11}$$

unlikely to hold true in applications. Because of this, Austin et al. (1994) and Austin and Gaywood (1994) express concern about the validity of the method. Fortunately, CCA appears extremely robust to deviation from these assumptions and other derivations do not necessarily rely on them. As always in statistics, with stronger assumptions stronger optimally properties of a method can be proven. Clearly, the ecological derivation of CCA in this paper requires minimal assumptions but on the other hand only guarantees that the dispersion (weighted variance of species centroids) is maximized. Equally undemanding in terms of underlying assumptions is a derivation by Takane, Yanai and Mayekawa (1991), based on work by Heiser (1987), in which CCA is a constrained unfolding method. Unimodality is the key assumption in these derivations. Even this assumption is not needed. As Sabatier et al. (1989) showed, CCA can be derived as a weighted form of the method of reduced rank regression (ter Braak, 1990a; ter Braak and Looman, 1994), which is also known under the names of redundancy analysis and principal component analysis with respect to instrumental variables (Rao, 1964). The key element in this derivation of CCA is that the relative abundance is a linear function of the environmental variables (relative here means relative to both the site total and the species total, i.e. $y_{ik}/y_{i+}y_{+k}$). This characterization (Box 3) is the basis of the least-squarer properties and associated biplot interpretation of the CCA ordination diagram in

rows 1 and 8 of Table 2, as discussed in the next section. As unimodality and compositional data often go hand in hand (ter Braak, 1995a), it emerges that the common element in all these derivations is that CCA models compositional (i.e. relative) abundance data instead of the absolute abundance data.

To summarize this section in more ecological terms, CCA models relative abundances. It thus takes the size of the sample taken at a site for granted. Usually the alpha diversity of a sample increases with its size. CCA takes that aspect of alpha diversity for granted and focuses, instead, on the beta-diversity (dissimilarity among sites). Sometimes the trend in alpha diversity coincides with beta-diversity, for example if species one by one disappear along a toxicity gradient. CCA is capable of extracting such trends (Iwatsubo, 1984: theorem 2).

Ordination diagrams and their interpretation

Introduction, interpretation of the ordination axes

The primary result of a CCA is an ordination diagram, i.e. a graph with a coordinate system formed by ordination axes (i.e. the synthetic gradients extracted by CCA). As illustrated in Figure 3, a CCA ordination diagram may consist of the following elements: points for species, sites and classes of qualitative environmental variables, and arrows for quantitative environmental variables. There exist a number of slightly different variants of the CCA diagram. In particular, axes may be differentially magnified or compressed with respect to one another. The differences in scaling of the diagram are unimportant if the eigenvalues of the axes are about equal. Table 2 summarizes the properties of the two variants discussed in this paper. We start with the species-conditional CCA biplot (third column of Table 2) which is the new standard in the computer program CANOCO version 3.1 (ter Braak, 1990b). The standard in earlier versions, Hill's scaling (second column of Table 2), is briefly discussed later on in a separate subsection.

The new standard ordination diagram naturally follows from the ecological derivation of CCA (Box 2), and is constructed and interpreted as follows. The coordinates of the site points are the values (termed scores) of the sites on the two best synthetic gradients (axes 1 and 2 in Fig. 3). Recall from the initial derivation of CCA that each gradient is standardized to zero weighted mean and unit weighted variance, and that species are represented by their niche centre along each axis, i.e. by the weighted average of the axis-scores of sites in which they occur (Fig. 2). Consequently, each species point in the diagram is at the centroid (weighted average) of the site points in which it occurs. The species points thus indicate the relative locations of the two-dimensional niches of the species in the ordination diagram. In principle, the niche breadths could be indicated also, namely by the weighted standard deviation on each synthetic gradient (Box 1), but this is not done in Fig. 3. From the definition of CCA, it would be natural to display the environmental variables by the weights that each variable has in the linear combinations that form the axes. With correlated environmental variables, these weights are often difficult to interpret (ter Braak, 1986; Eriksson et al., 1995). Instead, quantitative environmental variables are displayed by their correlations with the axes and qualitative

Table 2. Sub-tables of Fig. 1 (row numbers) that can be displayed by two differently scaled ordination diagrams in canonical correspondence analysis (CCA). Display is by the biplot rule unless noted otherwise. Hill's scaling (column 2) was the default in CANOCO 2.1, whereas the species-conditional biplot scaling (column 3) is the default in CANOCO 3.1. The weighted sum of squares of sites scores of an axis is equal to $\lambda/(1-\lambda)$ with λ its eigenvalue and equal to 1 in scaling −1 and scaling 2, respectively. The weighted sum of squares of species scores of an axis is equal to $1/(1-\lambda)$ and equal to λ in scaling −1 and scaling 2, respectively. If the scale unit is the same of both species and sites scores, then sites are weighted averages of species scores in scaling −1 and species are weighted averages of site scores in scaling 2. Tables in italic are fitted by weighted least-squares (rel. = relative; env. = environmental; vars = variables; cl. = classes; – = interpretation unknown)

Scaling		−1: focus on sites Hill's scaling	2: focus on species biplot scaling of CCA
1	species × sites[a]	rel. abundances[b,c]	*fitted rel. abundances*[b]
2	species × species	–	chi-square distances[d]
3	sites × sites	turnover distances[c,e]	[f]
Quantitative env. vars:			
4	sites × env. vars[g]	–	values of env. vars
5	species × env. vars	*weighted averages*	*weighted averages*
6	env. vars × env. vars	effects[h]	correlations
Qualitative env. vars:			
7	sites × env. classes[i]	membership[k]	membership[k]
8	species × env. classes	rel. total abund.[c,b]	*rel. total abund.*[b]
9	env. vars × env. cl.	–	mean values of env. vars
10	env. classes × env. cl.	turnover distances[c,e]	[f]

[a] Site scores are linear combinations of the environmental variables. The adjective "fitted" must be deleted if site scores are proportional to the weighted average of species scores, as in ter Braak (1986, 1987a, b)

[b] The centroid principle can be applied also if sites and species scores are plotted in the same units, i.e. in scaling −1, species that occur in a site lie around it, whereas in scaling 2, the species' distribution is centred at the species point

[c] The biplot rule cannot be applied

[d] In the definition of this coefficient, abundance must be replaced by fitted abundance values, because CCA is correspondence analysis of fitted abundance values

[e] No explicit formula known

[f] Chi-square distances, provided the eigenvalues of the axes are of the same magnitude

[g] Environmental scores are (intra-set) correlations in scaling 2; more precisely, the coordinate of an arrow head on an axis (i.e. the score) is the weighted product-moment coefficient of the environmental variable with the axis, the weights being the abundance totals of the sites (y_{i+}). The scores in scaling −1 are $\{\lambda(1-\lambda)\}^{1/2}$ times those in scaling 2

[h] Effect is defined as the change in site scores if the environmental variable changes one standard deviation in value (while neglecting the other variables)

[j] Enviromental points are centroids of site points

[k] Via centroid principle, not via biplot

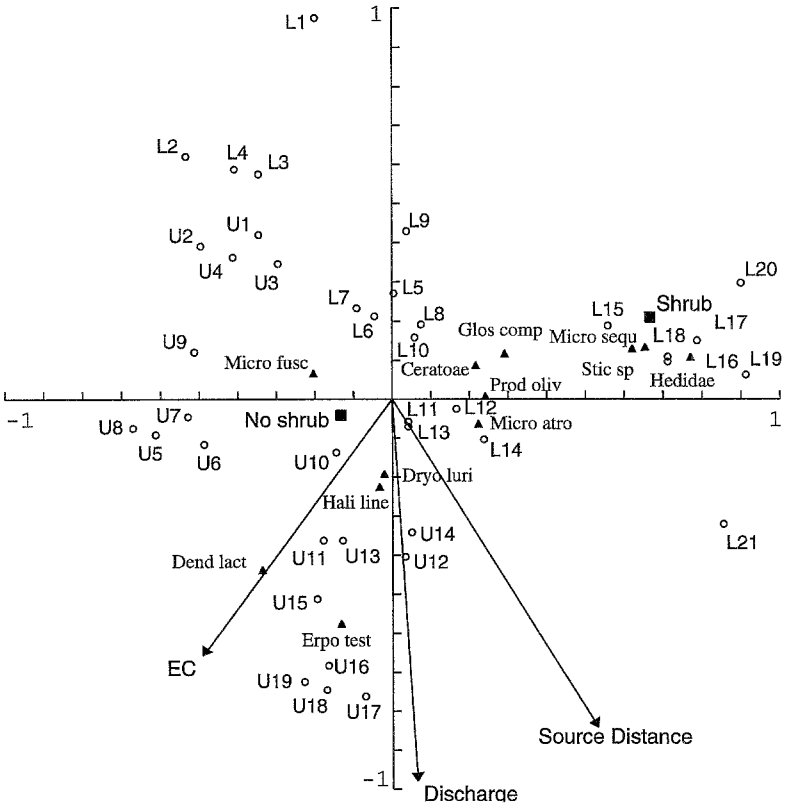

Figure 3. Species-conditional triplot based on a canonical correspondence analysis of the example macroinvertebrate datra displaying 13% of the inertia (= weighted variance) in the abundances and 69% of variance in the weighted averages and class totals of species with respect to the environmental variables. The eigenvalues of axis 1 (horizontally) and axis 2 (vertically) are 0.35 and 0.17, respectively; the eigenvalue of the axis 3 (not displayed) is 0.13. Sites are labelled with stream code (U, L) and are ranked by distance from the source (rank number within stream). Species (triangles) are weighted averages of site scores (circles). Quantitative environmental variables are indicated by arrows. The class variable shrub is indicated by the square points labelled Shrub and No shrub. The scale marks along the axes apply to the quantitative environmental variables; the species scores, sites scores and class scores were multiplied by 0.4 to fit in the coordinate system. Only selected species are displayed which have $N_2 > 4$ and small N_2-adjusted root mean square tolerance for the first two axes. The species names are abbreviated to the part in italic as follows *Ceratopogonidae*, *Dendrocoelum lacteum*, *Dryops luridus*, *Erpobdella testacea*, *Glossiphonia complanata*, *Haliplus lineatocollis*, *Helodidae*, *Micropsectra atrofasciata*, *Micropsectra fusca*, *Micropterna sequax*, *Prodiamesa olivacea*, *Stictochironomus sp.*

environmental variables by the centroids of their classes. More precisely, the arrow for an quantitative variable runs from the origin (centre) of the diagram to an arrow head, the coordinates of which are the correlations of the variable with the axes. A qualitative environmental variable consists of a number of classes that partition the sites; each class is naturally represented by a point in the diagram, namely the centroid of sites points belonging to the class. (The centroid is a weighted average, the weight being the total abundance of a site).

The place of each element in the diagram already gives ample scope for interpretation. For example, the positioning of the environmental variables in Fig. 3 shows that the first synthetic gradient (i.e. the main explainable variation in the faunal composition) is positively correlated with the source distance (ca. 0.5) and negatively with EC (ca. −0.5). The position and separation of the points for "shrubs" and "no shrubs" along the first axis indicate that sites with positive scores on the first axis (that lie at the right-hand site of the diagram) usually do not have shrubs (they border more intensively farmed land). The two classes of sites (with and without shrubs) thus differ systematically in faunal composition. The second axis is strongly negatively correlated (ca. −0.95) with the discharge rate. The site points are based on linear combinations of these environmental variables. For ease of interpretation of the configuration of the site points, each site is labelled by the first letter of its stream name (L or U; a distinction not used in the CCA) and its ranked distance from the source. The L-sites are well separated from the U-sites in the diagram, and thus differ in faunal composition, but the separation is much more pronounced downstream than upstream.

Interpretation of species and site points

So far, interpretation focused on the synthetic gradients in conjunction with an abstract notion of variation in faunal composition. In addition, the ordination diagram can be interpreted in much more definite terms, namely in terms of the data-tables used in the analysis and in terms of derived data-tables (Fig. 1). The ordination diagram in Fig. 3 summarizes the main structure of all ten tables of Fig. 1 (except the site × site table in the case when the ratio of eigenvalues of the axes differs strongly from 1). Table 2 presents an exhaustive list which is discussed sequentially in what follows.

The first three rows of the body of Table 2 concern the interpretation of the species and sites configurations. According to row 1 of Table 2, species points and site points jointly represent the species × site table of fitted relative abundances of species in sites. This fitted table replaces the observed one, that CCA was applied to (sub-table 1 in Fig. 1). It should therefore be possible to infer the approximate relative abundances from the diagram. There is an ongoing discussion in the literature on how this should be done precisely (Greenacre, 1989, 1993; ter Braak, 1985). This is no surprise, because there are at least four possible ways to infer the fitted relative abundances from this variant of the CCA diagram: (1) by means of the centroid principle, (2) the distance rule, (3) the biplot rule, and (4) the biplot rule for compositional data. In the first instance, attention is restricted to the centroid principle and the distance rule. These rules are easiest (but most qualitative) and

are most pertinent to the ecological applications of (canonical) correspondence analysis. Later on, a biplot rule, due to Greenacre (1993), is presented.

The centroid principle is as follows. Species are at the centroid of their niche, i.e. at the centroid of the points for sites in which they occur. Therefore, sites that contain a particular species are scattered around the point of that species. For example, *Erpobdella testaceae* is, in Fig. 3, at the centre of the downstream Uddel sites; so its distribution is centred on these sites. Similarly, the position of *Micropterna sequax* in Fig. 3 shows that the distribution of this species is largely confined to the downstream Leuvenum sites.

The centroid principle can be extended a little towards a distance rule. Because the centroid is actually a weighted average (with the weight being the abundance), the sites close to the species point tend to have a higher abundance than sites far from the species point. *The inferred abundance of a species is thus maximal if the site point coincides with the species point and decreases in all directions the farther away the site point is.* This is the *distance rule*, at least if the decrease is the same in all directions. (In the diagram discussed here, the decrease is, however, somewhat greater along the first axis than along the second axis, a difference that becomes important if the first eigenvalue is more than, say, twice the second). For example, from Fig. 3 we would infer that *Prodiamesa olivacea* has its maximum abundance in sites L11–L14 and decreases more upstream and downstream in the Leuvenum stream.

According to row 2 of Table 2, the species points among themselves represent the species × species table of chi-square distances (sub-table 2 in Fig. 1). This is a table that can be derived from the primary species × site table by a mathematical formula. The table is square and symmetric and is therefore indicated by a lower triangle in Fig. 1. The chi-square distance is a measure of the dissimilarity between the abundance profile across sites of one species and that of another. The most striking feature in the mathematical formula of the chi-square distance (Box 4) is that it is the relative abundances that are being compared. Differences in total abundance among species thus do not increase their dissimilarity as measured by the chi-square distance. The rule for inferring the chi-square distance from the diagram is simple: chi-square distance increases the further apart two species are in the diagram. Species that are close are thus expected to be similar in their distribution across the sites, whereas species that are far apart are expected to be dissimilar. Be aware that points that are close may show considerable dissimilarity if the

Box 4. Dissimilarity as measured by the chi-square distance.

The chi-square distance between the abundance profile of two species k and l is

$$\delta_{kl} = \left\{ \sum_{i=1}^{n} \frac{y_{++}}{y_{i+}} \left(\frac{y_{ik}}{y_{+k}} - \frac{y_{il}}{y_{+l}} \right)^2 \right\}^{\frac{1}{2}}. \tag{12}$$

The chi-square distance between the abundance profile of two sites is defined analogously by interchanging rows and columns of the matrix $\mathbf{Y} = \{y_{ik}\}$.

ordination fits badly, because the points may be far apart on ordination axes other than the ones shown in the diagram. Points that are far apart can, however, be trusted to be dissimilar.

The inter-sites distances are discussed now (row 3 of Table 2). If the eigenvalues of the axes are of the same magnitude, distances among sites in the diagram reflect the site × site table of chi-square distances (sub-table 3 in Fig. 1). This is another table that can be derived from the primary species × site table by a mathematical formula (Box 4). Note again from Box 4 that it is the relative abundances that are being compared. Differences in total abundance among sites thus do not necessarily increase the dissimilarity, although CCA may still pick up trends in species richness (as shown for correspondence analysis by Iwatsubo, 1984). If the second eigenvalue is a magnitude smaller than the first, the species-conditional distance diagram (last column of Table 2) overemphasizes the distances among sites along the second axis. Therefore, the approximation of chi-square distances among sites is not mentioned in the last column of Table 2. For a better representation of the chi-square distance among sites in Fig. 3, the second axis should be compressed with respect to the first axis, namely by multiplying the site scores of the second axis by a factor of $(\lambda_2/\lambda_1)^{1/2} = (0.17/0.35)^{1/2} = 0.70$. This yields a site-conditional scaling. Fortunately, this change does not influence the earlier global interpretation of between-site differences in Fig. 3 in terms of stream (U versus L) and Source distance.

Interpretations based on the environmental arrows

Rows 4–6 of the body of Table 2 concern interpretations that use the arrows for quantitative environmental variables. According to row 4 of Table 2, the site points and the environmental arrows jointly represent the site × environmental variable table, the second of the primary data-tables that CCA was applied to (sub-table 4 in Fig. 1). The points and arrows form a *biplot* (Gabriel, 1982), that is an ordination diagram with specific rules about how the points and arrows represent the data entries in the table. The most useful rules are summarized in the following (for more details see Gabriel (1982), Gabriel and Odoroff (1990), and ter Braak (1987b, 1994)). There is a useful symbolism in the use of arrows in biplots: *the arrow points in the direction of maximum change in the value of the associated variable, and the arrow length is proportional to this maximum rate of change. In the perpendicular direction the variable does not change in value.* This is illustrated in Fig. 3 for the variable Source distance, the arrow of which points South-South-East in Fig. 3. The sites are labelled within stream by the rank number of distance from the source. The rank number clearly increased most strongly in the direction indicated by the arrow. (For example, look at the line from L1 to L21). The sites U16–U19 are close and are therefore expected to be at about the same distance from the source. Site L19 is inferred to be at about the same distance as the sites U16–U19, because L19 and U16 do not deviate much in the direction of the arrow. This is verified geometrically by projecting the sites on the arrow. Although at about the same distance from the source, L19 and U16 differ strongly in faunal composition; this difference can be attributed to other environmental variables, notably EC. In the ranking of the projection points, the origin (0,0) indicates the mean of the

variable. In Fig. 3, L1–L10 and U1–U9 are inferred to be at a smaller than average distance from the source, whereas the remaining sites are inferred to be at a larger than average distance. The inference is not always perfect. For example, by definition L20 is farther from the source than L19, but by projecting these sites on the arrow for Distance the opposite is inferred. Generally, an ordination diagram does not display a data-table exactly. It cannot do so, because it uses only two dimensions whereas the data-table is usually multidimensional.

According of row 5 to Table 2, the species points and the environmental arrows jointly represent the species × environmental variable table of weighted averages (sub-table 5 in Fig. 1). This table summarizes the niche centres of the species along each of the environmental variables. The points and arrows again form a biplot. By projecting the species points on the arrow for EC in Fig. 3, we infer, for example, that the species *Dendrocoelum lacteum* and *Erpobdella testacea* have, of all the displayed species, the highest weighted averages for EC and thus occur at high EC values. *Micropsectra fusca* has a higher weighted average for EC than *M. atrofasciata*. In the ranking of projection points, the origin (0,0) indicates the global average of the variable. Thus *M. fusca* largely occurs at higher than average EC values and *M. atrofasciata* at lower than average values. The species close to *Micropterna sequax* have the lowest weighted average for EC; they occur on average at sites with low EC values.

An attractive feature of the diagram is that it takes the method of weighted averaging literally in the following sense. Species points are weighted averages of sites points, not only in the diagram as a whole, but also when projected on to any particular environmental arrow. What we had so far is that the projection points for sites and species display approximate values in sites and approximate weighted averages of species for the corresponding environmental variable. But now we have in addition that the projection points for species are exactly the weighted averages of the displayed environmental values. The method of weighted averaging is thus presented geometrically in the diagram.

According to row 6 of Table 2, the environmental arrows among themselves display the table of correlations among the quantitative environmental variables. This is a derived table (sub-table 6 in Fig. 1) with weighted product-moment correlation coefficients, the weights being the total abundances in sites. *The arrows form a biplot* among themselves: correlations with a particular environmental variable are inferred by projecting on the arrow the arrow heads of the other variables; the order of the projection points then gives the inferred ranking of the correlations. In the ranking, the origin (0,0) indicates zero correlation. Thus, projecting the arrow heads for EC and Distance on the arrow for Discharge shows that Distance is stronger correlated with Discharge than EC, both correlations being positive. An alternative, qualitative rule of interpretation is that the sign of a correlation coefficient between two variables is inferred from the angle between their arrows: if the angle is sharp the correlation is positive, if obtuse, negative (ter Braak, 1987b: 129).

Informally, the length of an environmental arrow indicates the importance of the variable. More formally, (1) the length is equal to the multiple correlation of the variable with the displayed ordination axes and thus indicates how well the values of the variable are displayed in the biplot of sites and environmental variables; this property follows from the facts that the coordinates of the arrow head are correla-

tions with the axes and that the axes are uncorrelated; (2) the length is equal to the maximum rate of change of the variable; variables with short arrows thus do not vary much across the diagram, and (3) the length is equal to the size of the effect that the corresponding variable has on the ordination scores while neglecting other variables (ter Braak, 1994: 140). A later section discusses a method for ranking the relative importance of environmental variables, which is not hindered by the fact that the ordination diagram represents only a two-dimensional view of the species-environment relationship.

Interpretations based on the environmental class points

Rows 7–10 of the body of Table 2 concern interpretations that use the points for classes of qualitative environmental variables. According to row 7 of Table 2, the points for sites and classes jointly represent the table of class memberships of sites. This table is the third of the primary data-tables that CCA was applied to (sub-table 7 in Fig. 1). The data of quantitative and qualitative environmental variables are usually supplied to CCA as a single environmental data-table; for interpretation purposes it is, however, convenient to divide the table. Sites that belong to a particular class are scattered around the class point, simply because, by definition, each class is at the centroid of the sites that it contains. This is yet another application of the centroid principle. The inference is fuzzy; one does not know for sure from the diagram to which class a site belongs.

A class stands for a group of sites. The class point is the weighted mean of the site points that it contains. Therefore the rules given for the interpretation of site points also apply to classes. If the environmental data consist of a single qualitative variable, the points for classes and species in the CCA diagram are identical to those obtained from a correspondence analysis applied to a table of species × classes, the entries of which are the total abundance of each species in each class. If there are more environmental variables, the class points in the CCA diagram are positioned *as if* the CCA had been applied to such a table (i.e. neglecting the other variables). The values of quantitative environmental variables are in this analysis the weighted class means of the quantitative variables (the weights being the total abundances of the sites). This is the clue to the understanding of the rows 8–10 of Table 2.

According to row 8 of Table 2, the points for species and classes jointly represent the table of relative total abundances of species in classes (sub-table 8 in Fig. 1). The interpretation is thus identical to the joint plot of species and site points.

According to row 9 of Table 2, the points for classes and arrows for quantitative environmental variables jointly represent the table of mean values of the quantitative variables in the classes (sub-table 9 in Fig. 1). In fact it is weighted means that are displayed, with the weights being the site totals (y_{i+}). The interpretation is identical to the biplot of site points and environmental arrows.

Inter-class distances (row 10 of Table 2) should be interpreted in this variant of the diagram with the same caution as for the inter-sites distances above (row 3 of Table 2). If the eigenvalues of the axes are of the same magnitude, distances among classes represent the class × class table of chi-square distances (sub-table 10 in Fig. 1), calculated on the basis of the table of total abundances of the species in the classes.

Interpretation via the biplot rule

A surprising, paradoxically feature of this variant of the CCA diagram is that species points together with the site points or the class points can not only be interpreted by the centroid principle, but also by the biplot rule. The diagram is a biplot visualizing transformed abundances (e.g. Greenacre, 1984: 119; ter Braak, 1985) Recently, Greenacre (1993) presented an alternative interpretation of this biplot, namely in terms of the relative abundances $\{y_{ik}/y_{i+}\}$ and $\{y_{ik}/y_{+k}\}$. The biplot rule for this interpretation is as follows. Draw an arrow for a particular species, the kth species, say, by connecting its point with the origin. This arrow points in the direction of maximum change in the relative abundance $\{y_{ik}/y_{i+}\}$ (for a given species k). After projecting the sites on the arrow, the order of projection points thus gives the inferred ranking of the relative abundances $\{y_{ik}/y_{i+}\}$. The biplot thus displays the share that this species has in the total abundance at each site. The role of species and sites can be interchanged in the above rule, thus allowing inference about the relative abundances $\{y_{ik}/y_{+k}\}$. The biplot thus also displays the share each site has in the total abundance of each species. The adjective "fitted" in Table 2 is a reminder that the abundances are fitted to a model based on the environmental data; the abundances are only displayed as far as they are fitted by this model (Lebreton et al., 1991; Box 3).

All tables (except those of rows 2 and 7) in the last column of the Table 2 can be visualized by the biplot rule in the CCA diagram. Because the diagram contains three kinds of entities (sites, species and environmental variables), each pair of which forms a biplot, it is a triplot (ter Braak, 1994; Smilauer, 1992, 1994). In addition, the distance rule can be applied per species, i.e. the diagram is species-conditional. Therefore we propose to name the diagram a "species-conditional CCA triplot". The tables printed in italic in Table 2 are represented optimally as judged by weighted least-squares critera (ter Braak, 1995b). The quality of display of these tables is expressed by the percentage variance accounted (see the legend of Fig. 3).

Ordination diagrams in Hill's scaling

ter Braak (1986, 1987a,b) originally presented another variant of the CCA diagram. This variant used Hill's scaling (Table 2), also used in the program DECORANA (Hill, 1979; Hill and Gauch, 1980). It was the default in older versions of the computer program CANOCO (version 2.1; ter Braak, 1988b). The two main points of difference with the diagram discussed so far are (1) the species scores were standardized to zero weighted mean and a weighted variance of $1/(1-\lambda)$ with λ the eigenvalue of the ordination axis (instead of having variance λ) and (2) site points were weighted averages of species points (instead of being a linear combination of the environmental variables). For a discussion of the second point of difference see Palmer (1993) and ter Braak (1994: 131). The old default is a site-conditional distance diagram (ter Braak, 1994). The centroid principle then implies that the species occurring in a particular site are scattered around the point of that site. In contrast, in a species-conditional diagram, sites that contain a particular species are

scattered around the point of that species. The asymmetry in interpretation (site-conditional versus species-conditonal) went unnoticed in ter Braak (1987a,b), despite early cautionary notes by Oksanen (1987) and Greenacre (1984: 181).

A diagram in Hill's scaling has the advantage that the site scores are expressed in "Standard Deviation units of species turnover" (SD). In this unit, sites that differ by more than about 4 SD in score are expected to have few species in common (ter Braak, 1987b) and the range of the site scores is a measure of beta diversity, termed the length of gradient. In addition, species points are interpreted as optima of response functions; species points could not be weighted averages of the site points, because they would then always fall inside the sampled region, whereas nature is likely to have placed some outside (Hill and Gauch, 1980). ter Braak (1987b: 103, 141) described how to obtain a diagram in Hill's scaling from the CCA algorithm (i.e. from a species-conditional CCA triplot). The properties of a diagram in Hill's scaling are summarized in Table 2. The only valid least-squares biplot is that of species points and environmental arrows visualizing weighted averages.

The species-conditional CCA triplot is valid for more purposes (Table 2), it is easier to use and it gives a more direct display of the weighted averaging principle underlying CCA than a diagram in Hill's scaling. Because environmental arrows are displayed by correlations, the species-conditional CCA triplot is easier to interpret quantitatively. These were the main reasons for adopting the species-conditional CCA triplot as the standard in CANOCO 3.1.

Practical points

A CCA diagram does not need to contain all the elements (species, sites, environmental variables). To avoid overcrowding of points, species and sites are often shown in separate diagrams that can, in principle, be overlain. Alternatively, selected points or variables are displayed. Selection is based on personal judgement, or is based on number of occurrence, total abundance, tolerance, or percentage fit.

The quality of the ordination diagram in displaying some of the tables in Fig. 1 (goodness-of-fit) is best described in the legend of the diagram (ter Braak, 1994). Each eigenvalue of CCA can be converted to a percentage variance accounted for by dividing the eigenvalue ($\times 100$) by the total inertia of the abundance data, inertia being a measure of weighted variance that is closely related to the chi-square statistic (Greenacre, 1984). This usage of the eigenvalues is not obvious from the ecological derivation of CCA. This usage derives from CCA as a weighted form of redundancy analysis (Sabatier et al., 1989; Box 3). For ecological data, the percentage-explained inertia is typically low ($<10\%$), especially for strong gradients. This is nothing to worry about; it is an inherent feature of data with a strong presence/absence aspect. As in applications of binary logit regression (Jongman et al., 1995), the percentage-explained is not very informative, and is probably best left unreported. Apparently, the importance of extracted gradients must be decided upon by other means. Decision criteria include the magnitude of the eigenvalues themselves (as a rule of thumb, eigenvalues >0.30 indicate strong gradients), the statistical significance as judged by Monte Carlo permutation tests and, even more importantly, the ecological interpretability. Each eigenvalue of CCA can also be

divided by the sum of all CCA eigenvalues and converted to a percentage. This percentage has two interpretations: (1) it is the percentage variance accounted for relative to the inertia of the fitted abundance values, and (2) the percentage variance accounted for relative to the total variance in the species × environment tables (sub-tables 5 and 8 in Fig. 1). These tables summarize the species × environment relations. In conclusion, the legend of the ordination diagram should contain the values of the eigenvalues of the axes and the percentage accounted for of the variance in the weighted averages and class totals (sub-tables 5 and 8 in Fig. 1).

In the example data, the first two eigenvalues are 0.35 and 0.17, the total inertia is 4.0, whereas the sum of all CCA eigenvalues is 0.75. Fig. 3 thus displays $100 \times (0.35 + 0.17)/4.0 = 13\%$ of the total inertia and $100 \times (0.35 + 0.17)/0.75 = 69\%$ of the variance in the weighted averages and relative class totals of these data. Consequently, Fig. 3 is not very faithful in displaying the observed abundances, but reasonably faithful in displaying the fitted abundance values, weighted averages and class totals.

Although always applicable, the centroid principle is of limited use if CCA does not strongly separate the species niches. As a rule of thumb the eigenvalues should be at least 0.4. The first two eigenvalues for the example data (0.35 and 0.17) are thus on the small side for interpretation via the centroid principle. The distance rule applies in so far as CCA is a good approximation to the fitting of (circular) bell-shaped response surfaces with the species scores being the optima (ter Braak, 1986). For example, by fitting a Gaussian response surface across the diagram for *Prodiamesa olivacea*, we found that its optimum lies inbetween L9 and L20, rather far from the point for this species in Fig. 3; also the tolerance is large. The fitting of Gaussian surfaces is a standard feature of the computer program CanoDraw (Smilauer, 1992, 1994).

If the eigenvalues are small, it is often attractive to magnify the configuration of species points with respect to that of the samples. If configurations or diagrams are in different scale units, the centroid and distance rules can no longer be used, but the biplot rule can still be used in the CCA biplot or triplot. The biplot rule appears more informative than the centroid rule with small eigenvalues, whereas the centroid rule and the distance rule appear more informative when unimodality is strong, as indicated by large eigenvalues (> 0.4, say) or large lengths of gradients (> 4 SD).

Ranking environmental variables in importance

It is often of interest to rank environmental variables in their importance for determining the species composition. A related aim is to reduce a large set of variables to a smaller set that suffices to explain the variation in species composition. Environmental variables can be ranked and selected in CCA in very much the same ways as predictors can be ranked and selected in (multivariate) regression. The reason is that CCA is a form of multivariate linear regression on transformed data (Sabatier et al., 1989; Lebreton et al., 1991; ter Braak et al., 1993; Box 3). The species and environmental variables take the roles of response variables and predictor variables, respectively. This does not mean that CCA and multivariate regression would yield identical rankings. CCA aims to explain the variation in the species

composition, i.e. in relative abundance values, whereas linear regression and related linear methods such as redundancy analysis and PLS aim to explain the variation in absolute abundance values.

Following suggestions by Escoufier and Robert (1979), the computer program CANOCO version 3.1 offers the method of forward selection. In the first step of this method, all environmental variables (including classes of qualitative variables) are ranked on the basis of the fit for each separate variable. The measure of fit is the first (and only) eigenvalue of the CCA *with each one variable as the only environmental variable.* Recall from Box 2 that the eigenvalue measures niche separation. The first column of Table 3 gives an example. For example, if CCA is applied to the example macrofauna data with EC as the only environmental variable, the first eigenvalue would be 0.20. The single variable giving the highest eigenvalue is the class variable Shrubs ($\lambda = 0.25$). The statistical significance of the effect of each variable is tested by a Monte Carlo permutation test (See Manly (1991) and ter Braak (1992) for an explanation of such tests) and the resulting significance level is given in Table 3 (step 1). At the 5% level, eight of the environmental variables are significantly related to the species data.

At the end the first step of the forward selection the best variable, here Shrubs, is selected. Hereafter, all remaining environmental variables are ranked on the basis of the fit that each separate variable gives in conjunction with the variable(s) already selected. The measure of fit is the sum of all eigenvalues of the CCA *with each variable as the only additional environmental variable.* The program reports the "extra fit", which is the change in the sum of all eigenvalues of CCA if the associated variable would be selected. (Eigenvalues of CCA are usually termed canonical eigenvalues to distinguish them from eigenvalues of correspondence analysis; see below). In the example, Source distance is the variable giving the highest change

Table 3. Ranking environmental variables in importance by their marginal (left) and conditional (right) effects of the macrofauna in the example data-set (Table 1), as obtained by forward selection. (λ_1 = fit = eigenvalue with variable j only; λ_a = additional fit = increase in eigenvalue; cum (λ_a) = cumulative total of eigenvalues λ_a; P = significance level of the effect, as obtained with a Monte Carlo permutation test under the null model with 199 random permutations; – additional variables tested; veg. = vegetation). Seasonal variation is partialled out by taking the month class variables as covariables

marginal effects (forward: step 1)				conditional effects (forward: continued)				
j	variable	λ_1	P	j	variable	λ_a	P	cum (λ_a)
1	Shrubs (1/0)	0.25	(0.01)	1	Shrubs	0.25	(0.01)	0.25
2	Source distance	0.22	(0.01)	2	Source distance	0.19	(0.01)	0.44
3	EC	0.20	(0.01)	3	Discharge	0.19	(0.01)	0.63
4	Discharge	0.17	(0.01)	4	EC	0.14	(0.03)	0.75
5	Total cover of veg.	0.16	(0.01)					
6	Shading	0.15	(0.01)	–	Cover emergent veg.	0.11	(0.10)	–
7	Soil grain size	0.14	(0.02)	–	Cover bank veg.	0.11	(0.12)	–
8	Stream width	0.14	(0.05)	–	Soil grain size	0.10	(0.13)	–
9	High weedy veg.	0.14	(0.08)					
10	Cover bank veg.	0.13	(0.11)					
–	U vs L stream	0.22	(0.01)	–	U vs L stream	0.09	(0.26)	–

(0.19). Notice that when taken singly (column 1 of Table 3) Source distance has a somewhat higher eigenvalue. This is because part of the effect of Distance is already explained by the variable Shrubs. The extra fit gives the conditional effect of Distance (namely given Shrubs), whereas the value in the first column gives the marginal effect, i.e. ignoring the other variables. The conditional effect is statistically significant ($P < 0.01$) as judged on the basis of a Monte Carlo test (199 random permutations). So in the second step, the variable Distance is selected.

The third and later steps in the forward selection proceed in the same way as the second one. In the example, the third and fourth best variables are Discharge and EC, respectively. Both have significant conditional effects. Notice the change in order compared with the marginal effects. The fifth variable to be selected, Cover of emergent plants, is not statistically significant ($P > 0.11$), neither are other variables with an extra fit of comparable magnitude.

The stream name (U-L) was not used as selectable predictor variable, because we were interested in which measured variables could account for the differences in macrofauna composition among the streams. The class variable stream name would be ordered second among the marginal effects with an eigenvalue of 0.22. As judged by the Monte Carlo test, macrofauna composition differed significantly among the streams ($P < 0.01$). After selecting four variables (Table 3), stream name could contribute 0.09 to the sum of the eigenvalues, but the additional effect was non-significant ($P = 0.26$). In conclusion, the four selected variables well explained the differences in macrofauna compositon among streams. The CCA ordination diagram with these variables is shown in Fig. 3.

The Monte Carlo tests replace the usual *F*- or *t*-tests in forward selection in multiple regression. The Monte Carlo test does not require the assumptions of normality. None of these tests controls the overall type I error. See Miller (1990: 50) for a discussion of this point and for a rough Bonferoni-type adjustment. In practical terms, this means that variables that are irrelevant will too easily be judged significant.

In the forward selection example (Table 3) the month class variables were specified as covariables. The seasonal variation was thus already accounted for. It is of some interest to compare the amount of seasonal variation with that of the environmental variation (cf. Sabatier et al., 1989). The sum of eigenvalues associated with months only is 0.58. The four selected variables add another 0.75. The environmental variation is thus of the same order of magnitude as the seasonal variation.

Because the sampling design is (nearly) balanced, each source of variation has an (almost) unique associated amount of variation. Variance decomposition for unbalanced data, i.e. in a general regression situation, is very interesting, but more complicated (Borcard et al., 1992; Økland and Eilertsen, 1994).

Relationships with other multivariate methods

Relationships with discriminant analysis

Canonical correspondence analysis has an early precursor in the ecological literature in the form of Green's (1971, 1974) multi-group discriminant analysis for quan-

tifying the multivariate Hutchinsonion niche of species. Green's method appeared rather ad-hoc and lacked a solid statistical basis (James and McCulloch, 1990). After critical discussions on particular proposals for measuring niche breadths (summarized by Carnes and Slade, 1982) interest in Green's method was lost, ironically in the same period in which the ordination method of correspondence analysis surged in popularity. At the time nobody recognized the relationship between the methods; the aims and domains of applications were different. Chessel et al. (1987) and Lebreton et al. (1988a) recognized the formal equivalence between canonical correspondence analysis and discriminant analysis on reformatted data (see also Takane, Yanai and Mayekawa, 1991). The details are as follows.

The derivation of canonical correspondence analysis is very similar to that of multi-group (linear) discriminant analysis, alias canonical variate analysis. Multiple discriminant analysis (Rao, 1952; Krzanowski, 1988; McLachlan, 1992) works on measurements of features on individuals belonging to different groups. The usual aim is to assign new individuals with unknown group membership to groups on the basis of the measured features. It is often convenient for explorative purposes to see whether the groups can be discriminated in less dimensions, i.e. on a few synthetic features. For this, the method finds canonical variates, that are linear combinations of the features that show maximum discrimination among groups, or in other words, that maximally separate the groups. Plotting the scores of the individuals on the first two canonical variates helps to see how well the groups can be discriminated. Replacement of "groups" by "niches of species" in the above yields similar definitions for discriminant analysis and canonical correspondence analysis. But there is an important difference: with discriminant analysis, the features of individuals are measured, whereas with canonical correspondence analysis, it is the (environmental) features of sites that are measured. Suppose now that the species data for CCA are counts of individuals at sites. Then the link between the methods can be completed by treating each individual counted as a separate unit, i.e. as a separate row in the data-table; see Lebreton et al. (1988a) for an example. The data for each individual counted are then the species to which it belongs and the measurements of the features of the site at which it occurs. Multi-group discriminant analysis carried out on data brought in this form is identical to canonical correspondence analysis (Lebreton et al., 1988a). There are minor differences in the default output. For example, if the eigenvalue of canonical correspondence analysis is λ, then the corresponding eigenvalue of the discriminant analysis if $\lambda/(1-\lambda)$ (ter Braak, 1988b: section 9.5); the scores of species and of the sites are linearly related. The scaling of the scores as used in discriminant analysis is a variant of Hill's scaling in (canonical) correspondence analysis (ter Braak, 1988b; Jongman et al., 1995: 103). It has the advantage that the mean within-species variance is equalized across dimensions, but also some disadvantages for quantitative interpretation of the ordination diagram. The difference with the standard Hill scaling (the first column of Table 2) is that the standard ordination diagram of discriminant analysis is not site-conditional but species-conditional (species points as weighted averages of site points).

Green (1974) proposed a multivariate niche analysis with temporally varying environmental factors. Our macrofauna data sampled in different months form an obvious example. The analysis by Green (1974) is identical to a partial canonical correspondence analysis applied to presence/absence data.

In summary, the main difference between CCA and discriminant analysis is that the unit of the statistical analysis in discriminant analysis is the individual, whereas it is the site in CCA. This is important for the way in which statistical tests need to be carried out. The statistical tests designed for discriminant analysis, as used by Green (1972, 1974), are invalid in the context of CCA, because these ignore the grouping of individuals within sites. Valid statistical tests can be based on Monte Carlo permutation of sites (instead of individuals) and are standard in the computer program CANOCO (ter Braak, 1988b).

Relationships to correspondence analysis (CA)

If in a particular study only biological assemblage data were collected, CCA cannot be applied. Nevertheless, one might want to construct a hypothetical, synthetic variable that maximises niche separation. This is what correspondence analysis does; it constructs the best variable "out of blue water" (from the species data only). In contrast, CCA constructs the best, synthetic variable by linearly combining the measured environmental variables. This has the advantage that the environmental basis of the ordination is guaranteed in CCA. There is one snag to this guarantee: if there are almost as many environmental variables as sites, the environmental basis may become very unstable or nonsensical, and CA and CCA produce about the same site and species ordination (ter Braak, 1986, 1987a). This is because CCA considers all linear combinations of the many variables and therefore has almost as much freedom as CA to construct the best variable, if there are many environmental variables compared to the sample size. The distinction between CA and CCA is thus nontrivial only with far fewer environmental variables than sites.

In the example data, there were 40 sites and 25 environmental variables. With this high number of environmental variables compared to the sample size, there is a great danger that CCA produces a noninterpretable environmental basis. This was the main reason for carrying out the forward selection of environmental variables. The selection reduced the number of environmental variables to a manageable number and, in addition, gave additional information on the importance of each of the variables.

Relationships to two-block PLS

PLS (Partial Least Squares Projection to Latent Structure) can be applied to the same type of primary input data as CCA (Fig. 1). If applied to two data-tables, PLS yields a model for predicting one data-table from the other. PLS originated with Herman Wold (1982) in econometrics as a "poor-man's-alternative" to structural equation modelling (e.g. LISREL, Saris and Stronkhorst, 1984; Lohmuller, 1988). It was developed by Harald Martens and Svante Wold for calibration and prediction in chemometrics (Geladi, 1988). In these later developments the earlier mode A, B and C methods (H. World, 1982) were integrated in a smart way into a single algorithmic framework. Cross-validation became an important tool. The following discussion applies to PLS as used in chemometrics (Höskuldsson, 1988;

Martens and Naes, 1989). After listing some similarities between CCA and PLS, we list the most importance dissimilarities among the methods. This section concludes with the key ideas needed for integrating members of the correspondence analysis family into the PLS framework.

CCA shares several properties with PLS. The methods are both asymmetric: species abundance is modelled as a function of the environmental variables. Both are thus regression methods. Both methods use the ideas of latent variables (the ordination axes, components or synthetic gradients), dimension reduction and associated graphical display (the ordination diagram). In both methods the latent variables are linear combinations of the environmental variables. Both methods are suited to analyze uncorrelated environmental variables, or variables that show a moderate amount of correlation. Both methods can meaningfully analyze any number of species, irrespective of the sample size n (the number of sites). The association or correlation (+ or –) among species may be arbitrarily high without affecting the usefulness of the results. The prize paid for this is that neither method is invariant to linear transformations of the species variables. If the abundance measurements are not commensurate (counts for one species, biomass for another, for example), the units in which each species abundance is expressed, need careful consideration in both methods. In CCA, one may want to equalize the abundance total per species (divide by the abundance total of each species). In PLS, species could be standardized to zero mean and unit variance (autoscaling).

CCA differs in some important aspects from PLS. Most importantly, the model underlying CCA is unimodal, whereas it is linear in PLS. The response data in CCA must be nonnegative; the data are abundances (e.g. counts or presence-absence) or compositional data, in the sense that only relative values are meaningful (ter Braak, 1988a, 1995a, b). Typical response data in PLS are quantitative (be they positive or negative, without special meaning attached to the value 0). PLS shares these differences with redundancy analysis (RDA), the linear analogue of CCA (ter Braak and Prentice, 1988; ter Braak, 1994). PLS is identical to RDA, if the predictor variables are uncorrelated, as in many designed experiments (e.g. Data set III in Eriksson et al., 1995).

PLS is a biased regression method. It aims primarily at prediction. CCA and RDA are based on unbiased regression. By providing a least-squares fit, they aim primarily at explanation and efficient description. This is also the difference between PLS and multiple regression (Eriksson et al., 1995). However, unbiased regression methods predict poorly if the predictor variables are very highly correlated (multicollinear), which happens trivially if the number of environmental variables (p) is of the same order of magnitude as the sample size n. PLS contains a special guard against multicollinearity. In this sense PLS is akin to ridge regression (de Jong and Farebrother, 1994). The first PLS component acts as if the environmental variables were uncorrelated (Frank and Friedman, 1993), and later components bring in more of the correlations among variables. If the number of components is maximal, PLS is identical to unbiased multivariate regression. The crux of PLS is the selection of the number of components so as to minimize the prediction error. This is done by cross-validation. The chosen number of components minimizes the prediction error as estimated by cross-validation. In contrast, the environmental weights of the first component (axis) in CCA and RDA are

already adversely affected by multicollinearity among the environmental variables. If environmental variables are highly correlated in CCA and RDA, the weights become instable and uninterpretable (but nevertheless the axes remain stable). For this reason, it is standard practice to abstain from interpreting weights and to focus on the correlations of the environmental variables with the axes as in Table 2 and Fig. 3. These still indicate how individual variables influence the species. The first two components of RDA generally extract more variance of the species data than the first two components of PLS. The ordination diagram thus displays (describes) more of the data. But, under the conditions that are favourable for PLS (notably if $p > n$), some or all of the displayed correlations with the environment may be spurious. The possibilities and limitations of interpretation of CCA are then precisely those of an indirect gradient analysis (carrying out a CA on the species data and subsequently interpreting the components in terms of the environmental data; see also the previous section on the relation with CA). With many environmental variables, there is a real danger of over-interpretation. Some statisticians require here the application of simultaneous testing procedures to counter the danger. The solution that PLS offers is to focus on prediction and associated procedures of cross-validation, rather than on statistical significance. With CCA and RDA, the solution must be sought by first invoking other methods that reduce the number of environmental variables. We used forward selection of variables in the example. A possibility that is more in line with the ideas of PLS and principal component regression (PCR), is to apply a preliminary principal component analysis to the environmental data and to treat the first few components of this analysis as the new environmental variables in CCA (Ruse, 1994). Most biologists are, however, not really interested in species relations to abstract environmental variables like principal component axes. This objection can be alleviated by adding the original environmental variables afterwards to the ordination diagram. This can be done in such a way that the biplot interpretations given in Table 2 continue to hold true. The program CANOCO contains facilities for carrying out the required analyses. We did not do so in the example; it might have increased the predictive properties of the analysis, but would have decreased our understanding of the major variables in the system.

Dimensionality often plays a different role in CCA and PLS. CCA and RDA aim to visualize the data in an ordination diagram. Two-dimensional diagrams are the easiest to construct and to inspect, leading to a strong bias for using two dimensions only. This may be too many or too few. To guard against interpretation of spurious axes, tests of statistical significance can be used. With the program CANOCO, the statistical significance of the first axis can be tested by a Monte Carlo permutation test. The significance of each additional axis can be judged similarly by carrying out a partial CCA with the previously tested axes as (extra) covariables. In the example data, both axes displayed in Fig. 3 were statistical significant ($P < 0.01$). Two more axes were significant and thus potentially contained interesting structure. These axes added detail to the main structure displayed in Fig. 3. We decided not to display the extra axes, because the effects that we wanted to demonstrate are already convincingly displayed in Fig. 3. The seasonal effects (treated as covariables) were not displayed either for lack of space. The use of covariables allowed us to display the effects of prime interest in two dimensions. The bias towards the usage of one or two dimensions probably also applies to PLS,

as far as PLS is used for producing ordination diagrams (such as Fig. 15 in Eriksson et al., 1995). More commonly, however, PLS focuses on prediction. Then, dimensions are added as long as they increase the predictive power of the model. Because a computer program is used to make the predictions, there is no other limit to the number of dimensions.

A technical difference between PLS and CCA is that CCA is invariant to linear transformations of the environmental variables. The display of standardized variables in the ordination diagram is for convenience; it does not affect other aspects of the analysis. In PLS, standardization (autoscaling) of variables has a nontrivial effect on the result.

As shown in the appendix, it is quite straightforward to combine the best of the worlds of PLS and CCA. The key ingredients for this are presented in ter Braak et al. (1993): whereas PLS selects linear combinations of species variables with certain optimal properties, CCA-PLS takes weighted averages with certain optimal properties. In principle, the method does not require a special computer program. We hope to give more details and an example elsewhere.

The aim of CCA-PLS is to predict the species data from the environmental data in accordance with the supposed causal flow: species respond to the environment. For the purpose of multivariate species-environment calibration, ter Braak et al. (1993) turned the problem upside down: the environment was to be predicted from the species data. For this aim, the ideas of correspondence analysis, weighted averaging and PLS were combined into a technique called Weighted Averaging Partial Least Squares (WA-PLS). The predictive power of this method was demonstrated on real and simulated data by ter Braak and Juggins (1993), ter Braak et al. (1993) and ter Braak (1995a).

Relationships to co-inertia analysis

Dolédec and Chessel (1994) proposed co-inertia analysis as a simple method for analyzing species-environment data with many species and many environmental variables. Co-inertia avoids the problems that CCA has with many environmental variables by totally disregarding the correlations among environmental variables. The methods amounts to an analysis of the species-environment sub-tables (Table 2) by a singular value decomposition. For the comparison of the singular value decomposition of such cross-product tables with PLS see de Jong and ter Braak (1994). The first axis of co-inertia analysis is identical to that of CCA-PLS. In contrast to PLS and CCA, co-inertia treats species and environmental data in a quite symmetric way; it analyzes covariation; there are neither regression models nor prediction models involved. From co-inertia ordination diagrams, one can infer about weighted averages and relative class totals, and less about the other tables. In co-inertia analysis, the more members a group of correlated environmental variables contains, the more the group is emphasized. Group size has less influence in CCA and CCA-PLS.

Discussion

The example data were collected to study the direct and indirect effects of intensive agricultural land-use on the macro-invertebrates in two morphologically similar streams. Agricultural land-use implies soil fertilization and thus eutrophication, here expressed by the electrical conductivity (EC). But it also implies increased run-off through improved drainage. As is well-known, increased discharge fluctuations have a major effect on the macro-invertebrate community through changes in discharge regime, current velocities, morphological structure of bottom and banks and erosion and siltation. The Uddel stream had, on average, a higher discharge and much stronger discharge fluctuations than the Leuvenum stream. Our analysis (Table 3) confirmed the effect of discharge on the macro-fauna community and demonstrated the additional effect of EC. Despite eutrophication and disturbed hydraulics, both streams showed a gradient from source to mouth, as indicated by the variable Source distance.

In classic pollution studies, the response of macro-invertebrates to eutrophication was mostly indicated by the use of a diversity, saprobic or biotic index (Washington, 1984). Studies on the relation of macro-invertebrates and discharge regimes or distances to source, reviewed by Hawkes (1975), mostly used non-numerical methods to highlight river zonation and to arrange species to zone classes. Some authors also used numerical methods, in particular, cluster analysis, for this. CCA allows eutrophication and zonation to be studied simultaneously. It produces an ordination diagram in which species, sites and environmental variables are arranged in a single diagram. The diagram serves to represent concisely the main results.

Canonical correspondence analysis and other members of the correspondence analysis family have their own niche in the space of available multivariate methods. Their usage is recommended if two or more of the following criteria are satisfied: (1) relationships are unimodal, (2) the data have positive values, but contain many zeroes or (3) the data are compositional in the sense that relative values are relevant to the problem. For many ecological data-sets in the aquatic sciences at least two of these criteria are fulfilled. Criterion (2) nearly always applies, but now suppose that trends in absolute abundance values are relevant to the problem at hand. Because of the zeroes, non-linear or generalized linear models are then required, but their multivariate extensions are not yet available for routine application. Canonical correspondence analysis is available and can take care of the nonlinearity caused by the zero values, but focuses on relative abundance values. Our suggestion is then to apply univariate regression to analyze the total abundance across species and to apply canonical correspondence analysis for the analysis of the community composition.

Appendix: On the PLS form of CCA

It is quite straightforward to combine the best of the worlds of PLS and CCA. The key ingredients for this are presented in ter Braak et al. (1993): whereas PLS selects linear combinations of species variables with certain optimal properties. CCA-PLS

takes weighted averages with certain optimal properties. In the following matrix algebra, sites correspond to rows (in contrast to Fig. 1). Let $\mathbf{X}^*$ and $\mathbf{Y}^*$ denote the predictor matrix and response matrix, respectively. In particular, the first component of PLS selects linear combinations of the environmental variables and of the species data, $\mathbf{t}^*_E = \mathbf{X}^* \mathbf{w}^*$ and $\mathbf{t}^*_S = \mathbf{Y}^* \mathbf{c}^*$, respectively, that have maximum covariance, subject to the constraints $\mathbf{w}^{*\prime}\mathbf{w}^* = \mathbf{c}^{*\prime}\mathbf{c}^* = 1$. The second and further components also maximize the covariance, but subject to the constraint that both the new $\mathbf{t}^*_E$ and the new $\mathbf{t}^*_S$ are orthogonal to the environmental components $\{\mathbf{t}^*_{E1}, \mathbf{t}^*_{E2}, \ldots\}$, that are already extracted (the orthogonality requirement replaces the calculation of residual matrices at each step in the usual PLS algorithms (Martens and Naes, 1989); note that orthogonality must be with respect to the environmental components, hence the asymmetry in PLS, see de Jong and ter Braak, 1994). For defining CCA-PLS, let $\mathbf{X}$ and $\mathbf{Y}$ denote the environmental data and species data, respectively, and let $\mathbf{R} = \mathrm{diag}(y_{1+}, \ldots, y_{n+})$ and $\mathbf{K} = \mathrm{diag}(y_{+1}, \ldots, y_{+m})$ with y_{i+} and y_{+k} the total abundance in site i and per species k, respectively. The first component of CCA-PLS can now be defined as selecting the linear combination of the environmental data, $\mathbf{t}_E = \mathbf{X}\mathbf{w}$, and the weighted average of the species data, $\mathbf{t}_S = \mathbf{R}^{-1}\mathbf{Y}\mathbf{u}$, that have maximum covariance in the metric defined by $\mathbf{R}$, i.e. maximum $\mathbf{t}_E'\mathbf{R}\,\mathbf{t}_S$, subject to the constraints $\mathbf{w}'\mathbf{w} = \mathbf{u}'\mathbf{K}\mathbf{u} = 1$. The second and further components also maximize the covariance, but subject to the constraint that both the new $\mathbf{t}_E$ and the new $\mathbf{t}_S$ are $\mathbf{R}$-orthogonal to the environmental components that are already extracted. As in PLS, the maximization to obtain the subsequent components in CCA-PLS amounts to solving subsequently for the first singular vectors of certain cross-product matrices. Alternatively, a NIPALS algorithm akin to the iterative algorithm of CCA (ter Braak, 1986) can be used. Fortunately, no special computer program is required to obtain the solution. The following three steps yield the solution: (1) preprocess $\mathbf{X}$ and $\mathbf{Y}$ as in ter Braak et al. (1993: (2.5)), i.e. $\mathbf{R}$-centre $\mathbf{X}$, so that $\mathbf{1}'\mathbf{R}\mathbf{X} = \mathbf{0}$, and calculate

$$\mathbf{Y}^* = \mathbf{R}^{-1/2}\mathbf{Y}\,\mathbf{K}^{-1/2} \text{ and } \mathbf{X}^* = \mathbf{R}^{1/2}\mathbf{X}, \tag{A.1}$$

and (2) carry out a PLS2 without additional centring or standardization with $\mathbf{Y}^*$ as response matrix and $\mathbf{X}^*$ as predictor matrix so as to obtain for each component the X-scores $\mathbf{t}^*_E$, the Y-scores $\mathbf{t}^*_S$, the X-weights $\mathbf{w}^*$, the Y-weights $\mathbf{c}^*$ and the X-loadings $\mathbf{p}^*$, the Y-loadings $\mathbf{q}^*$, and (3) postprocess these results so as to obtain for each component the corresponding entities for CCA-PLS

$$\mathbf{u} = \mathbf{K}^{-1/2}\mathbf{c}^*,\ \mathbf{q} = \mathbf{K}^{-1/2}\mathbf{q}^* \text{ and } \mathbf{t}_S = \mathbf{R}^{-1/2}\mathbf{t}^*_S, \tag{A.2}$$

and

$$\mathbf{w} = \mathbf{w}^*,\ \mathbf{p} = \mathbf{p}^* \text{ and } \mathbf{t}_E = \mathbf{R}^{-1/2}\mathbf{t}^*_E. \tag{A.3}$$

The proof is analogous to that in the Appendix of ter Braak et al. (1993). By reformatting the data as in the section on the relation of CCA with discriminant analysis, CCA-PLS can also be obtained as the PLS version of discriminant analysis (DA-PLS, i.e. PLS in which the response matrix is an indicator matrix of the group memberships).

ACKNOWLEDGEMENTS

We would like to thank H.J.B. Birks, P.J. van den Brink, H.F. van Dobben, H. van der Voet, P. Smilauer and M. Stapel for comments on the manuscript. This research profitted from discussions with J.-D. Lebreton,, D. Chessel and Y. Escoufier, made possible with financial support of the CNRS and ENSAM. Figure 3 was prepared with CanoDraw (Smilauer, 1992).

REFERENCES

Anderson, N.J., 1993. Natural versus anthropogenic change in lakes: the role of the sediment record. TREE 8:356–361.

Anderson, N.J., T. Korsman and I. Renberg, 1994. Spatial heterogeneity of diatom stratigraphy in varved and non-varved sediments of a small, boreal-forest lake. Aquat. Sci. 56:40–58.

Anderson, N.J., B. Rippey and C.E. Gibson, 1992. A comparison of sedimentary and diatom-inferred phosphorus profiles: implications for defining pre-disturbance nutrient conditions. Hydrobiologia 253:357–366.

Austin, M.P. and M.J. Gaywood, 1994. Current problems of environmental gradients and species response curves in relation to continuum theory. J. Veg. Sci. 5:473–482.

Austin, M.P., A.O. Nicholls, M.D. Doherty and J.A. Meyers, 1994. Determining species response functions to an environmental gradient by means of a β-function. J. Veg. Sci. 5: 215–228.

Bakker, C., P.M.J. Herman and M. Vink, 1990. Changes in seasonal succession of phytoplankton induced by the storm-surge barrier in the Oosterschelde (S.W. Netherlands). J. Plankton Res. 12:947–972.

Barker, P., 1994. Book review of "H. van Dam (Editor). Twelfth International Diatom Symposium. Kluwer, Academic Publ. Dordrecht". Eur. J. Phycol. 29:281–283.

Birks, H.J.B., S. Juggins and J.M. Line, 1990a. Lake surface-water chemistry reconstructions from palaeolimnological data. In: Mason B.J. (ed.), The Surface Waters Acidification Programme, Cambridge University Press, Cambridge, pp. 301–313.

Birks H.J.B., J.M. Line, S. Juggins, A.C. Stevenson and C.J.F. ter Braak, 1990b. Diatoms and pH reconstruction. Phil. Trans. Roy. Soc. London, Ser B 327:263–278.

Birks, H.J.B., S.M. Peglar and H.A. Austin, 1994. An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986–1993, Botanical Institute, Bergen, Norway, 58 pp.

Borcard, D., P. Legendre and P. Drapeau, 1992. Partialling out the spatial component of ecological variation. Ecology 73:1045–1055.

Carnes, B.A. and N.A. Slade, 1982. Some comments on niche analysis in canonical space. Ecology 63:888–893.

Carpenter, S.R., T.M. Frost and D.K.T.K. Heisey, 1989. Randomized intervention analysis and the interpretation of whole-ecosystem experiments. Ecology 70:1142–1152.

Charles, D.F. and J.P. Smol, 1994. Long-term chemical changes in lakes: quantitative inferences from biotic remains in the sediment record. In: Baker L. (ed.), Environmental Chemistry of Lakes and Reservoirs, American Chemical Society, Washington, pp. 3–31.

Chessel, D., J.-D. Lebreton and N. Yoccoz, 1987. Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie. Rev. Statist. Appl. 35:55–72.

Copp, G.H., 1992. An empirical model for predicting microhabitat of 0+ juvenile fishes in a lowland river catchment. Oecologia 91:338–345.

Cumming, B.F., J.P. Smol and H.J.B. Birks, 1992. Scaled chrysophytes (Chrysophyceae and Synurophyceae) from Adirondack drainage lakes and their relationship to environmental variables. J. Phycol. 28:162–178.

de Jong, S. and R.W. Farebrother, 1994. Extending the relationship between ridge regression and continuum regression. Chemometrics Intell. Lab. Syst. 25:179–181.

de Jong, S. and C.J.F. ter Braak, 1994. Comments on the PLS kernel algorithm. J. Chemometrics 8:169–174.

Descy, J.P., 1979. A new approach to water quality estimation using diatoms. Nova Hedwigia, Beiheft 64:305–323.

Dolédec, S. and D. Chessel, 1994. Co-inertia analysis: an alternative method for studying species-environment relationships. Freshwater Biol. 31:277–294.
Ellenberg, H., 1948. Unkrautgesellschaften als Mass für den Säuregrad, die Verdichtung und andere Eigenschaften des Ackerbodens. Ber. Landtech. 4:130–146.
Eriksson, L., J.L.M. Hermens, E. Johansson, H.J.M. Verhaar and S. Wold, 1995. Multivariate analysis of aquatic toxicity data. Aquat. Sci. this volume.
Escoufier, Y. and P. Roberts, 1979. Choosing variables and metrics by optimizing the RV-coefficient. In: Rustagi J.S. (ed.), Optimizing methods in Statistics, Academic Press, New York, pp. 205–219.
Fairchild, G.W. and J.W. Sherman, 1993. Algal periphyton response to acidity and nutrients in softwater lakes: lake comparison vs. nutrient enrichment approaches. J. N. Am. Benthol. Soc. 12:157–167.
Frank, I.E. and J.H. Friedman, 1993. A statistical view of some chemometric regression tools (with discussion). Technometrics 35:109–148.
Fritz, S.C., S. Juggins and R.W. Batterbee, 1993. Diatom assemblages and ionic characterization of lakes of the northern great plains, North America – a tool for reconstructing past salinity and climate fluctuations. Can. J. Fish. Aquat. Sci. 50:1844–1856.
Gabriel, K.R., 1982. Biplot. In: Kotz S. and N.L. Johnson (eds.), Encyclopedia of Statistical Sciences, Vol. 1, Wiley, New York, pp. 263–271.
Gabriel, K.R. and C.L. Odoroff, 1990. Biplots in biomedical research. Statist. Med. 9:469–485.
Gauch, H.G., 1982. Multivariate analysis in community ecology. Cambridge University Press, Cambridge, 298 pp.
Gause, G.F., 1930. Studies on the ecology of the Orthoptera. Ecology 11:307–325.
Geladi, P., 1988. Notes on the history and nature of partial least squares (PLS) modelling. J. Chemometrics 2:231–246.
Gower, A.M., G. Myers, M. Kent and M.E. Foulkes, 1994. Relationships between macroinvertebrate communities and environmental variables in metal-contaminated streams in south-west England. Freshwater Biol. 32:119–221.
Grantham, B.A. and B.J. Hann, 1994. Leeches (Annelida, Hirundinea) in the experimental lakes area, Northwestern Ontario, Canada – Patterns of species composition in relation to environment. Can. J. Fish. Aquat. Sci. 51:1600–1607.
Green, R.H., 1971. A multivariate statistical approach to the Hutchinsonian niche: bivalve mollucs of central Canada. Ecology 52:543–556.
Green, R.H., 1974. Multivariate niche analysis with temporally varying environmental factors. Ecology 55:73–83.
Green, R.H., 1979. Sampling design and statistical methods for environmental biologists. Wiley, New York, 257 pp.
Greenacre, M.J., 1984. Theory and applications of correspondence analysis, Academic Press, London, 364 pp.
Greenacre, M.J., 1989. The Carroll-Green-Schaffer scaling in correspondence analysis: a theoretical and empirical appraisal. J. Marketing Research 26:358–365.
Greenace, M.J., 1993. Biplots in correspondence analysis. J. Appl. Statist. 20:251–269.
Hawkes, H.A., 1975. River zonation and classification. In: Whitton B.A. (ed.), River Ecology. Studies in ecology, vol. 2, Univ. Calif. Press, pp. 312–374.
Heiser, W.J., 1987. Joint ordination of species and sites: the unfolding technique. In: Legendre P. and L. Legendre (eds.), Developments in numerical ecology. Springer-Verlag, Berlin, pp. 189–224.
Higler, L.W.G. and F. Repko, 1981. The effects of pollution in the drainage area of a Dutch lowland stream on fish and macro-invertebrates. Verh. Int. Verein. Limnol. 21:1077–1082.
Higler, L.W.G. and P.F.M. Verdonschot, in prep. The relation between macro-invertebrates, hydraulics and soil fertilization in two man-made tributaries of a Dutch lowland stream.
Hill M.O., 1973. Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61:237–249.
Hill, M.O., 1974. Correspondence analysis: a neglected multivariate method. Appl. Statist. 23:340–354.
Hill, M.O., 1979. DECORANA - A FORTRAN program for detrended correspondence analysis and reciprocal averaging. Ecology and Systematics, Cornell University, Ithaca, New York, 52 pp.

Hill, M. O. and H. G. Gauch, 1980. Detrended correspondence analysis, an improved ordination technique. Vegetatio 42:47–58.
Höskuldsson, A., 1988. PLS regression methods. J. Chemometrics 2:211–228.
Hutchinson, G. E., 1968. When are species necessary? In: Lewontin E. (ed.), Population biology and evolution, Syracuse Univ. Press, Syracuse, N.Y., pp. 177–186.
Iwatsubo, S., 1984. The analytical solutions of eigenvalue problem in the case of applying optimal scoring method to some types of data. In: Diday E. (ed.), Data Analysis and Informations III, North Holland, Amsterdam, pp. 31–40.
James, F. C. and C. E. McCullach, 1990. Multivariate analysis in ecology and systematics: panacea or Pandora's box. Ann. Rev. Ecol. Syst. 21:129–166.
Jones, V. J., S. Juggins and J. C. Ellis-Evans, 1993. The relationship between water chemistry and surface sediment diatom assemblages in maritime Antarctic lakes. Ant. Sci. 5:339–348.
Jongman, R. H. G., C. J. F. ter Braak and O. F. R. van Tongeren, 1995. Data analysis in community and landscape ecology, Cambridge Univesity Press, Cambridge, 299 pp.
Kautsky, H. and E. van der Maarel, 1990. Multivariate approaches to the variation in phytobenthic communities and environmental vectors in the Baltic Sea. Mar. Ecol. Progr. Ser. 60:169–184.
Kingston, J. C., H. J. B. Birks, A. J. Uutala, B. F. Cumming and J. P. Smol, 1992. Assessing trends in fishery resources and lake water aluminium from paleolimnological analyses of siliceous algae. Can. J. Fish. Aquat. Sci. 49:116–127.
Krzanowski, W J., 1988. Principles of Multivariate Analysis, Clarendon Press, Oxford.
Lebreton, J.-D., D. Chessel, R. Prodon and N. Yoccoz, 1988. L'analyse des relations espèces-milieu par l'analyses canonique des correspondances. I. Variables de milieu quantitatives. Acta Oecol. Gen. 9:53–67.
Lebreton, J.-D., D. Chessel, M. Richardot-Coulet and N. Yoccoz, 1988. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. II. Variables de milieu qualitatives. Acta Oecol. Gen. 9:137–151.
Lebreton, J.-D., R. Sabatier, G. Banco and A. M. Bacou, 1991. Principal component and correspondence analysis with respect to instrumental variables: an overview of their role in studies of structure-activity and species-environment relationships. In: Devillers J. and W. Karcher (eds.), Applied Multivariate Analysis in SAR and Environmental Studies, Kluwer, Dordrecht, pp. 85–114.
Line, J. M., C. J. F. ter Braak and H. J. B. Birks, 1994. WACALIB version 3.3 – a computer program to reconstruct environmental variables from fossil assemblages by weighted averaging and to derive sample-specific errors of prediction. J. Palaeolimnol. 10:147–152.
Lohmuller, J.-B., 1988. The PLS program system: latent variables path analysis with partial least squares estimation. Mult. Beh. R. 23:125–127.
Malmqvist, B. and M. Maki, 1994. Benthic macroinvertebrate assemblages in North Swedish streams – environmentla relationships. Ecography 17:9–16.
Manly, B. F. J., 1991. Randomization and Monte Carlo methods in biology, Chapman and Hall, London, 281 pp.
Martens, H. and T. Naes, 1989. Multivariate calibration, Wiley, Chichester, 419 pp.
McLachlan, G. J., 1992. Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York.
Miller, A. J., 1990. Subset Selection in Regression, Champan and Hall, London, 229 pp.
Odum, E. P., 1971. Fundamentals of Ecology 3rd Edition, W. B. Saunders Company, Philadelphia.
Økland, R. H. and O. Eilertsen, 1994. Canonical correspondence analysis with variation partitioning: some comments and an applications. J Veg. Sci. 5:117–126.
Oksanen, J., 1987. Problems of joint display of species and site scores in correspondence analysis. Vegetatio 72:51–57.
Palmer, M. W., 1993. Putting things in even better order: the advantages of canonical correspondence analysis. Ecology 74:2215–2230.
Pantle, R. and H. Buck, 1955. Die biologische Überwachung der Gewässer und die Darstellung der Ergebnisse. Gas- und Wasserfach 96:604.
Rao, C. R., 1952. Advanced Statistical Methods in Biometric Research, Wiley, New York.
Rao, C. R., 1964. The use and interpretation of principal component analysis in applied research. Sankhya A 26:329–358.

Reilly, S.B. and P.C. Fiedler, 1994. Interannual variability of dolphin habitats in the eastern tropical Pacific. 1. Research vessel surveys. Fish. Bull. 92:434–450.
Ruse, L.P., 1994. Chironomid microdistribution in gravel of an English chalk river. Freshwater Biol. 32:533–551.
Sabatier, R., J.-D. Lebreton and D. Chessel, 1989. Multivariate analysis of composition data accompanied by qualitative variables describing a structure. In: Coppi R. and S. Bolasco (eds.), Multiway data tables, North-Holland, Amsterdam, pp. 341–352.
Saris, W.E. and L.H. Stronkhorst, 1984. Causal modelling in nonexperimental research. An introduction to the LISREL approach, Sociometric Research Foundation, Amsterdam.
Shelford, V.E., 1911. Ecological succession: stream fishes and the method of physiographic analysis. Biol. Bull. (Woods Hole) 21:9–34.
Sládecek, V.E., 1986. Diatoms as indicators of organic pollution. Acta hydrochim. hydrobiol. 14:555–566.
Smilauer, P., 1992. CanoDraw User's Guide v. 3.0, Microcomputer Power, Ithaca, NY USA, 118 pp.
Smilauer, P., 1994. Exploratory analysis of palaeoecological data using the program CanoDraw. J. Paleolimnol. 12:163–169.
Snoeijs, P.J.M., 1989. Effects of increasing water temperatures and flow rates on epilithic fauna in a cooling-water discharge basin. J. Appl. Ecol. 26:935–956.
Snoeijs, P.J.M. and I.C. Prentice, 1989. Effects of cooling water discharge on the structure and dynamics of epilithic algal communities in the northern Baltic. Hydrobiologia 184:99–123.
Soetaert, K., M. Vincx, J. Wittoeck, M. Tulkens and D. Vangansbeke, 1994. Spatial patterns of Westerschelde meiobenthos. Estuarine Coastal and Shelf Science 39:367–388.
Stevenson, A.C., H.J.B. Birks, R.J. Flower and R.W. Battarbee, 1989. Diatom-based pH reconstruction of lake acidification using canonical correspondence analysis. Ambio 18:228–233.
Stewart-Oaten, A., W.M. Murdoch and K.P. Parker, 1986. Environmental impact assessment: "Pseudoreplication" in time? Ecology 67:929–940.
Sundbäck, K. and P. Snoeijs, 1991. Effects of nutrient enrichment on microalgal community composition in a coastal shallow-water sediment system: an experimental study. Bot. Mar. 34:341–358.
Takane, Y., H. Yanai and S. Mayekawa, 1991. Relationships among several methods of linearly constrained correspondence analysis. Psychometrika 56:667–684.
ter Braak, C.J.F., 1985. Correspondence analysis of incidence and abundance data: properties in terms of a unimodal reponse model. Biometrics 41:859–873.
ter Braak, C.J.F., 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology 67:1167–1179.
ter Braak, C.J.F., 1987a. The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio 69:69–77.
ter Braak, C.J.F., 1987b. Ordination. In: Jongman R.H.G., C.J.F. ter Braak and O.F.R. van Tongeren (eds.), Data analysis in communityy and landscape ecology, Pudoc, Wageningen (reprinted by Cambridge University Press, Cambridge, 1995), pp. 91–173.
ter Braak, C.J.F., 1988a. Partial canonical correspondence analysis. In: Bock H.H. (ed.), Classification and related methods of data analysis, North-Holland, Amsterdam, pp. 551–558.
ter Braak, C.J.F., 1988b. CANOCO – a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1). Report LWA-88-02., Agricultural Mathematics Group, Wageningen, 95 pp.
ter Braak, C.J.F., 1990a. Interpreting canonical correlation analysis through biplots of structural correlations and weights. Psychometrika 55:519–531.
ter Braak, C.J.F., 1990b. Update notes: CANOCO version 3.1, Agricultural Mathematics Group, Wageningen, 35 pp.
ter Braak, C.J.F., 1992. Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Jöckel K.-H., G. Rothe and W. Sendler (eds.), Bootstrapping and related techniques, Springer Verlag, Berlin, pp. 79–85.
ter Braak, C.J.F., 1994. Canonical community ordination. Part I: Basic theory and linear methods. Ecoscience 1:127–140.

ter Braak, C. J. F., 1995a. Non-linear methods for multivariate statistical calibration and their use in palaeoecology: a comparison of inverse (k-Nearest Neighbours, PLS and WA-PLS) and classical approaches. Chemometrics Intell. Lab. Syst. 28:165–180.
ter Braak, C. J. F., 1995b. Canonical community ordination. Part II: The correspondence analysis family. in prep.
ter Braak, C. J. F. and S. Juggins, 1993. Weighted averaging partial least squares regression (WA-PLS): an improved method for reconstructing environmental variables from species assemblages. Hydrobiologia 269:485–502.
ter Braak, C. J. F., S. Juggins, H. J. B. Birks and H. Van der Voet, 1993. Weighted averaging partial least squares regression (WA-PLS): definition and comparison with other methods for species-environment calibration. In: Patil G. P. and C. R. Rao (eds.), Multivariate Environmental Statistics, North-Holland, Amsterdam, pp. 525–560.
ter Braak, C. J. F. and C. W. N. Looman, 1994. Biplots in reduced-rank regression. Biom. J. 36:983–1003.
ter Braak, C. J. F. and I. C. Prentice, 1988. A theory of gradient analysis. Adv. ecol. res. 18:271–317.
ter Braak, C. J. F. and H. van Dam, 1989. Inferring pH from diatoms: a comparison of old and new calibration methods. Hydrobiologia 178:209–223.
Underwood, A. J., 1992. Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. J. Exp. Mar. Biol. Ecol. 161:145–178.
van Nes, E. H. and H. Smit, 1993. Multivariate analysis of macrozoobenthos in Lake Volkerak-Zoommeer (the Netherlands): changes in an estuary before and after closure. Achiv. Hydrobiol. 127:185–203.
Verdonschot, P. F. M., 1989. The role of oligochaetes in the management of waters. Hydrobiologia 180:213–227.
Verdonschot, P. F. M. and C. J. F. ter Braak, 1994. An experimental manipulation of oligochaete communities in mesocosms treated with chlorpyrifos or nutrient additions: multivariate analysis with Monte Carlo permutation tests. Hydrobiologia 278:251–266.
von Tümpling, W., 1966. Über die statistische Sicherheit soziologischer Methoden in der biologischen Gewässeranalyse. Limnologica (Berlin) 4:235–244.
Walker, I. R., J. P. Smol, D. R. Engstrom and H. J. B. Birks, 1991. An assessment of Chironomidae as quantitative indicators of past climatic change. Can. J. Fish. Aquat. Sci. 48:975–987.
Washington, H. G., 1984. Diversity, biotic and similarity indices: a review with special relevance to aquatic ecosystems. Water Res. 18:653–694.
Whittaker, R. H., S. A. Levin and R. B. Root, 1973. Niche, habitat and ecotope. Amer. Nat. 107:321–338.
van Wijngaarden, R. P. A., P. J. van den Brink, J. H. Oude Voshaar and P. Leeuwangh, 1995. Ordination techniques for analysing response of biological communities to toxic stress in experimental ecosystems. Ecotoxicol. 4:61–77.
Wold, H., 1982. Soft modeling: the basic design and some extensions. In: Joreskog K. G. and H. Wold (eds.), Systems under indirect observations II, North-Holland, Amsterdam, pp. 1–54.
Zelinka, M. and P. Marvan, 1961. Zur Präzisierung der biologischen Klassifikation der Reinheit fliessender Gewässer. Arch. Hydrobiol. 57:389–407.

Received 20 March 1995;
revised manuscript accepted 30 May 1995.