ÉCOSCIENCE
1 (2) : 127-140 (1994)

Canonical community ordination. Part I: Basic theory and linear methods[1]

Cajo J. F. ter BRAAK, DLO-Agricultural Mathematics Group and DLO-Institute of Forestry and Nature Research,
Box 100, NL-6700 AC Wageningen, The Netherlands, e-mail: c.j.f.ter.braak@glw.agro.nl

*Abstrac*t: Canonical community ordination comprises a collection of methods that relate species assemblages to their environment, in both observational studies and designed experiments. Canonical ordination differs from ordination *sensu stricto* in that species and environment data are analyzed simultaneously. Part I reviews the theory in a non-mathematical way with emphasis on new insights for the interpretation of ordination diagrams. The interpretation depends on the ordination method used to create the diagram. After the basic theory, Part I is focused on the ordination diagrams in linear methods of canonical community ordination, in particular principal component analysis, redundancy analysis and canonical correlation analysis. Special attention is devoted to the display of qualitative environmental variables.
Keywords: multivariate analysis, principal component analysis, redundancy analysis, canonical correlation analysis, biplot, ordination diagram, species-environment relations.

Résumé: L'ordination canonique des communautés renferme un ensemble de méthodes reliant les assemblages d'espèces à leur milieu et ce, autant avec des données de travaux empiriques qu'expérimentaux. L'ordination canonique diffère de l'ordination au sens strict en ce que les données d'assemblage d'espèces et les variables environnementales sont analysées simultanément. Ce premier article passe en revue de façon non-mathématique la théorie de l'ordination canonique des communautés. Nous insistons principalement sur de nouveaux aspects utiles dans l'interprétation des diagrammes d'ordination, laquelle dépend de la méthode d'ordination employée lors de l'analyse. La démonstration se poursuit en considérant les diagrammes d'ordination basés sur des méthodes linéaires d'ordination canonique des communautés, plus particulièrement l'analyse en composantes principales, l'analyse de la redondance et l'analyse de corrélation canonique. Une attention spéciale est accordée aux variables environnementales qualitatives.
Mots-clés: analyse multivariée, analyse en composantes principales, analyse de redondance, analyse de corrélation canonique, plan factoriel, relations espèces-milieu, variables instrumentales.

Introduction

Ordination is primarily a research tool for the interpretation of field data on plant and animal assemblages and their environment. The graphical result of an ordination, the ordination diagram, is an important aid in this interpretation. But, an ordination diagram is also a communication tool. In a research paper, a single ordination diagram can often replace a large table or even a number of tables or graphs. Compared to tables, ordination diagrams have the disadvantage that they require more skill of the reader. To be an effective tool for communication, every reader of a paper that uses ordination methods should know *how* to read ordination diagrams and, preferably, know the basic theory. Every user of ordination methodology needs to know, in addition, how to produce the diagram and how to describe it properly. This paper supplies the required knowledge.

In this paper, I summarize the basic theory of ordination in a non-mathematical way and highlight new developments since the publication of ter Braak & Prentice (1988) and ter Braak (1987a) – the ordination chapter in Jongman, ter Braak & van Tongeren (1987). These papers describe the state-of-art at the release of version 2.1 of the computer program CANOCO (ter Braak, 1988). An annotated bibliography by Birks & Austin (1992) for the period 1986-1991 lists 165 entries covering both the theory and applications of canonical correspondence analysis and related constrained ordination methods. In 1990, CANOCO version 3.1 was released (ter Braak, 1990a). The new developments fall under the main headings: ordination diagrams and their interpretation, ordination diagnostics, analysis of variance tables, and tests of statistical significance by Monte Carlo methods. Part I of this series of papers covers the basic theory and the general rules for the interpretation of ordination diagrams. The rules depend on the model behind the diagram. The link between data table, model and ordination diagram is elucidated. After this more general overview, Part I is focused on linear ordination methods, in particular principal component analysis, redundancy analysis and canonical correlation analysis. New theory is given for the display of qualitative variables. Part II will review the correspondence analysis family and Part III will focus on tests of significance and the analysis of designed experiments.

A typology of ordination methods

Community ecologists aim at understanding the occurrence and abundance of taxa (usually species) in space and time. Except at small spatial and temporal scales, the environment is thought to be the most important driving force. Sites (*e.g.* quadrats, volumes, stands, traps) with different environmental conditions are sampled. At each site, the presence and abundance of species of interest are

[1]Rec. 1994-03-29; acc. 1994-06-09.

determined, and the environment is characterized by a set of quantitative and/or qualitative variables. The primary data on the species assemblages and the environment are collected in species × sites and environment × sites data tables (Figure 1). From these primary tables, secondary tables can be derived, such as the site × site table of dissimilarities among sites (the Euclidean distances in Figure 1a), the species × species table of (dis)similarities among species (*e.g.* correlations among species), and environment × species tables (*e.g.* correlations and means for quantitative and qualitative environmental variables, respectively). Ordination is a tool to help ecologists understand the species-environment relationships from such data tables. Classification methods can also be used. Classification assumes from the outset that the species assemblages fall into discontinuous groups, whereas ordination starts from the idea that such assemblages vary gradually.

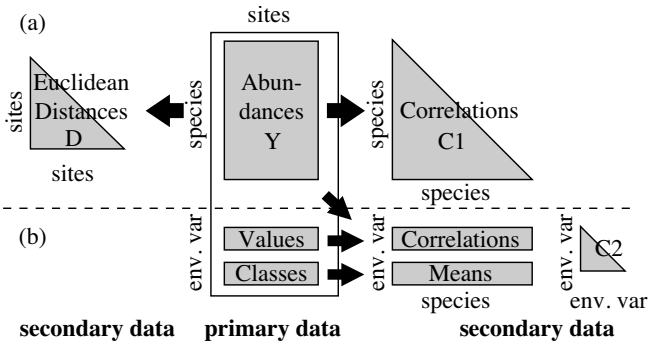

FIGURE 1. Primary and secondary data tables in an ecological study on species-environment relations. Indirect methods of ordination use the tables under (a). Direct methods also use the tables under (b). The primary data are the table of abundance values and the tables of values and class labels of quantitative and qualitative environmental variables (env. var), respectively. The secondary tables are named after the (dis)similarity coefficients they contain. The appropriate coefficients must be chosen by the ecologist. The coefficients shown in the figure are optimal when species-environment relations are linear. In (a) the secondary tables are derived from the abundance data only as the arrows indicate; in (b), the secondary table "Correlations" (containing the correlation coefficients between each species' abundance and each quantitative environmental variable) is derived from the abundance data and the values of the quantitative environmental variables; the table "Means" (containing the mean abundance of each species per environmental class) is derived from the abundance data and the class labels of the qualitative environmental variables and the table "C2" (correlations among environmental variables) is derived from the environmental data only.

Ordination methods can be divided in two main groups: direct and indirect methods. Direct methods use species and environment data in a single, integrated analysis. Indirect methods use the species data only. Both groups of methods can be further subdivided on the basis of the underlying model they use for the species responses along environmental gradients (linear or nonlinear, with unimodal response model being a case of particular ecological interest; see Jongman, ter Braak & van Tongeren, 1987: 31) and the form in which they use the species data (the species × sites table of abundances or a derived, symmetric table of (dis)similarities among sites or among species, *i.e.* in Figure 1a, **Y**, **D** or **C1**). If an ordination is derived from dissimilarities, it is often termed a multidimensional scaling (*e.g.* Clarke, 1993). Notice that there is a link between response models for species-environment relations and coefficients in the secondary tables. This link is two-way: an assumed response model determines, at least partly, the type of coefficient, and, the other way round, a chosen coefficient determines, at least partly, the type of response model. This link is illustrated most easily for the species × environment table (Figure 1b). If a linear response model is assumed, the relations between species and environmental variables are linear. Linear relations are best summarized by correlation coefficients, so that the optimal coefficient for use in the species × environment table is the correlation coefficient, at least for quantitative environmental variables (Figure 1b). In contrast, if a unimodal response model is assumed, the relationships are unimodal. Unimodal relations are usefully summarized by their modes or, more conveniently, weighted averages (ter Braak & Looman, 1986), so that a sensible coefficient for the species × environment table is the weighted average (Part II). The other way round, if a correlation coefficient is chosen, the implied response model is linear or approximately linear, and if the chosen coefficient is the weighted average, then the implied response model is unimodal (*i.e.* if the true model is bimodal, the ordination will fail, and if the true model is linear, the ordination will be inefficient). Assumptions about the response model and the choice of coefficients to use in secondary tables are thus interrelated. The divisions in the typology are explained in more detail below.

Direct and indirect methods of ordination

For illustration of indirect methods, Figure 2 (without the arrows) shows a typical ordination diagram that was the result of an ordination of epilithic algal communities at eleven sites in a shallow coastal ecosystem with cooling-water discharge from a nuclear power plant (Snoeijs & Prentice, 1989: Figure 10). An ordination diagram is a coordinate system formed by so-called ordination axes. Sites are represented by points, the coordinates of which are the site scores on two ordination axes. The scores are obtained by applying ordination analysis to the species data, in the example algal abundance data. The ordination analysis calculates the site scores in such a way that close sites in the diagram are similar in species composition and distant sites are dissimilar. In practice, the ordination analysis is carried out by a computer program, such as CANOCO, and the diagram is prepared by a graphical software such as CanoDraw (Smilauer, 1992). In indirect methods of ordination, data on the environment at the sites are not used to determine the locations of the sites in the diagram. But, to help the environmental interpretation, the sites in Figure 2 are labelled by the flow and heating of the water. The diagram shows a systematic difference between heated and unheated sites within a flow level and thus, indirectly, a difference between heated and unheated algal communities. Direct methods, in contrast, use both the species and environment data to arrange the sites along ordination axes. In a direct method, the axes (*i.e.* the site scores) are constrained: they must be a function of the environment. The environmental basis of the ordination is thus ensured.

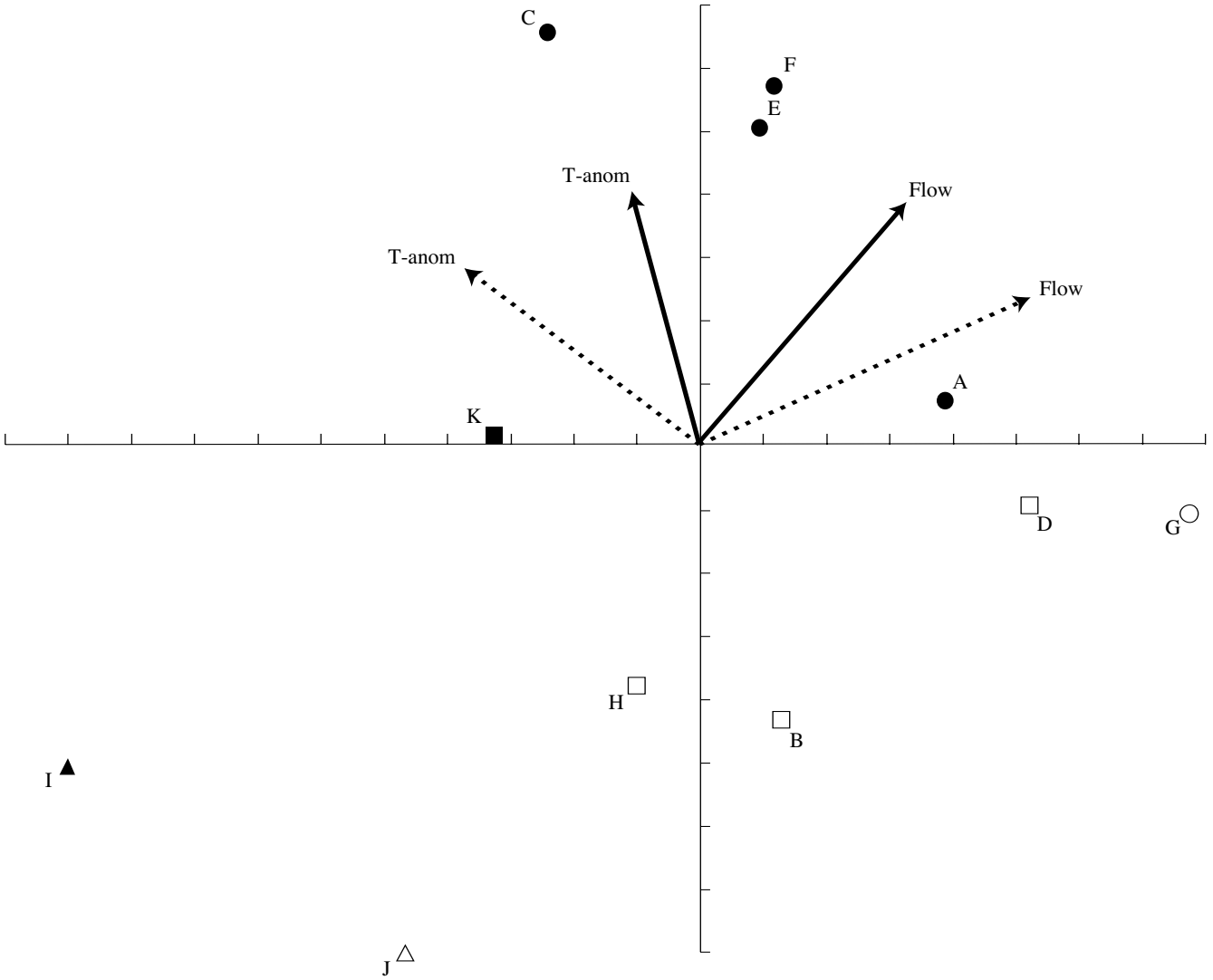

FIGURE 2. Indirect ordination of eleven sites in a shallow coastal ecosystem with cooling-water discharge showing the dissimilarity (chi-square distance) of their epilithic algal communities and interpreted in terms of flow and heating of the water (Snoeijs & Prentice, 1989). The water at sites is flowing (circles), quiescent (squares), or stagnant (triangles), and heated (solid) or unheated (open). The arrows for the flow factor (Flow) and the mean temperature anomaly (T-anom) are added afterwards by simple regressions (solid) and multiple regression (dashed). The arrows illustrate the difference between the marginal effects of variables shown in a standard ordination diagram of canonical ordination (Table Ib) and the conditional effects of variables shown in a regression biplot. For further explanation see text.

We limit our discussion to the case of most practical importance where each axis is constrained to be a linear combination of environmental variables. The constraints on the axes then can be formulated as a linear regression model without an error term, with the axis taking the role of response variable (dependent variable) and with the environmental variables as explanatory variables (independent variables). In this sense, direct ordination integrates indirect ordination and regression. The maximum number of constrained axes is equal to the number of environmental variables (or less than this number if there are linear dependencies among the variables).

Direct ordination can also be characterized as a multivariate form of regression analysis, in which the species data are modeled as a function of the environment data. The differences with regression are that all species are analyzed together and that the ordination axes act as intermediaries: the species data are modelled as a function of the ordination axes that are, in turn, a function of the environment. The effects of the environment on the species are channelled through the ordination axes.

Direct ordination lies actually in between multivariate regression and classical indirect ordination. At the one extreme, if the maximum number of constrained axes is used, direct ordination is equivalent with multivariate regression with the axes being just a linear transformation of the environmental variables. Direct ordination *sensu stricto* involves dimension reduction by using a few ordination axes only. On the other extreme, if the number of environmental variables is greater than the number of sites, direct ordination is equivalent with indirect ordination. The reason for this is as follows. If more and more environmental variables are used, the constraints on the axes become weaker and weaker, until the point that the constraints are effectively absent. This point is reached if the number of environmental variables is one less than the number of sites.

It is possible to extract more constrained axes than environmental variables. These further axes are unconstrained and extract the main patterns of variation that remain after the effects of the environmental variables have been removed. In CANOCO, these further unconstrained axes can be obtained directly by treating the environmental variables as so-called covariables. In an analysis with covariables (*i.e.* in a partial ordination), the effects of the covariables are removed. Common examples are the removal of observer effects, seasonal effects or spatial effects (Borcard, Legendre & Drapeau, 1992).

Direct ordination is also termed constrained ordination (ter Braak & Prentice, 1988) or canonical ordination (ter Braak, 1987a). These terms are often used interchangeably. I reserve the term 'canonical' for methods with external linear constraints, *i.e.* where the ordination axes are constrained to be a *linear* function of a set of predictor variables. A review of multidimensional scaling with external constraints is given by ter Braak (1992).

Direct ordination as defined above extends Whittaker's (1967) definition. In Whittaker (1967), each axis of a direct ordination represented a single environmental variable or a given environmental complex-gradient. In the case of a complex-gradient, the weight given to each environmental variable was determined by the researcher or the weight was defined implicitly (*e.g.* in case of geographic gradients and ecoclines), whereas in our definition of direct ordination, the weights are optimally determined by the analysis. Whittaker's direct ordination did not involve dimension reduction and was therefore formally equivalent with multivariate regression and calibration (ter Braak & Prentice, 1988). Whittaker's analysis consisted of two parts, a regression part and a calibration part. If measurements of the environmental variable were available, the species distributions (species responses) were graphed against that variable. By noting the relative locations of the modes (or centroids) of those distributions an ordination of the species was obtained. This is the regression part. If, for a new series of sites, the environmental variable had not been measured, an ordination of the sites was obtained from the previous ordination of the species. The ordination of sites (obtained by weighted averaging) was meant to reflect the unmeasured values of the environmental variable. This is the calibration part. In CANOCO, a direct ordination (*via* canonical correspondence analysis) with a single environmental variable and some sites without environmental data amounts to Whittaker's direct ordination. It gives a (one-dimensional) ordination of the species along that environmental variable and a joint ordination of the sites with and without environmental measurements. There exist more advanced methods for studies with one or just a few (< 5 let us say) environmental variables. For the regression part see Jongman, ter Braak & van Tongeren (1987), Yee & Mitchell (1991) and Huisman, Olff & Fresco (1993). For the calibration part see Birks *et al.* (1990) and ter Braak *et al.* (1993). Kingston *et al.* (1992) use direct ordination and calibration in a complementary way.

From Figure 2, it can not be inferred how the algal communities changed under the influence of flow and heating. To enable such inference, points for the algae can be added to the diagram. Snoeijs & Prentice (1989) added the algae in a separate set of diagrams to avoid crowding and concluded that a temperature increase favoured blue-green algae at the expense of red and brown algae. The interpretation of diagrams with one, two, or more sets of points is the topic of the next section.

Data tables, models, and ordination diagrams

Ordination diagrams visualize the main structure of multivariate data tables in two (or three) dimensions. It depends on the model used to create the diagrams how the data tables are visualized, and thus how the diagrams must be interpreted. In almost all applications in ecology, the interpretation is either *via* distances among points (distance diagrams) or via directions across the diagram (biplots). This is explained in more detail below. Which data tables are displayed depends, of course, on which of the data tables in Figure 1 the ordination is applied to or, if the ordination is applied to the primary data, as is usual in ecology, which of the primary or secondary data tables is the major focus of attention. In the latter case, the ordination method determines which data table is the focus. For example, a canonical ordination emphasizes the species × environment table. For a systematic discussion of the types of diagrams, I take the data table that has the focus as the leading entry. This section concludes with a subsection on how well the data table is displayed.

With the focus on the *inter-sites table* **D** in Figure 1a, a definition of ordination is that it arranges points (representing sites) in a diagram in such a way that distance between points corresponds as well as possible with the dissimilarity between sites. The interpretation is thus via inter-point distances: sites that are similar in species composition are close together and sites that are dissimilar lie far apart. Be aware that points that are close may show considerable dissimilarity in species composition if the 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 measure of dissimilarity used in the ordination of Figure 2 is the chi-square distance. [The chi-square distance is the implicit dissimilarity coefficient if the chosen ordination method belongs to the correspondence analysis family (see part II)].

In the above definition, ordination is equivalent to multidimensional scaling. The user must select an appropriate dissimilarity coefficient from the ones available. See Faith (1983), Gower & Legendre (1986), Faith, Minchin & Belbin (1987), and Shi (1993), among others, for theoretical and practical evaluations of dissimilarity coefficients. The implied model for the species abundance table, if any, remains hidden.

When the focus is on the *species × sites abundance table* **Y** in Figure 1a, a second, more ambitious definition of ordination is that it arranges both sites and species as points in a diagram in such a way that the species assemblages can be derived from it as well as possible. How the species assemblages can be derived depends on the type of diagram. There are two types of diagrams in common usage, distance

diagrams[2] and biplots. Distance diagrams (with points for species and sites) are based on a unimodal model or – as psychometricians call it – an unfolding model (DeSarbo & Rao, 1984; Heiser, 1987). Biplots are based on a linear model or, in case of nonlinear biplots, a generalization thereof (Gower & Harding, 1988). Distance diagrams and biplots are discussed in separate subsections below. One would hope that despite the focus on the abundance table, the diagram would also display the two derived tables of (dis)similarities among sites and among species. This seems possible theoretically in distance diagrams (I cannot think of a reason why not, but this aspect has not received attention in the literature), but it is impossible in biplots. The problem with biplots is that only one of the derived tables in Figure 1a can be optimally displayed in conjunction with **Y**. To determine which one, it depends on the scaling chosen by the user (Table Ia).

With a focus on the *species* × *environment tables* (Figure 1b), ordination arranges both species and environmental variables as points, or perhaps arrows, in a diagram in such a way that the species × environment tables can be derived from it as well as possible. If this was the definition of canonical ordination, it would simply be an ordination applied to another data table; it would not arrange sites in the diagram. Canonical ordination is more than this: it fits an ordination model to the species data with axes that are linearly constrained by the environment variables. Compared to an ordination of the species × environment table, it emphasizes the environmental variables that are influential and deemphasizes or, preferably, neglects the variables that do not influence the species assemblage (ter Braak, 1987b). An extremely useful by-product of canonical ordination is thus an optimally weighted ordination of the species × environment tables.

The complete canonical ordination diagram consists of three sets of points: sites, species, and environmental variables. Species and sites jointly display the fitted abundance data, whereas species and environmental variables display the species-environment relationships. Both displays are optimal in a statistical sense. In addition, sites and environmental variables may display the environment data table to some degree, but this display is by no means optimal. It was not meant to.

In conclusion, canonical ordination arranges sites, species and, environmental variables in a diagram in such a way that the fitted species × sites table and the species × environment tables can be derived from it as well as possible.

The algorithms for canonical ordination in Jongman, ter Braak & van Tongeren (1987) give two sets of site scores. One is derived from the species and the other from the environment variables. In the diagrams in ter Braak (1986) and Jongman, ter Braak & van Tongeren (1987), the first set of scores were used to display sites on the grounds that species points and site points better display the abundance data. But, the site scores of the second set are the constrained ones and thus central to the canonical ordination.

[2] The term "distance diagram" is introduced here to replace the less apt term "joint plot" in ter Braak (1987a) and in ter Braak & Prentice (1988).

TABLE I. Tables that can be displayed by two differently scaled biplots in principal components analysis (a) and redundancy analysis (b). The sum of squares of site scores of an axis is equal to its eigenvalue in scaling 1, and equal to 1 in scaling 2. The sum of squares of species scores of an axis is equal to 1 in scaling 1 and equal to its eigenvalue in scaling 2. Tables in bold are fitted by (weighted) least-squares (env. = environmental; vars = variables; cl. = classes)

Biplot scaling	1: focus on sites distance biplot	2: focus on species correlation biplot
(a) principal component analysis		
species × sites	**abundances**	**abundances**
sites × sites	Euclidean distances	-
species × species	-	**correlations**[a]
(b) redundancy analysis		
species × sites[b]	**fitted abundances**	**fitted abundances**
sites[b] × sites	Euclidean distances[c]	-
species × species	-	**correlations**[a,c]
Quantitative env. vars:		
species × env. vars[d]	**correlations**	**correlations**
sites × env. vars[d]	-	values of env. vars
env. vars[d] × env. vars	effects[e]	correlations
Qualitative env. vars:		
species × env. classes[f]	**means**	**means**
sites × env.classes[f]	g	g
env. classes[f] × env. cl.	Euclidean distances	-
env. vars × env. classes	-	means

[a] Automatic if abundance is standardized by species. If abundance is only centred by species, a post-hoc rescaling of the species scores is needed so as to account for the differences in variance among species.
[b] Site scores are a linear combination of the environment variables instead of being a weighted sum of species abundances as in Table Ia.
[c] In the definition of this coefficient, abundance must be replaced by fitted abundance.
[d] Environmental scores are intraset correlations in scaling 2 and $\lambda_s^{1/2}$ times those in scaling 1 with λ_s the eigenvalue of axis *s*. In CANOCO, the scores are termed biplot scores for environmental variables.
[e] Effect of the environmental variable on the ordination scores, while neglecting the other environmental variables; length of arrow is the effect size, *i.e.* the variance explained by the variable (Appendix B).
[f] Environmental classes are centroids of site points belonging to the class (Appendix A).
[g] Membership *via* centroid principle (see Part II), not *via* the biplot rules.

The first set of scores is supplementary and goes part of the way towards the set of unconstrained scores in an indirect analysis (Palmer, 1993). Although the first set approximates better the species abundance data as observed, the second set approximates better the fitted abundances. In conclusion, the second set of site scores, which are linear combinations of the environmental variables, must be plotted.

Distance diagrams and unimodal models

The first definition of ordination (arranging site points in a diagram in such a way that the distances between site points correspond as well as possible with the dissimilarity between sites) gives a distance diagram with a single set of points: the distance among pairs of points is positively related to the dissimilarity among pairs of sites. (As an aside, an ordination method is called metric if this relation is linear and nonmetric if the relation is just monotonic [*e.g.* Clarke, 1993]). With the second definition (arranging both sites and species as points in a diagram in such a way that

the species assemblages can be derived from it as well as possible), the diagram contains two sets of points, sites and species. It is a distance diagram if the distance between a site point and a species point is inversely related to the abundance value. This definition of a distance diagram, implies a symmetry between species and sites that does not often exist in the ecological context: species may differ in maximal abundance, thus only abundance values within a row (species) can be compared. The technical term for this asymmetry is "row conditionality" (Schiffman, Reynolds & Young, 1981). In the species-conditional distance diagram, the inferred abundance of a species is maximal if the site point coincides with the species point and decreases the farther away the site point is. The species point is thus the mode or optimum of a unimodal surface or – said less technically – the peak of an imaginary 'mountain' above the ordination plane. Notice that each species has its own imaginary mountain. The diagram thus reflects the ecological idea of the Hutchinsonian niche of a species and of niche-space partitioning (Whittaker, Levin & Root, 1973). Distance diagrams can be produced on the basis of ordination by unfolding (DeSarbo & Rao, 1984; Heiser, 1987) or ideal point discriminant analysis (Takane, Bozdogan & Shibayama, 1987). At present, these computationally demanding methods are beyond routine application to ecological data, because of the size of the data sets and the numerical problems with the methods. If the niches of species are reasonably well separated, correspondence analysis methods may be used instead (Jongman, ter Braak & van Tongeren, 1987; ter Braak & Prentice, 1988). The controversy associated with this claim (Oksanen, 1987) is discussed in Part II.

In conclusion, the distance diagram for a single set of points assumes a linear or monotonic model between distance and dissimilarity, whereas for two sets of points, it is closely related to the unimodal model for the abundance data.

Biplots and linear models

A biplot is based on a linear model and can consist of one or two or even more sets of points, even though the prefix 'bi' was chosen for the case with two sets of points (Gabriel, 1982). With more than two sets, each pair of sets forms a biplot. Let me start with explaining the biplot for two sets of points (let us say, species and sites). A biplot is a scatter diagram and an origin, usually at the centre of the plot. Formally, a biplot is a diagram in which the value of a species at a site is given by the inner product between the species point and the site point (Gabriel, 1982). Fortunately, the practical usage of a biplot is far easier than this definition. A biplot can simply be read as the usual scatter diagram of two variables (the XY-point-chart): for the variable set out horizontally, draw vertical lines from a point and read off the scale value at the intersection with the axis. Usually, the scale is not marked quantitatively so that only the ranking of the values at the points is known. The biplot differs from this in that each point, *e.g.* each species point, can be connected with the origin to form an imaginary axis. The values of that particular species at the sites are inferred by mentally rotating the plot so that this axis is horizontal and then by proceeding as if it were a normal scatter plot. The rotation can be avoided by drawing perpendicular lines instead of the vertical lines. By drawing the perpendicular lines from the sites (*i.e.* by projection), we thus obtain a ranking of the values of this species at the sites. The position of the origin along this axis indicates the value zero. If each species was centred in advance, a zero value thus corresponds to the mean value in the original data. Another difference with the scatter diagram is that values are displayed with some error. The differences between a biplot and the usual scatter plot can thus be summarized by noting that a scatter plot displays the values of two variables in an exact way by perpendicular axes whereas the biplot displays the values of many variables in an approximate way by oblique axes.

How to interpret a biplot is illustrated in Figure 3 for the species indicated as Pla lan. The line through its point is the imaginary axis for this species. By projecting the sites on to this axis, as illustrated for sites 11 and 12, we see that the inferred abundance is highest in site 2, closely followed by sites 5, 6, 7, 11 and 18, somewhat lower in sites 10 and 17, about average in sites 1 and 9, and less than average in the remaining sites and least in site 16. Notice that sites 7 and 18 have about the same inferred abundance of Pla lan; the difference in distance to the imaginary axis is immaterial for Pla lan — it tells us about differences in the abundances of other species. (For comparison, the horizontal coordinate of a point in an ordinary scatter diagram gives the value of the variable set out horizontally, irrespective of the value of the vertical coordinate). The imaginary axis defines the direction in the diagram along which the value of the species changes. It is thus directions that must be interpreted in a biplot, not site-to-species distances.

A biplot is often obtained by adding species or environment points to an existing distance diagram of sites. This means that the interpretation of the site configuration is *via* distances whereas the joint interpretation is *via* directions. This mix is called a distance biplot (ter Braak, 1983).

In a biplot with a single set of points, each point can take the role of the first set and the remaining points the role of the second set. Each particular point thus forms an imaginary axis on to which the other points of the set can be projected. For example, if a biplot with points for species is said to display correlations (correlation biplot), then the projection points on the imaginary axis yield a ranking of correlations of the species with the species that forms the imaginary axis. In this ranking, the origin indicates zero correlation. Figure 3 provides an example. By projecting and ranking the species points on to the imaginary axis drawn for Pla lan, we see that Ach mil has the highest inferred positive correlation with Pla lan, closely followed by Lol per, Poa pra, Leo aut, Bro hor and Bel per. Poa tri and Sal rep have near zero correlation with Pla lan, whereas Agr sto is the most negatively correlated with Pla lan.

Biplots for compositional data, *i.e.* data scaled so that each site total sums to 100%, require special treatment, because the directions that deserve interpretation do not run through the origin, but through pairs of points (Aitchison, 1986; 1990). Here follows why. If the site totals are constant or an artifact of the sampling procedure, then only ratios of abundances, *e.g.* y_{ki}/y_{li} for species k and l in site i,

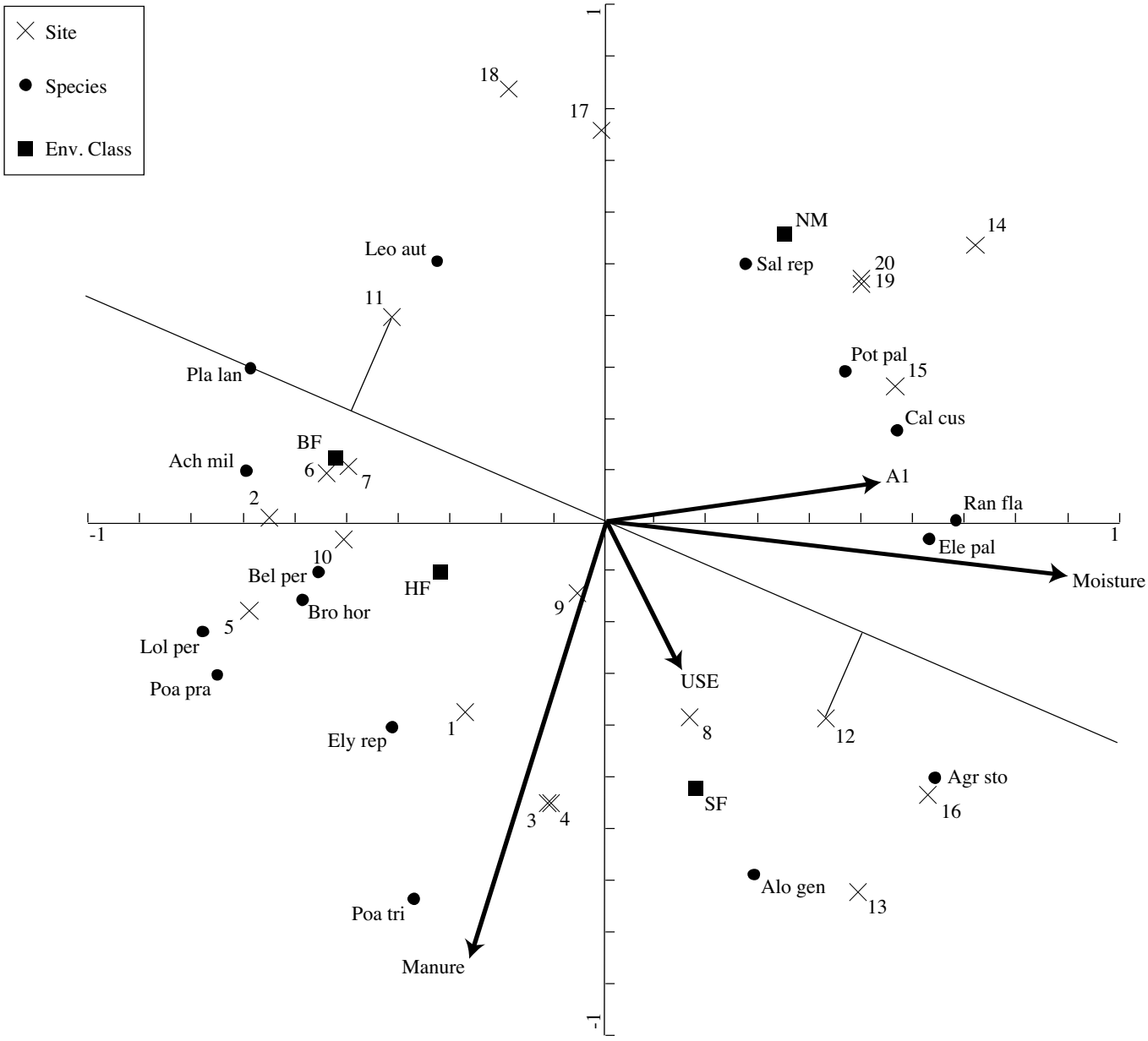

FIGURE 3. Correlation biplot based on a redundancy analysis of the Dune Meadow data displaying 43% of the variance in the abundances and 71% of the variance in the fitted abundances. Quantitative environment variables are indicated by arrows. The qualitative variable Management Type is indicated by the square points labelled SF, BF, HF, and NM. The displayed species are selected on the basis that more than 30% of their variance is accounted for by the diagram. Eigenvalues of the first three axes are 0.26, 0.17 and 0.07; the sum of all canonical eigenvalues is 0.61. The scale marks along the axes apply to the species and quantitative environmental variables; the site scores and class scores were multiplied by 0.46 to fit in the coordinate system. The abbreviations are given in Jongman, ter Braak & van Tongeren (1987). The rule for interpreting a biplot (projection on an imaginary axis) is illustrated for the species Pla lan and sites 11 and 12. For further explanation see text.

can be compared meaningfully among sites. Ratios have the disadvantage of being asymmetric in the species compared: if the abundance of species *k* goes to 0, the ratio goes to 0, whereas if that of species *l* goes to 0, the ratio goes to infinity. It is therefore much more attractive to use logratios. The problem this introduces with zero abundances can be circumvented by replacing zeroes by a small amount (Aitchison, 1990) or by replacing observed abundances by expected abundances as is done in generalized linear models (Dobson, 1990). The latter is worked out in ter Braak (1988: 66) and ter Braak *et al*. (1993: 534-535). Now, the logratio is a difference of the logs, namely

$$\log(y_{ki}/y_{li}) = \log(y_{ki}) - \log(y_{li}).$$

Differences can be read off from a biplot that displays $\log(y_{ki})$ (or a double-centred version thereof) as follows. First, connect the points for species *k* and *l* and extend it in both directions to form an imaginary axis for this pair of species. Second, project the sites on to this axis. The ordering of the projection points now gives a ranking of the logratios of these two species among sites. Examples and further details are given by Aitchison (1990) and Hermy & Lewi (1991).

QUALITY OF DISPLAY

An ordination diagram displays in two or three dimensions a multidimensional data table. Therefore, it generally does not display the data tables exactly. It "approximates" the data. How well the data are displayed is commonly expressed by the percentage variance accounted for. In methods that can be solved by eigenanalysis, this percentage can often be derived from the sum of the eigenvalues of the displayed axes and the total sum of eigenvalues. In eigenanalysis, the approximation criterion is always of the weighted least-squares form.

The quality of the display is best described in the legend of the diagram. In interpreting percentages variance accounted for, it must be kept in mind that the goal is not 100%, because part of the variance is due to noise in the data or to estimation error in correlation coefficients and fitted values. Even an ordination diagram that explains only a low percentage may be quite informative (ter Braak, 1986). The legend of Figure 3 gives the first three eigenvalues so as to give an idea of the stability of the ordination. The ordination is more stable, the smaller the third eigenvalue is compared to the second. In general, if s dimensions are used for the ordination, the ordination is unstable if the $(s + 1)$th eigenvalue is close to the sth eigenvalue.

The eigenvalues are a much better measure of the quality of the ordination and of the strength of the species-environment relationship than the so-called species-environment correlation (ter Braak, 1986). The reason is that even ordination axes that explain little of the species data may have a high species-environment correlation. The problem is similar to that of the interpretation of canonical correlations in canonical correlation analysis (Gittins, 1985).

Ordination diagrams in linear methods

Principal components analysis is the linear method of indirect ordination. In its standard form, the abundance table is centred by species, but other data transformations are possible. The implied derived tables contain the Euclidean distances among sites and correlations among species (Figure 1a) or covariances among species. Its ordination diagram is a biplot of sites and species that displays the (transformed) abundance data. The biplot can be normalized (scaled) in various ways, two of which are of special interest (Table Ia; Gabriel & Odoroff, 1990). In the first scaling the focus is on sites: inter-site distances in the biplot display the Euclidean distances among sites. In the second, scaling the focus is on species, because the species points form a biplot that displays the correlations among species.

The standard form of principal component analysis uses data centred, but not standardized, by species. Notwithstanding, after the ordination axes are obtained, each species can be standardized to zero mean and unit variance so as to obtain a biplot that is easy to interpret. This optional standardization of species is new compared to CANOCO version 2.1 and to the diagrams in Jongman, ter Braak & van Tongeren (1987) and has two advantages. First, the diagram displays standardized data, so that, with the focus on species (Table I), correlations are displayed instead of covariances. Correlations are easier to interpret because their values always lie between -1 and +1. Their sign and magnitude indicate the type and strength of the linear relation, respectively. In contrast, covariances have no fixed scale of reference; their magnitude increases with the variances of the variables. Second, the distance of a species point from the origin is the multiple correlation of the species with the ordination axes of the diagram, *i.e.* the distance tells how accurate the abundances can be read from the diagram. The standardization counteracts the unwanted effect (Hill, 1973) that dominant species lie far from the origin merely by their large abundance and associated large variance. A negative aspect of the post-hoc standardization is that it hides the problem that one or two dominant species may almost completely determine the ordination analysis, because of their large variance. It is therefore wise to check always the variances of the species on extreme values.

Redundancy analysis is the linear method of direct ordination. It also goes under the name of principal component analysis with respect to instrumental variables, which in our case are environmental variables (Sabatier, Lebreton & Chessel, 1989). It is also called least-squares reduced rank regression so as to emphasize its link with multivariate regression (ter Braak & Prentice, 1988; ter Braak & Looman, 1994). In statistical textbooks on multivariate analysis, redundancy analysis is usually neglected. Instead, canonical correlation analysis (Gittins, 1985) is presented as the standard method to relate two sets of variables. However, the latter method is useless if there are many species compared to sites, as in many ecological studies, because its ordination axes are very unstable in such cases. The ordination diagrams of canonical correlation analysis and redundancy analysis display the same data tables; the difference lies in the precise weighing of the species (ter Braak, 1987a; 1990b; ter Braak & Looman, 1994). Recent, good ecological examples of canonical correlations analysis, with many more sites than species, are Van der Meer (1991) and Varis (1991).

Redundancy analysis and canonical correlation analysis are linear methods. So, if well produced, their ordination diagrams are biplots or the superposition of biplots (a triplot). For illustration, I use the Dune Meadow Data from Jongman, ter Braak & van Tongeren (1987). The data were collected to study the relations between vegetation and management of dune meadows on the island of Terschelling. There are 20 sites, 30 species, and 5 environmental variables, one of which is qualitative (Management Type with four classes). Figure 3 is a correlation biplot (*i.e.* a biplot in scaling 2) based on a redundancy analysis of the species with respect to the environment. The last column of Table I states which data tables of Figure 1 are displayed: species and sites form a biplot of fitted abundance values, species and quantitative environmental variables (displayed by arrows) form a biplot of their pairwise correlations, species and environmental classes (displayed by filled squares) form a biplot of class means. These biplots hold true irrespective of the scaling of diagram. It is special for scaling 2 (correlation biplot) that the species form by themselves a biplot of correlations

among species, the environmental arrows form a biplot of correlations among quantitative environmental variables, and the sites and environmental variables form a biplot of the environment data table. Each of these biplots should be read as a scatter plot with oblique axes as explained in the previous section. The analysis is optimized for the biplots of the tables in bold type face in Table I; the remaining biplots are supplementary and should be interpreted with extra caution. The quality of the display is described in the figure legend and explained more fully below.

The quality of the display can be derived from the eigenvalues as follows. Eigenvalues of the first two axes are 0.26 and 0.17; the sum of all canonical (*i.e.* constrained) eigenvalues is 0.61. The biplot thus represents 43% (= 100 × [0.26 + 0.17]) of the variance in the species abundance data and 71% (= 100 × [0.26 + 0.17]/0.61) of the variance in the fitted species data. These calculations make use of the fact that in CANOCO the total variance in the species data is standardized to 1 (the sum of all constrained and unconstrained eigenvalues is thus 1), and that the sum of the canonical eigenvalues is equal to the variance in the table of fitted species abundances. The variance accounted for of the fitted abundances (71%) is at the same time the (weighted) variance accounted for in the tables of species × environment correlations and means (Appendix A) .

When a paper contains a correlation biplot, there is no need also to publish tables of correlations of either species or environmental variables with the ordination axes: these are already in a correlation biplot by way of the species and environment scores. For example, the coordinates of Moisture are about 0.9 and - 0.1; so its correlation with axis 1 is 0.9 and with axis 2 - 0.1. Further, the length of the arrow is equal to the multiple correlation R of Moisture with the ordination axes (so $R > 0.9$). The length of an arrow thus indicates how strongly the variable is related to the displayed ordination.

The optimal display of qualitative environmental variables has not previously received rigorous treatment. In ter Braak (1987a), I suggested that classes be presented either as quantitative (0/1) variables or as centroids (means) of the site points belonging to the class. In Appendix A, I show that the latter presentation is optimal for the approximation of the table of mean abundances in the classes. So, in the same way as abundances are averaged in the data, so site points are averaged in the diagram. Clearly, the class point plays the same role as a site point; abundances just become mean abundances. In the example of Figure 3, the sites numbered 2, 10 and 11 belong to the management class BF. The point for BF is at the centre of gravity of these three points in Figure 3. By projecting and ranking the class points on the imaginary axis for the species Pla lan, it is inferred that the mean abundance of this species is highest in class BF and decreases in the sequence BF, HF, NM, and SF.

The default redundancy analysis diagram in CANOCO 2.1 depicted covariances between species and environmental variables instead of correlations. In a covariance biplot, the correlation between a species and an environmental variable was then inferred from the angle between their (imaginary) arrows in the diagram (Jongman, ter Braak & van Tongeren, 1987; ter Braak & Prentice, 1988). (In CANOCO 3.1, the optional standardization of species to zero mean and unit variance – after the ordination axes are obtained – turns the covariance biplot into a correlation biplot). In a correlation biplot, the interpretation by angles still yields the correct sign of the correlation, but not the correct magnitude. The correct interpretation of the correlation biplot is by projection. For example in Figure 3, the correlation of species with the quantitative variable Manure decreases from positive in the sequence (using abbreviated species names) Poa tri, Alo gen, Poa pra, Ely rep, and Lol per to more negative in the sequence Cal cus, Leo aut and Sal rep.

The site × site table of Euclidean distances is the only table of Figure 1 that is not well displayed in the correlation biplot. If the focus is on inter-site comparisons, the choice of scaling should be scaling 1 (Table I), which results in a distance biplot. This scaling is particularly attractive with qualitative environmental variables. As explained above, classes play the same role in the diagram as sites, so inter-class comparisons are best made from a distance biplot. Euclidean distances among classes are defined as for sites, with mean abundance replacing abundance. Note from Table I that in a distance biplot, the correlations among species and among quantitative environmental variables are not well displayed. Also, sites and environment do not form a biplot of the environment data, as in scaling 2. The latter is surprising because the focus is on sites! The reason is that in the standard diagram of canonical ordination, the environmental arrows are added in such a way that, together with the species points, they optimally display the species × environment table. Fortunately, the environment arrows also have an interpretation in terms of the site scores. The coordinates of the arrow heads are the regression coefficients of the regression of the axes on the (normalized) environmental variable (Appendix B). The arrow thus points in the direction in which the site scores move if the value of this environmental variable increases, irrespective of the other variables; the arrow length is the distance over which the sites move if the value of the environmental variable increase one unit, *i.e.* 1 standard deviation for a normalized variable. (Note that in the diagram the site points are often shown on a smaller scale than the environmental arrows, see Figure 3). Before the development of canonical ordination in 1986, the way to add a quantitative environmental variable to an indirect ordination of sites used to be different: the variable was regressed on to the axes (Dargie, 1984; Jongman, ter Braak & van Tongeren, 1987: equation 5.12) so that, together with the site points, the arrows formed a biplot of the environment data. Both procedures happen to give the same arrows in the correlation biplot but different arrows in the distance biplot. The arrows in a standard distance biplot (Table Ib) possess an additional property that is central to the logic of canonical ordination: their length is a measure of their importance in explaining the species composition (Appendix B). These lengths better express importance than the lengths in the correlation biplot (Appendix B). Despite the elegance of the distance biplots in principal component analysis and redundancy analysis, it should be noted that inter-site distances in the diagram always underestimate the

Euclidean distances they are supposed to display (Meulman, 1986).

The choice between the two types of scaling of biplots is best made by first writing down what the diagram is meant to demonstrate, translating this in terms of the Tables in Figure 1 and then looking in Table I to see which scaling fits the purpose best. If most environmental variables are quantitative, scaling 2 is a good start, whereas if most environmental variables are qualitative then scaling 1 looks most promising, simply on the basis of number of valid interpretations in Table Ia,b. The choice of scaling is unimportant if the first two eigenvalues are of the same magnitude. A correlation biplot can be transformed into a distance biplot by multiplying the site scores and environment scores of each axis by the square root of the eigenvalue and dividing the species scores by the same value (ter Braak, 1983; Jongman, ter Braak & van Tongeren, 1987). The net result for Figure 3 is that the variation along the second axis decreases for sites and environmental variables by a factor of $(0.17/0.26)^{1/2} = 0.80$ compared to the variation along the first axis and for species increased by a factor of $1/0.80 = 1.24$. The qualitative difference is small; for example, Moisture and Manure make angles of 7° and 107° with the first axis in Figure 3, respectively, and of 6° and 110° in the corresponding distance biplot shown in Jongman, ter Braak & van Tongeren (1987: 146).

For the sake of completeness, here is a list of differences between Figure 3 and the Jongman figure. Apart from the scaling of the biplot, the Jongman figure differs from Figure 3 in that the site scores are a weighted sum of the species, the species points are not adjusted for the species variance, Management Types are shown by arrows instead of by centroid points, the direction of the second axis – which is immaterial anyway – is reversed, the arrow of Use is missing by mistake and all species are displayed whereas ill-represented species were left undisplayed in Figure 3. It is reassuring that despite this list of differences the overall features of the diagram remain intact!

Biplots of compositional data are obtained from principal component analysis or redundancy analysis by transforming the abundances to logarithms and centring the transformed data by species and by sites. Aitchison (1986, 1990) defines and explains summary statistics that are appropriate in the logratio analysis of compositional data. These statistics focus on the species, so that an attractive scaling is given by the covariance biplot (thus without the extra standardization for species; scaling -2 in CANOCO 3.1). In this scaling, distances between species points display the relative variation matrix. Further interpretative aids and examples are given in Aitchison (1990). An example of redundancy analysis of compositional data is given by ter Braak (1988: 64-65).

If the number of environmental variables is small (< 6, say) and their mutual correlations not too high (< 0.6, let us say), it may make sense to depict the environmental variables by their canonical coefficients with the axes. The canonical coefficients are the weights of the environmental variables in the linear combination that defines a constrained ordination axis. There is a close relation with multiple regression coefficients: the canonical coefficients of an axis are the coefficients of a multiple regression of the axis on to the environmental variables. This relation extends from the ordination axes to the species: if depicted by canonical coefficients, the environmental variables with the species together form a biplot of a table of multiple regression coefficients, namely those of the multiple regressions for each of the species on the environmental variables (ter Braak, 1990a, b; ter Braak & Looman, 1994). This biplot is termed a regression biplot to distinguish it from the standard biplots of Table 1b.

Whereas the coordinates of environmental variables follow from simple regressions in standard biplots (Table Ib, Appendix B), they follow from multiple regressions in a regression biplot. The distinction between the two is illustrated in Figure 2. Above, the configuration of sites in Figure 2 was interpreted in terms of flow and heating of the water. Quantitative variables for these are the flow factor and the mean temperature anomaly (Snoeijs & Prentice, 1989: Table 1). The solid arrows are the standard arrows arising from four simple regressions (of the site scores on each of the axes on each of the two variables). The dashed arrows follow from two multiple regressions (of the site scores of each of axes on the two variables). The solid arrow for the temperature anomaly points in the direction of more heating irrespective of the flow, whereas its dashed arrow points in the direction of more heating, given the flow. From the symbols for flow and heating classes, it is clear that the dashed arrow indicates more accurately the average shift in site scores due to heating within each flow class. This illustrates that, if interpreted with care, multiple regression coefficients (*i.e.* conditional effects) are more informative than simple regression coefficients or, for that matter, correlations (*i.e.* marginal effects). But multiple regression coefficients are unstable and, hence, often nonsensical if there are many environmental variables or if these have high mutual correlations. The correlation between the flow factor and the temperature anomaly was 0.5.

If redundancy analysis is used to produce a biplot of regression coefficients, this feature can be stressed by using its alternative name, reduced-rank regression. It is possible to enrich the regression biplot in a simple way to show approximate Student *t*-ratio of regression coefficients (ter Braak, 1990b; ter Braak & Looman, 1994). The display of qualitative variables in a regression biplot is explained in ter Braak & Looman (1994).

Discussion

The more applied an ecological study is, the more the emphasis is on the effects on ecological communities of particular environmental factors, for example pollutants, management regimes, and other human-induced changes in the environment. This emphasis should already have been translated in the research design, albeit observational or manipulative. Correspondingly, the statistical analysis should not 'just' show the major variation in the species assemblage, but focus on the effects on the variables of prime interest. Applied studies thus call for direct methods of ordination, typically with a very limited number of (qualitative or quantitative) environmental variables.

The range of community variation in an applied study tends to be quite small compared with that in the early ordination studies (*e.g.* Whittaker, 1956; Hill & Gauch, 1980). Consequently, the linear method of direct ordination, redundancy analysis, can often efficiently display the interesting effects, and there is no need for methods that also work when the range of community variation is larger (such as canonical correspondence analysis) nor for unconstrained nonmetric multidimensional scaling.

The data from which Figure 2 was produced were samples taken every third week throughout one year at each of the eleven sites. Figure 2 was in fact the result of a constrained ordination, but it could serve as an illustration of indirect ordination, because the configuration of sites was left unconstrained. The ordination of all (18 × 11) samples was constrained by an additive model of site and date (*i.e.* by two qualitative "environmental" variables, represented by two series of dummy variables, indicating to which site and date a particular sample belongs). This ensured that the ordination diagram displayed the systematic differences among sites and among date rather than differences between individual samples that are less easy to interpret. Each point in Figure 2 is the centroid of points representing the 18 samples taken at the same site. Seasonal variation could have dominated the site differences in the diagram. In that case, it would have been better to apply a partial canonical ordination: an ordination constrained by site, while eliminating seasonal variation by specifying date as a series of dummy covariables. This was not needed in the example, perhaps because site differences were partly governed by temperature.

In the statistical literature, the linear regression model has been extended to the generalized linear model and the generalized additive model (for ecological examples, see Yee & Mitchell, 1991). These new methods give new tools for the analysis of presence–absence and count data and for the building of nonlinear regression models. These new models can also be used in the context of (constrained) ordination (*e.g.* Israëls, 1992). Much of the general theory explained in this paper is directly applicable to these generalizations.

May this series of papers serve to enhance the understanding and the proper and creative use of ordination methods in community ecology.

Acknowledgements

I thank M.-J. Fortin and P. Dixon, in their capacity of reviewers, and also H.J.B. Birks, Sandra Clerkx, Petr Smilauer, Margriet Stapel, Han van Dobben, Ad van Hees, Ole R. Vetaas, Hilko van der Voet, and René van Wijngaarden for comments on the manuscript. Figures 2 and 3 were produced by CanoDraw 3.0 (Smilauer, 1992) with post-manipulation of its postscript output so as to illustrate the biplot projection rule.

Literature cited

Aitchison, J., 1986. The Statistical Analysis of Compositional Data. Chapman and Hall, London.

Aitchison, J., 1990. Relative variation diagrams for describing patterns of compositional variability. Mathematical Geology, 22: 487-511.

Birks, H. J. B. & H. A. Austin, 1992. An Annotated Bibliography of Canonical Correspondence Analysis and Related Constrained Ordination Methods 1986-1991. Botanical Institute, Bergen, Norway .

Birks, H. J. B., J. M. Line, S. Juggins, A. C. Stevenson, & C. J. F. ter Braak, 1990. Diatoms and pH reconstruction. Philosophical Transactions of the Royal Society of London, Series B: 327, 263-278.

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

Clarke, K. R., 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology, 18: 117-143.

Dargie, T. C. D., 1984. On the integrated interpretation of indirect site ordinations: a case study using semi-arid vegetation in south-eastern Spain. Vegetatio, 55: 37-55.

DeSarbo, W. S. & V. R. Rao, 1984. GENFOLD2: A set of models and algorithms for the general unfolding analysis of preference/dominance data. Journal of Classification, 1: 147-186.

Dobson, A. J., 1990. An Introduction to Generalized Linear Models. Chapman & Hall, London.

Faith, D. P., 1983. Asymmetric binary similarity measures. Oecologia, 57: 287-290.

Faith, D. P., P. R. Minchin & L. Belbin, 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio, 69: 57-68.

Gabriel, K. R., 1982. Biplot. Pages 263-271 *in* S. Kotz & N. L. Johnson (ed.). Encyclopedia of Statistical Sciences. Vol. 1, Wiley, New York.

Gabriel, K. R. & C. L. Odoroff, 1990. Biplots in biomedical research. Statistics in Medicine, 9: 469-485.

Gittins, R., 1985. Canonical Analysis. A Review With Applications in Ecology. Springer-Verlag, Berlin.

Gower, J. C. & S. A. Harding, 1988. Nonlinear biplots. Biometrika, 75: 445-455.

Gower, J. C. & P. Legendre, 1986. Metric and Euclidean properties of dissimilarity coefficients. Journal of Classification, 3: 5-48.

Heiser, W. J., 1987. Joint ordination of species and sites: The unfolding technique. Pages 189-224 *in* P. Legendre & L. Legendre (ed.). Developments in Numerical Ecology. Springer-Verlag, Berlin.

Hermy, M. & P. J. Lewi, 1991. Multivariate ratio analysis. A graphical method for ecological ordination. Ecology, 72: 735-738.

Hill, M. O., 1973. Reciprocal averaging: An eigenvector method of ordination. Journal of Ecology, 61: 237-249.

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

Huisman, J., H. Olff & L. F. M. Fresco, 1993. A hierarchical set of models for species response analysis. Journal of Vegetation Science, 4: 37-46.

Israëls, A., 1992. Redundancy analysis for various types of variables. Statistica Applicata, 4: 531-542.

Jongman, R. H. G., C. J. F. ter Braak & O. F. R. van Tongeren, 1987. Data Analysis in Community and Landscape Ecology. Wageningen: Pudoc. New edition: 1994, Cambridge University Press, Cambridge.

Kingston, J. C., H. J. B. Birks, A. J. Uutala, B. F. Cumming & J. P. Smol, 1992. Assessing trends in fishery resources and lake

water aluminium from paleolimnoligical analyses of siliceous algae. Canadian Journal of Fisheries and Aquatic Sciences, 49: 116-127.

Lebreton, J. D., D. Chessel, M. Richardot-Coulet & N. Yoccoz, 1988. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. II. Variables de milieu qualitatives. Acta Oecologia Generalis, 9: 137-151.

Meulman, J., 1986. A Distance Approach to Nonlinear Multivariate Analysis. DSWO Press, Leiden.

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.

Sabatier, R., J.-D. Lebreton & D. Chessel, 1989. Multivariate analysis of composition data accompanied by qualitative variables describing a structure. Pages 341-352 *in* R. Coppi & S. Bolasco (ed.). Multiway Data Tables. North-Holland, Amsterdam.

Schiffman, S. S., M. L. Reynolds & F. W. Young, 1981. Introduction to Multidimensional Scaling. Theory, Methods and Applications. Academic Press, London.

Shi, G. R., 1993. Multivariate data analysis in palaeoecology and palaeobiogeography. A review. Palaeogeography, Palaeoclimatology, Palaeoecology, 105: 199-234.

Smilauer, P., 1992. CanoDraw. Ithaca, Microcomputer Power, New York.

Snoeijs, P. J. M. & I. C. Prentice, 1989. Effects of cooling water discharge on the structure and dynamics of epilithic algal communities in northern Baltic. Hydrobiologia, 184: 99-123.

Takane, Y., H. Bozdogan & T. Shibayama, 1987. Ideal point discriminant analysis. Psychometrika, 52: 371-392.

ter Braak, C. J. F., 1983. Principal components biplots and alpha and beta diversity. Ecology, 64: 454-462.

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. Ordination. Pages 91-173 *in* R. H. G. Jongman, C. J. F. ter Braak & O. F. R. van Tongeren (ed.). Data Analysis in Community and Landscape Ecology. Pudoc, Wageningen.

ter Braak, C. J. F., 1987b. The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio, 69: 69-77.

ter Braak, C. J. F., 1988. CANOCO – a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1). Agricultural Mathematics Group, Report LWA-88-02, Wageningen.

ter Braak, C. J. F., 1990a. Update Notes: CANOCO version 3.1. Agricultural Mathematics Group, Wageningen.

ter Braak, C. J. F., 1990b. Interpreting canonical correlation analysis through biplots of structural correlations and weights. Psychometrika, 55: 519-531.

ter Braak, C. J. F., 1992. Multidimensional scaling and regression. Statistica Applicata, 4: 577-586.

ter Braak, C. J. F. & C. W. N. Looman, 1986. Weighted averaging, logistic regression and the Gaussian response model. Vegetatio, 65: 3-11.

ter Braak, C. J. F. & C. W. N. Looman, 1994. Biplots in reduced-rank regression. Biometrical Journal, 36: 983-1003.

ter Braak, C. J. F. & I. C. Prentice, 1988. A theory of gradient analysis. Advances in Ecological Research, 18: 271-317.

ter Braak, C. J. F., S. Juggins, H. J. B. Birks & H. van der Voet, 1993. Weighted averaging partial least squares regression (WA-PLS): Definition and comparison with other methods for species-environment calibration. Pages 525-560 *in* G. P. Patil & C. R. Rao (ed.). Multivariate Environmental Statistics (chapter 25). North-Holland, Amsterdam.

Van der Meer, J., 1991. Exploring macrobenthos-environment relationship by canonical correlation analysis. Journal of Experimental Marine Biology and Ecology, 148: 105-120.

Varis, O., 1991. Associations between lake phytoplankton community and growth factors – a canonical correlation analysis. Hydrobiologia, 210: 209-216.

Whittaker, R. H., 1956. Vegetation of the Great Smoky Mountains. Ecological Monographs, 26: 1-80.

Whittaker, R. H., 1967. Gradient analysis of vegetation. Biological Reviews of the Cambridge Philosophical Society, 49: 207-264.

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

Yee, T. W. & N. D. Mitchell, 1991. Generalized additive models in plant ecology. Journal of Vegetation Science, 2: 587-602.

Appendix A

TABLE OF MEANS AND THE DISPLAY OF QUALITATIVE VARIABLES BY CENTROIDS IN REDUNDANCY ANALYSIS AND CANONICAL CORRELATION ANALYSIS.

Following the notation in ter Braak (1987a: section 5.9.3), let $\mathbf{Y}$ be an $m \times n$ matrix in which the kth row contains the centred abundance values of the kth species (*i.e.* $y_{k+} = 0$) and let $\mathbf{Z}_u$ and $\mathbf{Z}$ be $q \times n$ matrices in which the jth row contains the uncentred and row-centred values of the jth environmental variable, respectively. Define $\|\mathbf{R}\|^2 = \text{trace}(\mathbf{RR'})$, the sum of squares over all elements of the matrix $\mathbf{R}$, and $[\mathbf{R}]_r$ the matrix consisting of the first r columns of the matrix $\mathbf{R}$. Further, let $S_{11} = \mathbf{Y}\,\mathbf{Y'}$ to obtain canonical correlation analysis and $\mathbf{S}_{11} = \mathbf{I}$ for redundancy analysis. An interesting intermediate case between canonical correlation analysis and redundancy analysis (ter Braak, 1990b; ter Braak & Looman, 1994) is to set $\mathbf{S}_{11} = \text{diag}(\sigma_1^2, \ldots \sigma_m^2)$, with σ_k^2 the (estimated) error variance of the abundance of the kth species ($k = 1 \ldots m$). The latter choice yields error-weighted redundancy analysis.

If the environmental variables are qualitative, then $\mathbf{Z}_u$ is a matrix of dummy variables, one for each class indicating by the values 1 and 0 whether a site belongs or does not belong to the class, respectively. With $\mathbf{N}_c = \text{diag}(\mathbf{ZZ'})$, a diagonal $q \times q$ matrix with the number of sites in each class on the diagonal, the classes × species table of means is then given by

$$\mathbf{M}_c = \mathbf{N}_c^{-1}\,\mathbf{Z}_u\,\mathbf{Y'} = \mathbf{N}_c^{-1}\,\mathbf{Z}\,\mathbf{Y'} \qquad (A.1)$$

where the latter equality holds, because $y_{k+} = 0$. For producing a biplot of $\mathbf{M}_c$ in r dimensions (usually $r = 2$), we need $q \times r$ and $m \times r$ matrices $\mathbf{E}$ and $\mathbf{F}$ so that $\mathbf{M}_c \approx \mathbf{E}\,\mathbf{F'}$. An unweighted least-squares approximation by minimizing $\|\mathbf{M}_c - \mathbf{EF'}\|^2$ has the disadvantages that all classes receive equal weight, irrespective of their size, and that linear transformation of $\mathbf{Z}$ and of $\mathbf{Y}$ would change the solution. To circumvent these disadvantages, we take a weighted least-squares approximation with weight matrices $\mathbf{N}_c^2(\mathbf{ZZ'})^{-1}$ and $\mathbf{S}_{11}^{-1}$. This choice weighs the classes proportional to their size and makes the approximation criterion independent of linear transformations of $\mathbf{Z}$ and, for canonical correlation analysis, of $\mathbf{Y}$. In error-weighted redundancy analysis, it makes the criterion independent of linear rescaling of the abundances of each individual species (row of $\mathbf{Y}$). Thus, we seek the minimum over $\mathbf{E}$ and $\mathbf{F}$ of

$$\|(\mathbf{ZZ'})^{-1/2}\,\mathbf{N}_c\,(\mathbf{M}_c - \mathbf{EF'})\,\mathbf{S}_{11}^{-1/2}\|^2 =$$
$$\|(\mathbf{ZZ'})^{-1/2}\,\mathbf{ZY'}\mathbf{S}_{11}^{-1/2} - ([\mathbf{ZZ'}]^{-1/2}\,\mathbf{N}_c\mathbf{E})(\mathbf{S}_{11}^{-1/2}\mathbf{F})'\|^2 \qquad (A.2)$$

As follows from the Eckhart-Young theorem (*e.g.* Gittins, 1985), the minimum is obtained from the singular value decomposition

$$(\mathbf{ZZ'})^{-1/2}\,\mathbf{ZY'}\mathbf{S}_{11}^{-1/2} = \mathbf{Q}\,\mathbf{\Lambda}^{1/2}\,\mathbf{P'}, \qquad (A.3)$$

where $\mathbf{Q}$ and $\mathbf{P}$ are orthonormal matrices of order $q \times t$ with $t = \min(q,m)$ containing the singular vectors and $\mathbf{\Lambda}$ a diagonal matrix with the squared singular values on the diagonal, arranged in decreasing order ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_t \geq 0$). This singular value decomposition is one computational route for obtaining a canonical correlation analysis and redundancy analysis (ter Braak, 1987a: (5.33)). The minimum of (A.2) is $\lambda_{r+1} + \ldots + \lambda_t$ and is attained by setting

$$(\mathbf{ZZ'})^{-1/2}\,\mathbf{N}_c\mathbf{E} = [\mathbf{Q}]_r \text{ and } \mathbf{S}_{11}^{-1/2}\mathbf{F} = [\mathbf{P}\mathbf{\Lambda}^{1/2}]_r. \qquad (A.4)$$

Equivalently,

$$\mathbf{E} = \mathbf{N}_c^{-1}\,(\mathbf{ZZ'})^{1/2}\,[\mathbf{Q}]_r = \mathbf{N}_c^{-1}\,\mathbf{ZZ'C} = \mathbf{N}_c^{-1}\,\mathbf{ZX} = \mathbf{N}_c^{-1}\,\mathbf{Z}_u\mathbf{X} \qquad (A.5)$$

and

$$\mathbf{F} = \mathbf{S}_{11}^{1/2}[\mathbf{P}\mathbf{\Lambda}^{1/2}]_r = \mathbf{Y}\,\mathbf{X}, \qquad (A.6)$$

where, following Equations 5.36 and 5.37 from ter Braak (1987a), $\mathbf{C} = (\mathbf{ZZ'})^{-1/2}\,[\mathbf{Q}]_r$ and $\mathbf{X} = \mathbf{Z'}\,\mathbf{C}$ are the $q \times r$ and $n \times r$ matrices of canonical weights and site scores, respectively. Equation (A.5) shows that the optimal classes points are at the centroid of the scores and equation (A.6) shows that the optimal species points are the usual species scores. Equations (A.4) use biplot scaling 2, but can be changed to scaling 1 by moving $\mathbf{\Lambda}^{1/2}$ from the second to the first equation in (A.4). Equations (A.5) and (A.6) can then be modified accordingly.

This derivation assumes that all environmental variables are qualitative. For a mixture of quantitative and qualitative variables a combined table of correlations and class means requires approximation (Figure 1). In ter Braak (1987a) and ter Braak (1990b) it was shown that the weighted least-squares approximation of correlations lead to the same singular value decomposition. The weight matrices are also identical, except for the factor N_c^{-1}. Apart from notational problems, it is thus straightforward to formulate a joint optimization criterion.

What difference does it make whether classes are plotted as centroids or as arrows? The coordinates of the arrow heads for Z are the rows of the matrix (ter Braak 1987a: Equation 5.36)

$$\mathbf{E}_a = \mathbf{Z}\,\mathbf{X}, \qquad (A.7)$$

so that $\mathbf{E} = \mathbf{N}_c^{-1}\,\mathbf{E}_a$. However, for ease of comparison, arrows are usually given for standardized environmental variables ($\mathbf{Z}\,\mathbf{Z'} = \mathbf{I}$). Compared to an arrow for a standardized variable, we have the relation: "the class centroid is equal to the arrow times $\text{sd}(z)/\bar{z}$", where $\bar{z}$ and $\text{sd}(z)$ are the mean and the standard deviation of the uncentred, unnormalized environmental variable (ter Braak, 1988: (4.19)). [For this, note that $\bar{z} = \mathbf{N}_c/n$]. If a fraction p of the sites belongs to the class, $\text{sd}(z)/\bar{z} = \{(1 - p)/p\}^{1/2}$. The arrows and the centroids thus point in the same directions, but differ more and more in their relative length, the more the sizes of the classes differ among each other. Drawing an imaginary axis for a species and projecting arrows (standardized or not) for classes on this axis would not yield a correct ranking in terms of mean abundance. It is thus essential to plot the centroids for classes instead of arrows.

For qualitative variables that are in $\mathbf{Z}$, the centroids of the site scores that are linear combinations of the environmental variables, coincide with the centroids of the site scores that are weighted sums of the species scores (ter Braak, 1988: 53). Therefore, the positions of the centroids do not depend on the question which set of site scores is plotted.

Appendix B

EFFECTS, ARROWS AND CENTROIDS OF ENVIRONMENTAL VARIABLES.

Let $\mathbf{x}_s$ contain the site scores on the *s*th ordination axis, z_l the *l*th normalized (quantitative) environmental variable ($\mathbf{z}_l'\mathbf{z}_l = 1$) and $\mathbf{y}_k$ the abundance of the *k*th species. Further let r_{ls} be the correlation of the *l*th environmental variable with the *s*th axis having eigenvalue λ_s. The standard deviation of a variable is indicated by sd(.).

The regression of the ordination scores $\mathbf{x}_s$ on the normalized environmental variable $\mathbf{z}_l$ gives a regression coefficient that is equal to $\mathbf{x}_s'\mathbf{z}_l = sd(\mathbf{x}_s)\, r_{ls}$, *i.e.* $\lambda_s^{1/2} r_{ls}$ in scaling 1 and r_{ls} in scaling 2. These are precisely the coordinates of the arrow-head of an environmental variable in the biplot (Table Ib). For both scaling 1 and scaling 2, each arrow can thus be interpreted as the effect that the corresponding environmental variable has on the site scores (neglecting in other variables). The arrow length is the effect size.

This definition of effect size carries through from the site scores to the species abundance data for scaling 1, but not for scaling 2. This is because in scaling 1 the variance of the site scores of an axis is equal to the eigenvalue, which is, in its turn, equal to the variance in the species data explained by the axis. In contrast, axes have equal variance in scaling 2. To be precise, the sum of squares (ss) in the species data explained by $\mathbf{z}_l$ *via* ordination axis $\mathbf{x}_s$ is equal to the sum of squares explained by $\mathbf{x}_s$ times the fraction of that sum of squares explained by $\mathbf{z}_l$, *i.e.*

$$\sum_k ss(\mathbf{y}_k \mid \mathbf{x}_s, \mathbf{z}_l) = \sum_k ss(\mathbf{y}_k \mid \mathbf{x}_s)\, r_{ls}^2 = \lambda_s r_{ls}2 \qquad (B.1)$$

In scaling 1, the squared length of an arrow is thus equal to the sum of squares explained by the variable via the axes.

Another interpretation of the length of an arrow in scaling 2 is that it is the multiple correlation of the environmental variable with the axes. The length of the arrow in scaling 1 thus links up nicely with the aim of canonical ordination, namely to explain the species data via the ordination axes by the environment, whereas in scaling 2 it says how well the environmental variable is explained or displayed! The importance of the environmental variables for the ordination is thus better expressed in scaling 1 than in scaling 2.

The above discussion of quantitative variables generalizes to qualitative environmental variables by replacing regression by analysis of variance. The primary motivation for adding classes as centroids is that, together with the species points, they display the species × environmental table of means (Appendix A). The additional interpretation in terms of the site scores is: centroids are the result of a one-way analysis of variance of the ordination axis with respect to the qualitative variable. The percentage variance explained by this analysis of variance is a measure of the importance of this qualitative variable for an ordination axis (Lebreton *et al*., 1988). Consequently, as the importance of a qualitative variable for the species assemblage gets higher, the more its centroids spread out in the distance biplot.

Because an ordination diagram is a low-dimensional approximation, it does not necessarily display the full effect of a variable. It is therefore impossible to derive the full effect size from the ordination diagram. It is of interest to note that the full effect can be obtained in CANOCO 3.1 after choosing forward selection; it is the 'extra fit' listed prior to the first selection.