Ministerie van Landbouw en Visserij
Directoraat-Generaal Landbouw en Voedselvoorziening
Directie Landbouwkundig Onderzoek

GROEP LANDBOUWWISKUNDE

CANOCO - a FORTRAN program for canonical community ordination by [partial] [detrended] [canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1).

Cajo J.F. Ter Braak

Agricultural Mathematics Group
Box 100, 6700 AC Wageningen
The Netherlands

This report is reprinted (with permission, and with corrections and some additions) from the technical report with number 87 ITI A 11 of the TNO Institute of Applied Computer Science, Statistics Department Wageningen, which is the former affiliation of the author.

Technical report: LWA-88-02
January 1988

GLW
Postbus 100
6700 AC Wageningen

CONTENT

OVERVIEW

Aim

A common problem in community ecology and ecotoxicology is to discover how a multitude of species respond to external factors such as environmental variables, pollutants and management regime. Data are collected on species composition and the external variables at a number of points in space and time. Statistical methods available so far to analyse such data either assumed linear relationships or were restricted to regression analysis of the response of each species separately. To analyse the generally non-linear, non monotone response of a community of species, one had to resort to the data-analytic methods of ordination and cluster analysis - "indirect" methods that are generally less powerful than the "direct" statistical method of regression analysis. Recently, regression and ordination have been integrated into techniques of multivariate direct gradient analysis, called canonical (or constrained) ordination. The use of canonical ordination greatly improves the power to detect the specific effects one is interested in. One of these techniques, canonical correspondence analysis, escapes the assumption of linearity and is able to detect unimodal relationships between species and external variables. The computer program CANOCO is designed to make these techniques available to ecologists studying community responses. CANOCO can carry out most of the multivariate techniques described in Ter Braak (1987) and Ter Braak and Prentice (1988) using a general iterative ordination algorithm.

Researchers in other fields may find CANOCO useful as well, for example, to analyse percentage data/compositional data, nominal data or (dis)-similarity data in relation to external explanatory variables. Such use is explained in separate sections in the manual. CANOCO is particularly suited if the number of response variables is large compared to the number of objects.

Techniques covered

1. CANOCO is an extension of DECORANA (Hill, 1979). CANOCO formerly stood for canonical correspondence analysis (Ter Braak, 1986a, b) and included weighted averaging, reciprocal averaging/[multiple] correspondence analysis, detrended correspondence analysis and canonical correspondence analysis. The program has been extended to cover also principal components analysis (PCA) and the canonical form of PCA, called redundancy analysis (RDA). Redundancy analysis (Van den Wollenberg, 1977; Israëls, 1984) is also known under the names of reduced-rank regression (Davies and Tso, 1982), PCA of y with respect to x (Robert and Escoufier, 1976) and mode C partial least squares (Wold, 1982). For these linear methods there are options for centring/standardization by species and by sites and for the method of scaling the species and site scores for use in the biplot. The eigenvalues reported in PCA/RDA are fractions of the total variance in the species data (percentage variance accounted for). Principal coordinates analysis and canonical variates analysis are also available.

2. CANOCO can also carry out "partial" analyses in which the effects of particular environmental, spatial or temporal "covariables" are eliminated from the ordination. A partial analysis allows one to display the residual variation in the species data and to relate the residual variation to the variables one is specifically interested in. Partial canonical correspondence analysis is the appropriate technique for the analysis of permanent plot data or for the joint analysis of data from several locations.

3. CANOCO allows one to test statistically whether the species are related to supplied environmental variables. The test provided is a Monte Carlo permutation test (Hope, 1968). The effect of a particular environmental variable can be tested after elimination of possible effects of other (environmental) variables by specifying the latter as covariables. For the analysis of randomized-block experiments or data from several locations, there is an option to restrict the permutation to permutations among samples-within-blocks or samples-within-locations.

4. CANOCO provides an alternative method of detrending which is intended to solve the problems reported to occur with the method used in DECORANA. CANOCO allows one to remove polynomial relations between ordination axes (up to order 4). Use of the old method of detrending by segments (Hill and Gauch, 1980) in partial and canonical analyses is not recommended.

5. CANOCO has an option for nonstandard analyses. In one possibility, the reciprocal averaging algorithm is modified so that at each iteration the species and/or site scores are replaced by rank numbers. This procedure circumvents what is known as the "deviant sample/rare species problem" in correspondence analysis.

Data input

CANOCO can read species data, environmental variables and covariables that are either in Cornell condensed format or in full format. The machine readable copy of the analysis can be used again as input for subsequent analyses. This possibility allows one, for example:
- to use principal components extracted from environmental data as input for a later canonical analysis of species data,
- to extract more than four ordination axes - simply by supplying the extracted ordination axes as covariables in a subsequent analysis.

Output options

CANOCO can supply:
- means, variances and correlations of environmental variables,
- eigenvalues, the percentages of variance accounted for by the biplot of species-environment relations,
- scores of species and sites on the ordination axes,

- canonical coefficients or regression coefficients of environmental variables with associated t-values,
- correlations of environmental variables with the ordination axes,
- scores of environmental variables for constructing the arrows in the species-environment biplot,
- centroids (weighted averages) of environmental variables in the ordination diagram (for variables with positive values). In particular, classes of nominal environmental variables are more naturally displayed by their centroid in the ordination diagram than by arrows. This option is also useful for displaying the results of a cluster analysis in an ordination diagram.

CANOCO allows interactive data analysis: results of an analysis can be displayed at the terminal and after inspection the analysis can be pursued, for example,

- by changing from an indirect gradient analysis to a direct gradient analysis,
- by dropping environmental variables,
- by reading other environmental variables to be related to the current ordination axes or to be used in further canonical analyses,
- by changing detrending options,
- by changing scaling options of the ordination scores.

Practical information

CANOCO is written in standard FORTRAN 77 and can be supplied on 5.25 inch diskette for IBM-compatible PC's, on magnetic tape (800/1600 bpi, ASCI-code) or via BITNET/EARN. On an IBM-compatible PC with 640 Kb, CANOCO can analyse ca. 750 samples, 600 species, 60 environmental variables and 100 covariables (see Table 3.4). The one-time costs are specified on the order form. Researchers from countries with valuta problems may send in a request for a free copy. Documentation will be sent with the program.

1. INTRODUCTION

1.1 General objective

CANOCO, an acronym of **CANO**nical Community Ordination, is designed for data analysis in community ecology. Researchers in other disciplines should consult Table 1.1 for the terminoloy used in this manual. Canonical ordination is a class of techniques for relating the composition of species communities to their environment. Data analysis by canonical ordination can either be exploratory or confirmatory. When used in an exploratory way, it leads to an ordination diagram of samples, species and environmental variables, which optimally displays how community composition varies with the environment. When used in a confirmatory way, it leads to statistical tests of the effects of particular environmental variables on community composition taking into account effect of other variables. The theory of this is given in Ter Braak and Prentice (1987) and Jongman et al. (1987).

1.2 Models, methods and algorithm

Canonical ordination is a combination of ordination and multiple regression. Ordination techniques such as principal components and correspondence analysis (= reciprocal averaging) are commonly used to reduce the variation in community composition to the scatter of samples and species in an ordination diagram. Subsequently the diagram is interpreted with help of external data, for example, by calculating correlation coefficients between environmental variables and ordination axes, or by multiple regression of the ordination axes on environmental variables. A difficulty here is that the ordination axes are just particular orthogonal directions in the ordination diagram. Other directions may well be better related to the environmental variables. Canonical ordination is a solution to this difficulty. The regression model is inserted in the ordination model. As a result the ordination axes appear in order of variance explained by linear combinations of environmental variables.

The ordination technique of correspondence analysis was introduced in ecology by way of the reciprocal averaging algorithm (Hill, 1973) or, for abundance data, the two-way weighted averaging algorithm. It is an iterative ordination algorithm: from initial arbitrary sample scores, species scores are obtained, from which new samples scores are derived, from which new species scores are derived, and so on.
Principal components analysis can be obtained by a similar algorithm by taking weighted sums, instead of weighted averages (Jongman et al. 1987: section 5.3). Canonical ordination techniques can be obtained by carrying out multiple regressions within the iterative algorithm: each time new sample scores are derived, they are regressed on the environmental variables (instead of just once after an ordination). CANOCO uses this kind of iterative ordination algorithm.

The resulting species scores are parameters of response curves of species with respect to the ordination axis. In linear methods to which

Table 1.1 Terminology used in CANOCO, with commonly used synonyms.

Community	a set of species occurring together in a sample.
Sample	sampling unit, individual, object, site.
Species	response variable, dependent variable in a regression equation, internal variable.
Abundance/response	value of a response variable, usually positive or 0; proximity.
Environmental variable	explanatory variable (of prime interest), independent variable in a regression equation, external variable, stimulus variable, treatment variable.
Covariable	concomitant variable, background variable, explanatory variables corresponding to incidental parameter or nuisance parameters, block factor in experimental design.
Indirect gradient analysis	internal analysis, "factor analysis", unconstrained ordination, unconstrained multidimensional scaling, possibly followed post-hoc by an regression analysis on external variables.
Direct gradient analysis	external analysis, canonical ordination, ordination constrained by external variables, constrained multivariate regression, reduced-rank regression.
Ordination	see indirect gradient analysis.
Ordination axis	eigenvector, latent variable, theoretical explanatory variable.
Ordination diagram	scatter plot of the eigenvector scores; used both for biplots and joint plots.
Canonical ordination	an ordination in which the axes are constrained to be linear combinations of environmental variables.
Canonical axis	an ordination axis that is constrained to be a linear combination of environmental variables.

Table 1.1: continued

Eigenvalue	importance measure of an ordination axis (section 4.2).
Species score	eigenvector coefficient; loading in PCA, center of species curve in CA and DCA.
Sample score	Value of eigenvector in a sample.
Biplot	an ordination diagram of two kinds of entities, e.g. species and environmental variables, which has particular rules of interpretation because it is based on a bilinear model. Interpretation proceeds by projecting points on directions defined by arrows in the biplot (e.g. Fig. 4.2).
Joint plot	an ordination diagram of two kinds of entities based on a weighted averaging method.
Linear method	method based on a linear model, e.g.: linear regression, multiple regression, principal components analysis, redundancy analysis.
Weighted averaging method	method based on a unimodal response model (= unimodal trace line) of which the optimum (mode, ideal point) is estimated by weighted averaging, e.g. correspondence analysis.

principal components analysis belongs, the response "curves" are straight lines (Fig. 1.1) and the species score is the slope parameter. In weighted averaging methods to which correspondence analysis belongs, the response curves are unimodal (Fig. 1.2) and the species score can be considered as the center of the curve or, for "central species", the optimum of the curve.

On the basis of the linear and unimodal response models in Figs. 1.1 and 1.2, we introduce six types of data-analysis problems (see Table 1.2). When there is just a single, known explanatory variable, the slope of each line in Fig. 1.1 would have been estimated by simple linear regression and the center of each curve in Fig. 1.2 by weighted averaging "regression". Estimating such species parameters is a regression problem. If there are some samples for which the value of the explanatory variable is missing, the values can be estimated from the species composition of those samples by seeking for each such sample the value of the environmental variable that is most likely to give the observed species composition as judged by the response curves in the Figures. This is a calibration problem: linear calibration in Fig. 1.1 and weighted averaging "calibration" in Fig. 1.2. When all values of the explanatory variable are missing, one could still attempt to construct a theoretical variable that best fits the species data according to a linear model or a unimodal model. This is an ordination problem. The theoretical variable is the first ordination axis found by the iterative ordination algorithm. The algorithm is essentially a converging sequence of regressions and calibrations. The sample scores are the values that the theoretical variable takes in the samples. The theoretical variable/ordination axis has no environmental basis. Canonical ordination is ordination with the additional constraint that the ordination axis must be a linear combination of environmental variables. Canonical ordination is thus a particular form of constrained ordination. It has an environmental basis. One can also apply ordination to the variation in the community data that remains after known environmental variables have been fitted by regression. Ordination of residual variation is called partial ordination: the effect of particular variables is "partialled out" (eliminated) from the ordination. The variables that are partialled out are called covariables. Finally, when axes of a partial ordination are constrained to be linear combinations of particular environmental variables, we obtain a partial constrained ordination. CANOCO can perform the techniques listed in Table 1.2 in the columns "linear/least squares" and "unimodal/weighted averaging". They are obtainable as special cases of the iterative ordination algorithm used in CANOCO (section 8). Under particular conditions the weighted averaging methods are a close approximation of the methods listed in the column "maximum likelihood" (Ter Braak 1986a). These are more formal statistical methods which require heavy computation and which are therefore less attractive for routine use. One cannot obtain them with CANOCO. But CANOCO is useful to obtain starting values for the maximum likelihood methods.

It may come as a surprise that canonical correlation analysis (Gittins, 1985) is missing in Table 1.2, as this is the standard linear multivariate technique for relating two sets of variables (in our case, the set of species and the set of environmental variables). In its place comes

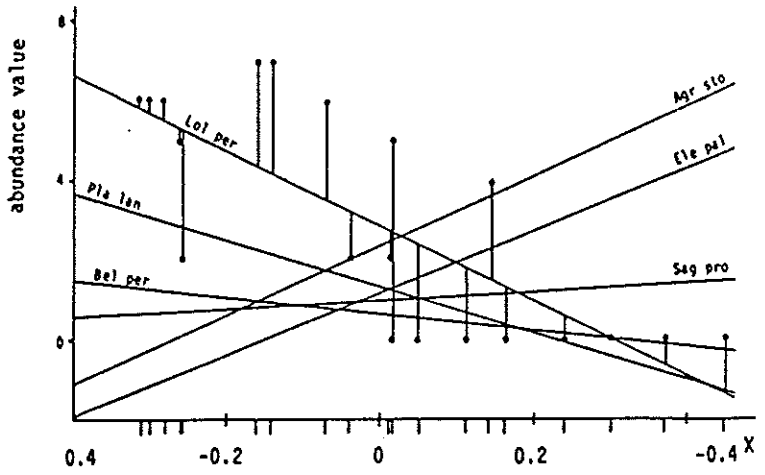

Fig.1.1 Straight lines for the abundance of six plant species along the first axis of principal components analysis (x), applied to the dune meadow data (Table 2.1). In linear ordination methods, the species score is the slope of the line of the corresponding species. The sample scores are shown by ticks below the abscissa. Also shown are the abundance values of Lolium perenne and their deviations from the fitted straight line. Abbreviations of species names are given as underlying in Table 2.1.

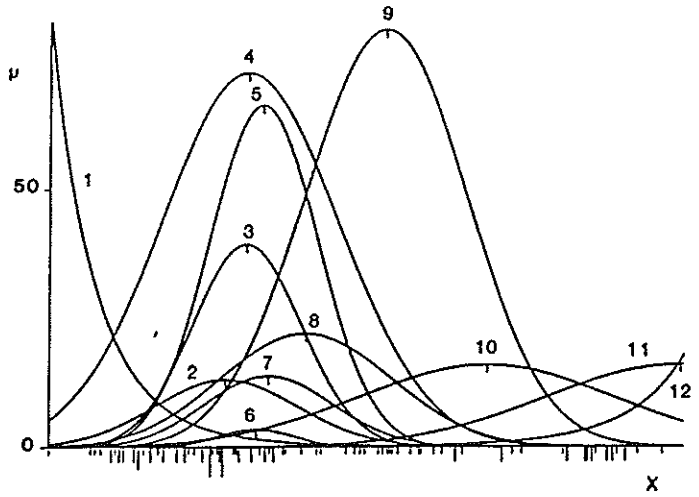

Fig.1.2 Unimodal response curves for the count (μ) of 12 species of wolf spiders in a dune area along the first axis of detrended correspondence analysis (x) applied to data of Van der Aart and Smeenk-Enserink (1975). The position of the optimum of each curve is indicated by a tick near the maximum. In weighted averaging methods the species score is a rough estimate of the center of its response curve. The sample scores are indicated by ticks below the abscissa (length proportional to number). The species are:
1 = Pardosa lugubris; 2 = Zora spinimana; 3 = Pardosa nigriceps; 4 = Trochosa terricola; 5 = Pardosa pullata; 6 = Arctosa lutetiana; 7 = Aulonia albimana; 8 = Alopecosa cuneata; 9 = Pardosa monticola; 10 = Alopecosa accentuata; 11 = Alopecosa fabrilis; 12 = Arctosa perita (after Ter Braak, 1985).

Table 1.2 Classification of gradient analysis techniques by type of problem, response model and method of estimation. The techniques listed under "linear/least-squares" and "unimodal/weighted averaging" can be carried out with CANOCO.

	RESPONSE MODEL: linear	unimodal	
METHOD OF ESTIMATION:	least-squares	maximum likelihood	weighted averaging
TYPE OF PROBLEM:			
regression	multiple regression	Gaussian regression	weighted averaging of site scores (WA)
calibration	linear calibration; "inverse regression"	Gaussian calibration	weighted averaging of species scores (WA)
ordination	principal components analysis (PCA)	Gaussian ordination	correspondence analysis (CA)[5]; detrended correspondence analysis (DCA)
constrained[1] ordination	redundancy analysis (RDA)[4]	Gaussian canonical ordination	canonical correspondence analysis (CCA); detrended CCA
partial ordination[2]	partial components analysis	partial Gaussian ordination	partial correspondence analysis; partial DCA
partial constrained ordination[3]	partial redundancy analysis	partial Gaussian canonical ordination	partial canonical correspondence analysis; partial detrended CCA

1) = constrained multivariate regression = canonical ordination
2) = ordination after regression on covariables
3) = constrained ordination after regression on covariables = constrained partial multivariate regression
4) = "reduced-rank regression" = "PCA of y with respect to x"
5) = multiple correspondence analysis = dual scaling = homogeneity analysis

the lesser known technique of redundancy analysis. The most important difference between these techniques is that redundancy analysis can analyse any number of species, whereas in canonical correlation analysis the number of species must be less than n-q with n the number of samples and q the number of environmental variables. The latter restriction makes canonical correlation analysis unattractive for most studies in commuy ecology. More details about the difference are given in section 7.3.

1.3 Terminology

The terminology (Table 1.1) used in CANOCO stems from typical applications in community ecology. CANOCO operates on species, environmental variables and covariables (Table 1.1). Ordination is applied to the species data, which are typically data on abundances or incidences of a set of species in a set of samples. The variation in the species data is to be explained via the ordination axes by environmental variables and covariables. Environmental variables are the explanatory variables of prime interest. Covariables are concomitant variables whose effect must be partialled out when estimating the effects of the environmental variables. When one wants a constrained ordination, the number of environmental variables and covariables must be smaller than the number of samples. For a greater number of explanatory variables, constrained ordination and unconstrained ordination coincide. There is no restriction on the number of species.

Of course, there is nothing special about the terms used. They have a formal meaning only (Table 1.1). For example, if one wants an ordination of environmental variables, then this is easily done by entering the name of the data file containing these variables at the point where one usually specifies a file with species data.

1.4 CANOCO's efficiency for ordination of community data

CANOCO is particular efficient for ordination of "sparse" data sets (data containing many zero values compared to the number of nonzero values). It is quite common in community data that the average number of species present in a sample is in the order of 10-30, whereas the total number of species in the data set is in the order of 100-1000. Because the abundance of an absent species is, of course, zero, the data are sparse. By not storing zero values, a large saving of memory space and of computer time is achieved. The iterative ordination algorithm used by CANOCO (section 8) is specially designed so as to make storage of zero values unnecessary. It uses methods of calculation that are efficient for sparse data. This design makes CANOCO efficient also for ordination of nominal data (section 7.2).

1.5 Outline of the manual

The sections 2, 3 and 4 are the kernel of the manual. Section 2 (input) should be consulted before running CANOCO. Section 3 helps the user to run CANOCO and section 4 describes the output. Section 6 is intended to

give the user a flavour of what can be achieved with CANOCO. The remaining sections including section 5 give further information on the capabilities of CANOCO and are somewhat more technical. The usual types of data in community ecology are abundance data and incidence data. The theory for applying ordination to these types of data is described in extenso in Ter Braak and Prentice (1987) and Jongman et al. (1987). Some theory for other types of data (percentage data/compositional data, nominal data and dissimilarity data) is given in section 7. It also discusses the relationships among the linear methods of regression analysis, redundancy analysis and principal components analysis. The description of the iterative ordination algorithm (section 8) requires a firm understanding of mathematics, but is not essential for the general user. Sections 9 and 10 may be of help when installing CANOCO on a computer and when the user wants to make small modifications to the functioning of the program.

2. DATA INPUT

The data input of CANOCO consists of species data and, optionally, environmental data and covariables. CANOCO will ask for the names of the computer files containing these types of data. This may be the same file for all three types, but this is not advisable. Often it is convenient to prepare one file for species data and a second file for both the environmental data and the covariables.
CANOCO can read data that are either in Cornell condensed format or in full format. Species data which contain many zero abundance values (absences of species) are most efficiently stored on a file in Cornell condensed format. Environmental data are most efficiently stored on a file in full format, unless they contain many nominal variables. Cornell condensed format is also convenient when analyzing nominal variables by (multiple) correspondence analysis (section 7.2).

In the preparation of data files it is useful to know that CANOCO has options for deleting samples, species and environmental variables from the analysis, some facilities for transformation of the species data, but that it has no facilities for transformation of the environmental data (see section 3.5 and 3.6). Note that the order or the numbering of covariables is important in Monte Carlo tests based on restricted permutation (see Q40 in section 3.8).

2.1 Cornell condensed format

For users who are familiar with the program DECORANA (Hill, 1979), we can be short: CANOCO can use the same data files as DECORANA. The format as required by CANOCO is more flexible in one respect: it is allowed to specify the number of couplets per line in free format on line 3 instead of in columns 69-70 on line 2.

Table 2.1 shows the dune meadow data used as example data in Jongman et al (1987) and Ter Braak (1986b) and Table 2.2 shows the same data in Cornell condensed format. In this format each sample and each species is given a number. Each line of the data begins with a sample number followed by a number of "couplets", each consisting of a species number and an abundance value. For example, line 9 of the data file shown in Table 2.2 begins with the number 3, which means that this line gives data of the sample which is given number 3, shortly sample 3. This number is followed by the couplet 2 4.0, which says that the species which is given number 2 (species 2) has abundance value 4.0 in sample 3. The next couplet says that species 4 has abundance value 7.0 in sample 3, etc. There are five couplets on this line. The next line again begins with a 3 followed by only one couplet (32 6.0), as an example that the number of couplets may vary among lines and that species numbers do not need to be arranged in increasing order. The data of sample 3 are on three lines. Species whose number does not appear on these lines are absent in sample 3. The program will assume that their abundance value is zero.

By inspecting Table 2.2 one sees that the sample numbers are increasing. For CANOCO the samples must indeed be arranged in increasing order, but they don't need to be consecutive. After sample 30 there is "notional" sample 0, which indicates the end of the data. Thereafter, there

Table 2.1 Dune meadow vegetation data of the island of Terschelling, The Netherlands (subset from M. Batterink and G. Wijffels, unpubl.). The table shows the abundance values (response values) of 33 plant species (rows) in 21 sample plots (columns of one digit width). The values are on a 1-9 scale and replace the original codes of the Blaun-Blanquet scale. A blank (space) denotes absence.

samples	111111112223
	12345678901234567890
species	
1 Achillea millefolium	13 222 4 2
2 Agrostis stolonifera	48 43 45447 5
3 Aira praecox	2 3
4 Alopecurus geniculatus	272 53 85 4
5 Anthoxanthum odoratum	432 4 4 4
6 Bellis perennis	3222 2 52
7 Bromus hordaceus	4 32 2 4
8 Chenopodium album	1
9 Cirsium arvense	2
10 Eleocharis palustris	4 458 4
11 Elymus repens	44444 6
12 Empetrum nigrum	2
13 Hypochaeris radicata	2 2 5
14 Juncus articulatus	44 33 4
15 Juncus bufonius	2 4 43
16 Leontodon autumnalis	52233332352222 2 562
17 Lolium perenne	75652664267 2
18 Plantago lanceolata	555 33 2 3
19 Poa pratensis	445423444444 2 143
20 Poa trivialis	2765645454 49 2
21 Potentilla palustris	22
22 Ranunculus flammula	2 2222 4
23 Rumex acetosa	563 2 2 3
24 Sagina procumbens	5 22 242 3
25 Salix repens	335
26 Trifolium pratense	252
27 Trifolium repens	52125223633261 22
28 Vicia lathyroides	12 1
29 Brachythecium rutabulum	2226222244 44 634
30 Calliergonella cuspidata	4 3 3
31 Hippophae rhamnoides	1 21
32 Poa annua	3364 2 232 3 4
33 Ranunculus acris	232 2 11

Table 2.2 The data of Table 2.1 in Cornell condensed format. Each sample and each species is given a number (see Table 2.1). Abbreviations of species names below the data are given as underlining in full names in Table 2.1. The names of the samples show the number used for the corresponding samples in Jongman et al. (1987). In the analysis sample 20 named PAS SAMP and the species 31, 32 and 33 are made passive. The file has in the example of section 3.9 the name "DUNEMEAD.SPE".

```
SPECIES - DUNE MEADOW DATA (M. BATTERINK AND G. WIJFFELS, 1983)
(I10,X,5(I4,F5.0))
 5
         1    1  1.0   11  4.0   17  7.0   19  4.0   20  2.0
         1   32  3.0
         2    1  3.0    4  2.0    6  3.0    7  4.0   11  4.0
         2   16  5.0   17  5.0   19  4.0   20  7.0   27  5.0
         2   32  3.0
         3    2  4.0    4  7.0    6  2.0   11  4.0   16  2.0
         3   32  6.0
         3   17  6.0   19  5.0   20  6.0   27  2.0   29  2.0
         4    2  8.0    4  2.0    6  2.0    7  3.0    9  2.0
         4   11  4.0   16  2.0   17  5.0   19  4.0   20  5.0
         4   24  5.0   27  1.0   29  2.0   32  4.0
         5    1  2.0    5  4.0    6  2.0    7  2.0   11  4.0
         5   16  3.0   17  2.0   18  5.0   19  2.0   20  6.0
         5   23  5.0   33  2.0   26  2.0   27  2.0   29  2.0
         6    1  2.0    5  3.0   16  3.0   17  6.0   18  5.0
         6   19  3.0   20  4.0   23  6.0   26  5.0   27  5.0
         6   29  6.0   33  3.0
         7    1  2.0    5  2.0    7  2.0   15  2.0   16  3.0
         7   17  6.0   18  5.0   19  4.0   20  5.0   23  3.0
         7   26  2.0   27  2.0   29  2.0   32  2.0   33  2.0
         8    2  4.0    4  5.0   10  4.0   14  4.0   16  3.0
         8   17  4.0   19  4.0   20  4.0   22  2.0   24  2.0
         8   27  2.0   29  2.0
         9    2  3.0    4  3.0   11  6.0   14  4.0   15  4.0
         9   16  2.0   17  2.0   19  4.0   20  5.0   23  2.0
         9   24  2.0   27  3.0   29  2.0   31  1.0   32  2.0
         9   33  2.0
        10    1  4.0    5  4.0    6  2.0    7  4.0   16  3.0
        10   17  6.0   18  3.0   19  4.0   20  4.0   27  6.0
        10   28  1.0   29  2.0   32  3.0
        11   13  2.0   16  5.0   17  7.0   18  3.0   19  4.0
        11   24  2.0   27  3.0   28  2.0   29  4.0   32  2.0
        12    2  4.0    4  8.0   15  4.0   16  2.0   20  4.0
        12   23  2.0   24  4.0   27  3.0   29  4.0
        13    2  5.0    4  5.0    8  1.0   15  3.0   16  2.0
        13   19  2.0   20  9.0   22  2.0   24  2.0   27  2.0
        13   32  3.0
        14    2  4.0   10  4.0   16  2.0   21  2.0   22  2.0
        14   27  6.0   30  4.0   33  1.0
        15    2  4.0   10  5.0   14  3.0   16  2.0   21  2.0
        15   22  2.0   27  1.0   29  4.0   33  1.0
        16    2  7.0    4  4.0   10  8.0   14  3.0   20  2.0
        16   22  2.0   29  4.0   30  3.0
        17    1  2.0    3  2.0    5  4.0   13  2.0   16  2.0
        17   18  2.0   19  1.0
        20    6  5.0   19  4.0   23  3.0
        28    6  2.0   16  5.0   17  2.0   18  3.0   19  3.0
        28   25  3.0   27  2.0   28  1.0   29  6.0   31  2.0
        28   32  4.0
        29    3  3.0    5  4.0   12  2.0   13  5.0   16  6.0
        29   24  3.0   25  3.0   27  2.0   29  3.0   31  1.0
        30    2  5.0   10  4.0   14  4.0   16  2.0   22  4.0
        30   25  5.0   29  4.0   30  3.0
         0
ACH MIL AGR STO AIR PRA ALO GEN ANT ODO BEL PER BRO HOR CHE ALB CIR ARV ELE PAL
ELY REP EMP NIG HYP RAD JUN ART JUN BUF LEO TAU LOL PER PLA LAN POA PRA POA TRI
POT PAL RAN FLA RUM ACE SAG PRO SAL REP TRI PRA TRI REP VIC LAT BRA RUT CAL CUS
HIP RHA POA ANN RAN ACR
......1 ......2 ......3 ......4 ......5 ......6 ......7 ......8 ......9 .....10
.....11 .....12 .....13 .....14 .....15 .....16 .....17              PAS SAMP
                                                          .....18 .....19 .....20
```

are abbreviations of the species names, for example ACH MIL stands for Achillea millefolium which is species 1. Each abbreviation must take 8 positions including spaces and there are 10 abbreviated names on a line. (In the output the 8 positions are printed as 2 × 4 positions with a space inbetween.) Species 10 is thus ELE PAL and species 22 is EMP NIG (see legend Table 2.1). The highest species number in this data set is 33. CANOCO expects therefore four lines with abbreviations of species names (even if not all of these species are actually present in the data). Thereafter CANOCO expects sample names in the same format as the species names. (Samples names are printed in the output without an extra space). From Table 2.2 we see, for example, that sample 20 has the name "PAS SAMP" (abbreviation of passive sample; in our example in section 3.9 this sample is made passive) and sample 28 has the name " .....18" (which demonstrates that numbers in names have meaning for the user only).

The first three lines of a file in Cornell condensed format must look like this:

Line 1 must contain a title. The title is reproduced (except for the last space) in the output to remind the user which data where used in the analysis.

Line 2 must contain a FORTRAN format which specifies how the data are stored on a line. Any FORTRAN format of maximal 80 positions which specifies the reading of the sample number and a number of couplets (species numbers and abundances) from a single line of 80 positions is acceptable. For example, the FORTRAN format in Table 2.2 is (I10, 1X, 5(I4, F5.0)). The format should always be enclosed between brackets. The letters I, X and F should be in upper case.
- "I10" means that the sample number is in the first 10 positions of a line (with the last digit in position 10),
- "1X" means that the eleventh position is skipped,
- "5(I4, F5.0)" means that there are a maximum of five couplets on a line, each with 4 positions for the species number and 5 positions for its abundance value. The species number (and sample number) must be a whole number or Integer, whence "I4"; the abundance value is considered as a real value and must be given as an "F" followed by the number of positions, whence "F5.0" (the ".0" is required and guarantees that the values in the data file are being read without modification).

Line 3 contains the maximum number of couplets on a line. It is in "free format", which means that it can appear anywhere on the line.

Line 4 is the beginning of the data. The data end with a notional sample 0 (zero), without further data. The species names and sample names follow thereafter as described above.

Table 2.3 Environmental variables in Cornell condensed format used as explanatory variables for the species data in Table 2.1. The variables are numbered as follows: 1 = Thickness of A1 horizon; 2 = moisture; 3 = quantity of manure; 4 = hayfield; 5 = haypasture; 6 = pasture; 7 = Standard Farm; 8 = Biodynamic Farm; 9 = Hobby Farm; 10 = Nature Manegement. The sample numbers and names are as in Table 2.2, except that sample 20 is missing. For explanation see text.

```
ENVIRONMENTAL DATA IN CONDENSED FORMAT - DUNE MEADOW DATA
 (I5,1X,I1,F5.0,3(2X,I1,F2.0),2X,I2,F2.0)
  5
    1 1   2.8  2 1   3 4   5 1    7 1
    2 1   3.5  2 1   3 2   5 1    8 1
    3 1   4.3  2 2   3 4   5 1    7 1
    4 1   4.2  2 2   3 4   5 1    7 1
    5 1   6.3  2 1   3 2   4 1    9 1
    6 1   4.3  2 1   3 2   5 1    9 1
    7 1   2.8  2 1   3 3   6 1    9 1
    8 1   4.2  2 5   3 3   6 1    9 1
    9 1   3.7  2 4   3 1   4 1    9 1
   10 1   3.3  2 2   3 1   4 1    8 1
   11 1   3.5  2 1   3 1   6 1    8 1
   12 1   5.8  2 4   3 2*  5 1    7 1
   13 1   6.0  2 5   3 3   5 1    7 1
   14 1   9.3  2 5   3 0   6 1   10 1
   15 1  11.5  2 5   3 0   5 1   10 1
   16 1   5.7  2 5   3 3   6 1    7 1
   17 1   4.0  2 2   3 0   4 1   10 1
   28 1   4.6* 2 1   3 0   4 1   10 1
   29 1   3.7  2 5   3 0   4 1   10 1
   30 1   3.5  2 5   3 0   4 1   10 1
    0
      A1MOISTURE  MANUREHAYFIELDHAYPASTU PASTURE      SF       BF       HF       NM
.......1 ......2 ......3 ......4 ......5 ......6 ......7 ......8 ......9 .....10
......11 .....12 .....13 .....14 .....15 .....16 .....17                PAS SAMP
                                                              .....18 .....19 .....20
```

It is also allowed to specify, as in DECORANA (Hill, 1979), the maximum number of couplets in the positions 69-70 of line 2 (or, if it is a number of one digit, in column 70). In this case line 3 is the beginning of the data. (Note that some FORTRAN-implementations allow to read more than 80 positions per line; consult your FORTRAN-manual for this and see section 10).

The Cornell condensed format can also be used for environmental data and covariables. The description of the format given above is still valid. Just replace the term "species" by "environmental variable" or by "covariable" and "abundance value" by "value". Values of environmental variables or of covariables may be negative. Note however that numbers of variables that do not appear in a sample receive in the calculations the value 0 (zero), which is highly undesirable in most cases. For missing values, one should insert a best possible guess or the mean value of the corresponding variable. For nominal environmental variables, Cornell condensed format may be quite convenient (a nominal variable specifies a classification of samples; its "values" are thus classes).

In the case of the dune meadow data five "environmental" variables were recorded at each site, two of which are nominal. The (semi-) quantitative variables are (1) A1: thickness of the A1 horizon (in mm), (2) MOISTURE: moisture content of of the soil scored on a five-point scale, (3) MANURE: quantity of manuring, also scored on a five-point scale. The nominal variables are (4) agricultural use, with the three classes hayfield, haypasture and pasture, and (5) management regime, with the four classes standard farming (SF), bio-dynamical farming (BF), hobby farming (HF) and nature management (NM). In CANOCO each class of a nominal variable must be coded as a separate variable. Table 2.3 shows the data in Cornell condensed format. Column 1 gives the sample numbers, which correspond to those in Table 2.2 except that sample 20 is absent. Columns 3, 5 and 7 give the values of the variables A1, MOISTURE, MANURE, which are the variables 1, 2 and 3, respectively. Columns 8 and 9 specify the agricultural use (variable 4 = hayfield, variable 5 = haypasture and variable 6 = pasture) and columns 10 and 11 the management regime of the meadows (variable 7 = SF, variable 8 = BF, variable 9 = HF, variable 10 = NM). For example, sample 3 is a haypasture of a standard farm because the variables 5 (= haypasture) and 7 (= standard farming) have the value 1 (columns 9 and 11). The other classes of agricultural use and management regime are absent in sample 3 resulting in the value 0. The asterisks in Table 2.3 are there as a reminder for us that the preceding value is an insertion for a missing value. The asterisks in the data file present no problem in this case because their positions on the line are skipped by virtue of the FORTRAN format specified on line 2.

When species data are in Cornell condensed format, the value 0 in a couplet may denote that the species is present with a small quantity. In the analysis these zeroes can be distinguished from true zeroes (absences), if desired, by choosing a particular transformation of the species data (see example in section 3.6: Question Q24). With environmental data or covariable data the value 0 in a couplet cannot be distinguished from zeroes in absent couplets.

Table 2.4 Environmental variables of Table 2.3 in full format. For explanation see text. The file has in the example of section 3.9 the name "DUNEFUL.ENV".

```
ENVIRONMENTAL DATA IN FULL FORMAT- DUNE MEADOW DATA
(I5,F5.0,1X,2F3.0,3X,3F2.0/18X,4F2.0)
 10
   1  2.8   1  4    0 1 0
                  1 0 0 0
   2  3.5   1  2    0 1 0
                  0 1 0 0
   3  4.3   2  4    0 1 0
                  1 0 0 0
   4  4.2   2  4    0 1 0
                  1 0 0 0
   5  6.3   1  2    1 0 0
                  0 0 1 0
   6  4.3   1  2    0 1 0
                  0 0 1 0
   7  2.8   1  3    0 0 1
                  0 0 1 0
   8  4.2   5  3    0 0 1
                  0 0 1 0
   9  3.7   4  1    1 0 0
                  0 0 1 0
  10  3.3   2  1    1 0 0
                  0 1 0 0
  11  3.5   1  1    0 0 1
                  0 1 0 0
  12  5.8   4  2*   0 1 0
                  1 0 0 0
  13  6.0   5  3    0 1 0
                  1 0 0 0
  14  9.3   5  0    0 0 1
                  0 0 0 1
  15 11.5   5  0    0 1 0
                  0 0 0 1
  16  5.7   5  3    0 0 1
                  1 0 0 0
  17  4.0   2  0    1 0 0
                  0 0 0 1
  28  4.6*  1  0    1 0 0
                  0 0 0 1
  29  3.7   5  0    1 0 0
                  0 0 0 1
  30  3.5   5  0    1 0 0
                  0 0 0 1
   0  0.0   0  0    0 0 0
                  0 0 0 0
    A1MOISTURE  MANUREHAYFIELDHAYPASTU PASTURE      SF      BF      HF      NM
.....1 .....2 .....3 .....4 .....5 .....6 .....7 .....8 .....9 .....10
.....11 .....12 .....13 .....14 .....15 .....16 .....17                PAS SAMP
                                                              .....18 .....19 .....20
```

2.2 Full format

Full format means that all values must be given, including all zero values, as is common in most statistical packages. In the full format of CANOCO a sample number appears once, followed by its values of all variables in a fixed format and a fixed order. Abbreviations of the names of the variables and of the samples are given at the end of the data, just as in Cornell condensed format. Table 2.4 shows the environmental data of Table 2.3 in full format. Variable 1 is A1, variable 2 is MOISTURE, ..., and variable 10 is NM (for Nature Management). The values of the 10 variables of sample 3 are given on the lines 8 and 9: the first 6 variables on line 8 and 4 more variables on line 9. In sample 3 the A1 horizon was thus 4.3 mm, its moisture content was scored the value 2, etc., ... Continuing on line 9 we see that sample 3 is of a standard farm: its value is 1 whereas the other management classes have the value 0 in sample 3.

We see the same sample numbers in Table 2.4 as in Table 2.3. After sample 30 there is again a "notional" sample 0, which indicates the end of the data. But, because each sample occupies two lines in this example, the notional sample 0 must also occupy two lines. For clarity its values are given the value 0, but these values could equally well be replaced by spaces ("blanks"). The rule is that the notional sample 0 in a full format file must occupy the same number of lines as the other samples. How many lines a sample occupies is implicit in the FORTRAN-format on line 2 and the number of variables specified on line 3.
Missing values are not allowed. For missing values one may insert a best possible guess or the mean value of the corresponding variable.
The first three lines of a data file in full format must look like this:

Line 1 must contain a title (see section 2.1).
Line 2 must contain a FORTRAN format which specifies how the data are stored for a sample. Any FORTRAN format of maximal 80 positions which specifies the reading of a sample number (by an "I-format") and the values (in "F-formats") of all variables is acceptable, with the proviso that one cannot read more than 80 positions per line. For example, the FORTRAN format in Table 2.4 is (I5, F5.0, 1X, 2F3.0, 3X, 3F2.0/18X, 4F2.0).
- "I5" means that the sample number is in the first 5 positions of the first line for a sample (with the last digit in position 5).
- "F5.0" reads a value from the next five positions (A1).
- "1X" means that position 11 is skipped.
- "2F3.0" reads two values of three positions each (MOISTURE and MANURE).
- "3X, 3F2.0" skips the next three positions and reads three values of two positions each (HAYFIELD, HAYPASTURE and PASTURE).
- "/18X" means go to the beginning of the next line and skip the first 18 positions of this line; thereafter
- "4F2.0" reads four values of two positions each (SF, BF, HF, NM).

A FORTRAN format of a full format file thus contains precisely one I-format.

Table 2.5 Nominal data to be subjected to multiple correspondence analysis in Cornell condensed format. The file gives data on three nominal variables A, B and C with 3, 4 and 2 categories, respectively. The value 1 in each couplet is omitted (see text for explanation).

```
DATA ON THREE NOMINAL VARIABLES TO BE SUBJECTED TO CA
    (I2, 1X, 3(I2, F1.0))
  3
 1 12 21 32
 2 13 23 32
 3 12 21 32
 4 11 22 31
 5 11 23 31
 6 13 24 32
 7 12 24 31
 8 12 23 31
 9 13 21 31
10 12 22 32
11 11 24 31
 0
(this line is for names of the non-existing variables 1-10)
 A1        A2        A3
 B1        B2        B3        B4
 C1        C2
.......1.......2.......3.......4.......5.......6.......7.......8.......9.......10
.......11
```

Line 3 contains in free format the number of variables read by F-formats in the FORTRAN format of line 2. (The data file may contain more variables, if they are skipped by the FORTRAN format).

Line 4 is the beginning of the data. The data end with a notional sample 0, which occupies as many lines as a normal sample. The names of the variables and sample names follow thereafter, as in Cornell condensed format.

It is also allowed to specify the number of variables in the positions 69-70 of line 2 (or, if it is a number of one digit, in column 70). In this case line 3 is the beginning of the data.
Full format can also be used for species data and for covariables. Just replace "variable" by "species" or "covariable" and "value" by "abundance value" or "value of covariable".

Important: CANOCO will transform full format species data internally to condensed format for reasons described in section 1.4. Because zero abundance values are not stored in the computer memory, zero values cannot be transformed to a non-zero value later on in the program, even if requested so in Q24-Q25. If the user wants to transform a zero abundance value to a non-zero value in full format species data, the zeroes should be specified in the full format data by a small quantity, e.g. 0.01.

2.3 Presence/absence and nominal data for ordination

When storing presence/absence data or purely nominal data in condensed format, each couplet consists of the number of a variable and the value 1. The 1's are then more of less redundant and can be omitted if the data are to be subjected to ordination only. Table 2.5 shows an example. The FORTRAN format on line 2 specifies that each couplet consists of three positions, namely two for the number of the variable present and one for the abundance value. As there is a blank after each variable number, CANOCO will read the value 0 for each abundance. Because each couplet in a condensed format file is stored, these zeroes are stored as well and can be transformed to the required value 1 by typing in response to question Q24 (section 3.6):

0	1
1	1
-1	0

However, when the data are in this form, they cannot be entered as environmental data or covariable data.

2.4 Linking up samples in different data files

Important: Species data, environmental data and covariable data are commonly stored in different files. CANOCO determines which samples in different files correspond to the same physical sampling unit on the basis of the sample number (not the sample name). The sample names in the species

data file serve to label the samples in the output. The sample names in the environmental data and covariable data are used merely to check whether the sample numbers in different files have the same name. If differences in names are detected, an error message is given but the program continues regardless.

3. TERMINAL DIALOGUE

3.1 How to activate CANOCO

CANOCO is designed as an interactive computer program: the program asks questions and the user types answers in response. At first use, this dialogue between CANOCO and the user can best take place at a computer terminal (console or personal computer). Later on, it may be useful to run CANOCO in batch. How to activate CANOCO depends on how CANOCO is installed at your computer; see Table 3.1. The person installing CANOCO (see Installation Notes) is requested to complete Table 3.1.

3.2 Input and output

During the dialogue CANOCO asks for names of input and output files. The answers must be valid names of at most 40 characters. CANOCO itself creates one file which gives an annotated copy of the answers entered at the terminal. This file may help the user to do further analyses, for example, in batch jobs (see Table 3.1). Input for CANOCO are the answers entered at the terminal, and one to three data files (see DATA INPUT). Output of CANOCO is the output file that contains a comprehensive copy of the terminal dialogue and also all numerical results of CANOCO or a part of it, depending on the user's answers. This file has a maximum of 132 positions per line. Additional output files, containing a machine readable copy of the ordination results, are optional. One to many of such files can be created after each activation of CANOCO. These files have a maximum of 80 positions per line and can be used as data input for CANOCO instead of a file containing environmental data or covariables. They are also useful as input to programs for preparing ordination diagrams. All output can also be displayed at the terminal.

3.3 Ways to answer the questions

In posing a question CANOCO indicates the range of valid answers by ending the question with a phrase like this:

```
Range of valid answers: 0 [1] 3
Type your answer or merely press RETURN for default, indicated by [].
```

In this example valid answers are obtained by typing one of the values 0, 1, 2 and 3 followed by pressing the RETURN key (sometimes termed the ENTER key). If one merely presses the RETURN key, the implied answer is the "default" answer indicated by "[]" in the range. In the example the default answer is the value 1. If the range is indicated like this:

```
Range of valid answers: [1] 3
```

then the default value is 1 and coincides with the minimum value of valid answers. It the range is given as "1 [3]", the default value is 3 and

Table 3.1. How to activate CANOCO at the terminal and in batch.

To be filled in by the person who installs CANOCO

Computer :.....

To activate CANOCO at a terminal type:.....

The answers entered while running CANOCO appear in annotated form on a file named: .....

To activate CANOCO in batch, you must submit a batch job. The file which specifies the batch job must begin with the following lines

(command)	(comment)
.....	.....
.....	.....
.....	.....
.....	.....

The job is submitted by typing:

.....

Programs that use the machine readable copy of CANOCO/DECORANA are

(name)	(purpose)	(contact)
.....	.....	.....
.....	.....	.....
.....	.....	.....

coincides with the maximum value of valid answers. If an invalid answer is given, the question is posed again. Real values like 2.5 are permitted only when the values in the range have a decimal point, e.g.

Range of valid answers: 1.0 [3.0]

If the program indicates that the answer must consist of two values, the values must be entered on the same line and be separated by one or more blanks (spaces) and/or by a comma. In this case the range of each of the values is indicated separately like this:

Ranges of valid answers: [-1] 10 , and [0] 10

By pressing RETURN, the implied answers are thus the values -1 and 0. If a comma is used in the answer, missing entries are replaced by the default value. In the example the answer ",3" is interpreted as the values -1 and 3.

CANOCO starts with the question

```
TYPE 0 FOR INPUT FROM CONSOLE
     1 FOR INPUT FROM FILE
Range of valid answers: [0] 1
Type your answer(s) or merely press RETURN for default, indicated
by [].
```

By pressing RETURN (or typing 0) CANOCO continues the terminal dialogue by asking further questions (see the next sections). But if the answer is 1, then the user needs to answer only one more question from the terminal:

```
TYPE NAME OF FILE WITH ANSWERS TO THE QUESTIONS
```

If the user types, for example, CANOCO.CON, then the program reads the answers to subsequent questions from the file CANOCO.CON and the analysis proceeds automatically. Commonly, the file specified here is a modification of the file which contains the annotated copy of answers of an earlier analysis. An example of such a file is given later on (Table 3.3). It may happen that the file contains too few answers, for example, because some of the answers were found invalid. In that case, the terminal dialogue with user starts again at the question left unanswered.

3.4 Questions to specify the type of analysis and in- and output files

Which questions are posed depends on the type of analysis. The questions following those of the previous section are numbered as Q1, Q2, ..., in the order in which they are posed. The answer to, for example, question Q5 is indicated by "Q5 = ...", where "...." is the answer.

An example is given in section 3.9. The file with the annotated copy of answers entered from the terminal contains the answers from question Q1 onwards (see Table 3.3).

Q1. TYPE NAME OF OUTPUT FILE

If the name is a valid name, CANOCO attempts to "open" a new file with this name and, if opened, write output to this file. The lines of this output file have a maximum length of 132 characters.

Q2. *** TYPE OF ANALYSIS ***

MODEL	GRADIENT ANALYSIS		
	INDIRECT	DIRECT	HYBRID
LINEAR	1=PCA	2= RDA	3
UNIMODAL	4= CA	5= CCA	6
"	7=DCA	8=DCCA	9
	10=NON-STANDARD ANALYSIS		

TYPE ANALYSIS NUMBER
Range of valid answers: 1 [5] 10

The analysis types are arranged in a 3 × 3 table of type-of-model by type-of-gradient-analysis. For more information than can be supplied here consult Jongman et al. (1987) and Ter Braak and Prentice (1987).
The first column refers to indirect gradient analysis techniques: ordination techniques, which search for major gradients in the species data irrespective of any environmental variables. The entries under this heading are

1 = PCA = Principal Components Analysis
4 = CA = Correspondence Analysis
7 = DCA = Detrended Correspondence Analysis

PCA assumes a linear model (row 1) for the relationship between the responses of each species and the ordination axes; CA and DCA assume a unimodal model (rows 2 and 3) for the relationship between the responses of each species and the ordination axes. Ordination axes can be thought of as being theoretical environmental variables or underlying gradients. The linear model is fitted by the method of two-way weighted summation which leads to the least-squares solution. The unimodal model is fitted by the method of two-way weighted averaging. Use of DCA is advised if an ordination by CA shows the arch effect, i.e. if the sample scores on the second ordination axis are approximately a quadratic function of the sample scores on the first axis. (The arch effect is also termed the Guttman effect.) Use of PCA is advised in particular if in ordinations by CA or DCA the range of the sample scores is less than 1.5 SD (see section 4.3). This advice is applicable to each choice between techniques of rows 1, 2 and 3.

Other names for CA are reciprocal averaging and - outside ecology, in particular when analysing nominal response variables (section 7.2) - dual scaling, optimal scaling, homogeneity analysis and multiple correspondence analysis (Gifi, 1981; Greenacre, 1984).

The choice between variants of PCA, like non-centred PCA, species-centred PCA, standardized PCA, double centred PCA, is relegated to section 3.6 as these variants are obtainable by transformation of the species data. Principal coordinates analysis can also be obtained as a variant of PCA (see section 7.4).

CANOCO calculates in a simple run at most four ordination axes. See Q7 for a method to obtain more ordination axes with CANOCO.

The second column refers to (multivariate) direct gradient analysis techniques (canonical ordination). They attempt to explain the species responses by ordination axes that are constrained to be linear combinations of supplied environmental variables. The ordination diagram obtained from a direct gradient analysis has therefore a known environmental basis. The entries under this heading are

2 = RDA = Redundancy Analysis
5 = CCA = Canonical Correspondence Analysis
8 =DCCA = Detrended Canonical Correspondence Analysis

When CCA is applied to nominal response variables it is termed redundancy analysis for qualitative variables (section 7.2; Israëls, 1984).

The maximum number of constrained ordination axes (= canocial axes) is in general equal to the number of environmental variables, unless "detrending" is in force (see for this exception Q9). Because CANOCO calculates in a single run at most four ordination axes, CANOCO will in general determine four constrained ordination axes, unless the number of environmental variables (q) is less than 4. If q is less than 4, CANOCO will extract, after the q constrained ordination axes, one or more unconstrained ordination axes (see below).

The third column refers to hybrid direct/indirect gradient analysis techniques. If the user chooses a technique from this column, CANOCO will ask later on, how many canonical axes are to be extracted. If two such axes are required, for example, the first two ordination axes will be "canonical", i.e. be constrained to be linear combinations of supplied environmental variables and the third and fourth ordination axis will be unconstrained, apart from being uncorrelated to the first two ordination axes. The unconstrained ordination axes represent residual variation in the species data that remains after extracting the constrained axes, and are therefore "partial" ordination axes. Another method to obtain partial ordination axes is by specifying covariables (see Q6).

Analysis number 10 stands for nonstandard analysis in which the user can specify unusual options or unusual combinations of options (see section 5).

The methods of row 1 will be called linear methods while the methods of rows 2 and 3 will be called weighted averaging methods, in accordance with the terminology in Ter Braak and Prentice (1987).

Q3. TYPE NAME OF FILE WITH SPECIES DATA

Ordination (PCA, CA, DCA, RDA, CCA, etc.) is applied to the data of the file specified here. In community ecology it are typically the species data, but if one wants, for example, a PCA of environmental data, the environmental data file should be specified here.

If the name is a valid name, CANOCO will attempt to read the species data from the file. The data can be either in Cornell condensed format or in full format (see DATA INPUT).

Note that the data should not contain negative values if a weighted averaging method (CA, DCA, CCA, DCCA) is chosen; if a negative value is encountered, CANOCO stops with an error message saying so.

Q4. TYPE 1 IF THE ORDINATION AXES - AFTER EXTRACTION - ARE TO BE RELATED TO ENVIRONMENTAL VARIABLES
ELSE TYPE 0
Range of valid answers: [0] 1

This question is posed only if an indirect gradient analysis technique is chosen at the first time Q2 appears. If the answer is 1, a file with environmental data is asked for (Q5) and CANOCO will calculate, among other things, correlation coefficients between the unconstrained ordination axes and the environmental variables of this file. Because environmental variables entered in this way do not influence the ordination axes (these remain unconstrained), the result is a "passive" analysis of environmental variables. Another method to obtain passive analyses is via Q35.

Q5. TYPE NAME OF FILE WITH ENVIRONMENTAL DATA

This question is posed only for analysis techniques requiring environmental data (see Q2) or if Q4 = 1. The data of the file specified here are used to interpret or to constrain the ordination of the data of Q3. In community ecology, it are typically environmental data. In general the file should contain "external" explanatory variables, i.e. variables by which one wants to explain the variation in the data specified in Q3. Explanation proceeds by way of a multiple regression of each ordination axis on the explanatory variables and by way of correlation coefficients. In a partial RDA and CCA the file should contain the explanatory variables of prime interest.

If the name is a valid name, CANOCO will attempt to read the environmental data from the file. The data can either be in Cornell condensed format or in full format. Here the user may also specify a file containing the machine readable copy of a previous analysis. The sample scores on the ordination axes will then be treated as the values of four variables named AX1, AX2, AX3 and AX4.

The file with environmental data may contain more variables than can actually be analysed by CANOCO, provided the excess variables are deleted later on in Q20.

```
Q6. TYPE 1 IF YOU HAVE COVARIABLES, ELSE TYPE 0
    *EXPLANATION* COVARIABLES ARE:
                  VARIABLES WITH KNOWN OR UNINTERESTING EFFECTS ON THE
                  SPECIES.
                  THEIR EFFECTS ARE ELIMINATED WHEN EXTRACTING ORDINATION
                  AXES.
    Range of valid answers:   [0]   1
```

If the answer is 1, then the types of analyses in Q2 need the prefix "Partial": the effect of covariables is partialled out from the ordination diagram. With covariables, CANOCO will give an ordination of the residual variation in the species data that remains after fitting the effects of the covariables. The ordination axes will be made uncorrelated to the covariables. Further, environmental variables (if present) will be regressed on the covariables and the residuals of these multiple regressions will take the place of the original environmental values. In this way, the effect of the environmental variables on the species is "corrected" for the effect that the covariables have on the species. Constrained ordination axes will therefore represent the effect that is "uniquely" attributable to the environmental variables- and not to (linear combinations of) covariables. With environmental variables in the analysis, covariables play the rôle of concomitant regressors in the multiple regression of the ordination axes on the explanatory variables specified in Q5. See sections 4.11 and 6 for examples of the use of covariables.

```
Q7. TYPE NAME OF FILE WITH COVARIABLES
```

This question is posed if Q6 = 1. See Q6 for explanation. If the name is a valid name, CANOCO will attempt to read the covariables from the file. See Q5 for the data formats allowed. By specifying a machine readable copy of a previous analysis CANOCO can extract further axes beyond the first four: if, for example, the covariables are the first four ordination axes of the species data, then the four ordination axes to be extracted will be made uncorrelated to these covariables and will thus be equivalent to ordination axes 5 to 8 of the previous analysis.

The file with covariables may contain more variables than can actually be analysed by CANOCO, provided de excess variables are deleted later on in Q22.

Q8. TYPE NUMBER OF CANONICAL AXES (1, 2 OR 3)
Range of valid answers: 1 [2] 3

This question is posed only for hybrid gradient analyses. See Q2 for explanation.

Q9. TYPE 1 FOR DETRENDING BY SEGMENTS
2 FOR DETRENDING BY 2ND ORDER POLYNOMIALS []
3 FOR DETRENDING BY 3RD ORDER POLYNOMIALS
4 FOR DETRENDING BY 4TH ORDER POLYNOMIALS
Range of valid answers: 1 [2] 4

This question is asked in DCA, DCCA and analysis numbers 9 and 10. Detrending is a method for removing the arch effect in CA and CCA (see Q2). Detrending-by-segments is the method of detrending proposed by Hill and Gauch (1980) and used in the computer program DECORANA (Hill, 1979). Minchin (1986) found that this method sometimes flattens out some of the variation associated with one of the underlying gradients. He ascribed this to an instability in the detrending-by-segments method (see also Kenkel and Orlóci, 1986). Detrending-by-polynomials is a more stable method of detrending. In the usual reciprocal algorithm of CA, trial site scores for a particular axis are made uncorrelated to the ordination axes already extracted in each iteration step. With detrending-by-polynomials they are also made uncorrelated to k-th order polynomials of the axes already extracted (k = 2, 3 or 4) and to first-order cross products of these axes.

When the arch effect crops up, the second CA-axis is approximately a quadratic function (= a second-order polynomial) of the first CA-axis. Detrending by second-order polynomials therefore specifically removes the arch effect. But this may not be enough because when there is a dominant first gradient, the third CA-axis is also a function of the first axis, namely a cubic function; and the fourth axis is a quartic function, etc., which may also obscure a true second underlying gradient. As the eigenvalues of these polynomial axes steadily decrease, detrending by fourth-order polynomials is presumably sufficient in most applications.

In DCCA and partial DCA, the method of detrending-by-segments is unattractive on theoretical grounds, but the method of detrending-by-polynomials can be modified into an acceptable method (see Appendix A). Use of detrending-by-segments is therefore not recommended in DCCA and partial DCA. When detrending-by-polynomials is used in a direct gradient analysis, then the number of canonical axes that can be extracted is less than without detrending. Less than four canonical axes can be extracted if there are less than 10, 13 or 16 environmental variables depending on whether the order of polynomials is 2, 3 or 4 respectively. Detrending is, however, almost never needed in CCA if only a few environmental variables are included in the analysis. Moreover if the arch effect does occur in a CCA,

it is an indication that some environmental variable is superfluous.

```
** Questions Q10 - Q13 are posed only if detrending-by-     **
** segments is asked for and in some nonstandard analyses. **
```

```
Q10. SPECIFY NUMBER OF SEGMENTS FOR USE IN THE DETRENDING PROCESS
     Range of valid answers:  10   [26]  46
```

This question is familiar to users of DECORANA (Hill, 1979). The default value is 26. The maximum permissible value is 46.

```
Q11. IS NONLINEAR RESCALING OF AXES REQUIRED?
     TYPE Ø (NO RESCALING), OR NUMBER OF TIMES TO BE DONE
     Range of valid answers:   0   [4]  20
```

This question has the same effect as the corresponding one in DECORANA (Hill, 1979). As in DECORANA, the default value is 4. The nonlinear rescaling of an ordination axis attemps to equalize the breadth of species response curves along the axis by means of equalizing the within-sample variances of the species scores. For this purpose a heuristic method is used in which the axis is divided into small segments; segments with samples with a small within-sample variance are expanded whereas segments with samples with a large within-sample variance are contracted. For further details see Hill (1979).

```
Q12. SPECIFY RESCALING THRESHOLD
     Range of valid answers:   [0.0]   100.0
```

Hill (1979) writes: "If the rescaling threshold is set to t, then axes with length less than t SD will not be rescaled, while those with length greater than t will be rescaled. The default value is t = 0". Here SD stand for the Standard Deviation unit, a measure of the length of an ordination axis compared to the average breadth of the species' response curves; (see also section 4.5 around equation (4.7)).

```
Q13. TYPE NUMBER (1-4) OF AXES FOR SPECIES-ENVIRONMENT BIPLOT
     Range of valid answers:   1   [2]   4
```

Answer here the number of axes of a planned ordination diagram. The question is needed when detrending-by-segments is in force, because the ordination axes are then in general slightly correlated. The optimal biplot

scores for the environmental variables will therefore depend on the number of axes chosen.

```
Q14. *** SCALING OF ORDINATION SCORES ***
     1 = SAMPLE  SCORES ARE WEIGHTED MEAN SPECIES SCORES []
     2 = SPECIES SCORES ARE WEIGHTED MEAN SAMPLE SCORES
     3 =  SYMMETRIC SCALING
     Range of valid answers:   [1]   3
```

This question is posed for weighted averaging methods but not if detrending-by-segments is in force (because then Q14 = 1 is implied). This question is needed because there is some arbitrariness in CA and derived methods of how to scale the sample scores with respect to the species scores (Ter Braak 1985; Heiser 1986). It was shown in Ter Braak (1985) that under particular conditions CA provides an approximate solution to a unimodal model. When such a model holds, the distances in the ordination diagram between sample points and a species point are inversely related to the abundance of the corresponding species. The scaling of odination scores asked for in this question influences sample-species distances in the ordination diagram. Unfortunately, CA gives no clue to what the optimal scaling is (this is an awkward consequence of the approximate nature of the solution provided by CA) and CANOCO contains no facility to automatically choose the optimal scaling (see Ter Braak (1985) for a proposal for a solution to this problem). The choice of scaling is the less critical the higher the eigenvalues of the ordination axes. Limited guidance is as follows: Answer 1 is standard in DECORANA and assumes that some species' optima lie outside the range of the sample scores. Answer 1 ensures that the range of the species scores is greater than that of the sample scores. Answer 2 assumes that the species' optima all lie inside the range of the sample scores (which is unrealistic in many ecological applications). Answer 3 is a compromise between 1 and 2.

Note that, if the answer is 1 and covariables are present, the sample scores that are "weighted mean species scores" are also made uncorrelated to the covariables. For more details about the scaling see sections 4.5 - 4.6.

```
Q15. *** SCALING OF ORDINATION SCORES ***
     1 = EUCLIDEAN DISTANCE BIPLOT []
     2 = COVARIANCE BIPLOT
     3 = SYMMETRIC SCALING
     Range of valid answers:   [1]   3
```

This question, posed for linear methods, addresses the same arbitrariness noted in the previous question. The answer to this question affects the inter-sample and inter-species interpretation of the ordination diagram,

but the joint interpretation of the sample and species points by the rules of the biplot is unaffected. Answer 1 leads to a biplot which is optimal for interpreting distances between samples as these approximate Euclidean distances in species-space (see e.g. Ter Braak, 1983). Answer 2 leads to a biplot which is optimal for interpreting angles between arrows of species as these angles approximate (linear) correlations between species (for a somewhat more precise statement see Corsten and Gabriel, 1976 or Jongman et al., 1987, section 5.3.4). Answers 1 and 2 are asymmetric scalings similar to those in Q14, whereas a symmetric scaling can be obtained by the answer 3. For more details about the scaling see sections 4.5 - 4.6.

Q16. TYPE 1 FOR A MACHINE READABLE COPY OF THE SOLUTION
Range of valid answers: [0] 1

For explanation see next question and section 3.2.

Q17. TYPE NAME OF FILE FOR MACHINE READABLE COPY

The question is posed if the previous answer is affirmative. If the name is a valid name, CANOCO attemps to "open" a new file with this name. This question is repeated later on if further analyses are wished. If the same name is entered then, CANOCO will not open a new file and will copy output to the existing file. If requested later on, CANOCO will write the ordination results to this file.

Q18. TYPE NAME OF FILE FOR MACHINE READABLE COPY OF THE ENVIRONMENTAL SCORES

N.B. This question is deleted in the latest version of CANOCO. All ordination is written to the file specified in Q17.

This question is posed if there are environmental variables in the analysis and Q16 = 1. The remarks of the previous question apply also to this file. If requested later on, CANOCO will write part of the ordination results to this file: biplot scores of environmental variables, centroids of environmental variables in the ordination diagram and the sample scores which are linear combinations of environmental variables.

3.5 Questions to omit samples and to manipulate environmental variables and covariables

---

Q19. ENTER NUMBER (NOT NAMES) OF SAMPLES TO BE OMITTED
ONE AT A TIME, ENDING LIST WITH A ZERO
Range of valid answers: [0] n

---

n = highest sample number in the species data

CANOCO asks this question after reading the file with the species data. Type only one sample number per line. For example, if samples 4, 7 and 10 are to be omitted, then these numbers should be entered as follows:

```
 4
 7
10
 0
```

If no samples are to be omitted, then it suffices to press RETURN. Samples can also be omitted at a later stage (see Q28). The advantage of doing it here, is that omitted samples are skipped when reading the environmental data and covariables.

Q20. ENTER NUMBERS (NOT NAMES) OF ENVIRONMENTAL VARIABLES
TO BE OMITTED ONE AT A TIME, ENDING LIST WITH A ZERO
Range of valid answers: [0] q

---

q = highest number of environmental variable

CANOCO asks this question just before the actual reading of the file with the environmental data (if there is one). Type one number per line as in the previous question. Deleted variables do not occupy data space in the computer. If a nominal variable has k classes, only $k-1$ dummy variables should be in the analysis to avoid multicollinearity among the environmental variables. The user should therefore delete one of the k dummy variables at this point; which one is arbitrary from the mathematical point of view (but, for numerical stability it is advisable to delete a class with many samples; and for ease of interpretation of regression coefficients it may be helpful to delete the class which is the natural reference class). If this is forgotten, CANOCO will automatically delete the highest class number of each nominal variable. The automatic procedure costs extra computation.

Q21. *** INTERACTIONS OF ENVIRONMENTAL VARIABLES ***
ENTER PAIRS OF NUMBERS OF ENVIRONMENTAL VARIABLES IF YOU WISH TO DEFINE PRODUCT VARIABLES
ENTER -1 Ø IF NO (FURTHER) PRODUCT VARIABLES ARE DESIRED
Range of valid answers: [-1] q, and [0] q

---

q = highest number of environmental variable

CANOCO also asks this question before the actual reading of the file with the environmental data. Type two numbers per line only. For example, suppose that a data file contains 20 environmental variables, number 1-20. Suppose that variable 2 is MOISTURE and variable 3 is MANURE; then entering

```
 2   3
 2   2
 2  22
-1   0  (or merely press RETURN)
```

has the effect that CANOCO creates three new variables with numbers 21, 22 and 23. Variable 21 is obtained by calculating for each sample the product of its MOISTURE value and its MANURE value. Variable 22 will contain squared moisture values and variable 23 will contain $(\text{MOISTURE})^3$. It is also possible to use numbers of variables that were deleted in the previous question.

By defining product variables, the user can investigate in very much the same way as in multiple regression analysis whether the effect of one variable depends on the value of another variable (see Jongman et al., 1987, section 3.5.4). In other words, this is a way to investigate interaction of effects. In the example just given, the effect that MANURE has on the species can be shown to depend on the value of MOISTURE if the first eigenvalue of the analysis turns out to be considerably higher than in the analysis without this product variable and if the t-value associated with this product variable is appreciably larger than 2 in absolute value. Inclusion of squared variables may alleviate the restriction that only linear combinations of environmental variables are considered in the analyses provided by CANOCO. The user should, however, be cautious in defining too many product variables, to avoid "data dredging".

After reading, CANOCO standardizes the environmental variables and their products (if defined), to mean 0 and variance 1.

Q22. ENTER NUMBER (NOT NAMES) OF COVARIABLES TO BE OMITTED ONE AT A TIME, ENDING LIST WITH A ZERO
Range of valid answers: [0] p

---

p = highest number of covariable

This question is analogous to Q20, but is asked now for covariables, when present. The remarks on nominal variables in Q20 also apply to the present question.

```
Q23. *** INTERACTIONS OF COVARIABLES ***
     ENTER PAIRS OF NUMBERS OF COVARIABLES IF YOU WISH TO DEFINE
     PRODUCT VARIABLES
     ENTER -1 Ø IF NO (FURTHER) PRODUCT VARIABLES ARE DESIRED
     Range of valid answers:  [-1]  p,  and  [0]  p
```

p = highest number of covariables

This question is analogous to Q21 but is asked now for covariables, when present. Squares (and products) of covariables may be useful in partial CA or partial DCA to prevent that the ordination axes from being a quadratic function of the covariables. This may happen if the covariables represent a long gradient in the species data and subsequent gradients are much shorter. See the DETRENDING question Q9.

After reading the covariables, CANOCO makes the covariables mutually uncorrelated by the Gram-Schmidt orthogonalization process (Rao, 1973: section 1a.4). If environmental variables are present, they are each regressed on the covariables and their values are replaced by the residuals of these regressions (without an extra standardization).

3.6 Questions to specify transformation of species data

The questions in this section are posed in order of their execution by CANOCO. This order settles any ambiguity in case of order dependency.

```
Q24. *** TRANSFORMATION OF SPECIES DATA ***
     TYPE -1 Ø IF NO TRANSFORMATION IS REQUIRED [ ]
     TYPE -2 Ø FOR THE SQUAREROOT-TRANSFORMATION
     TYPE -3 Ø FOR LN(Y+C)-TRANSFORMATION
     OR
     ENTER COUPLETS OF OLD AND NEW VALUES FOR PIECEWISE
     LINEAR TRANSFORMATION, ENDING WITH -1 Ø
     IF NO TRANSFORMATION IS DESIRED, MERELY PRESS RETURN
    Range of valid answers: -3.0 [-1.0] 999.9, and xxx [0.0] 999.9
```

xxx = -999.9 in linear methods
xxx = 0.0 in weighted averaging methods

The transformation that is chosen is applied to all species values (in general terms: to all response variables). If a logarithmic transformation is chosen by typing -3 0, the value of c is asked for next (Q25). Natural logarithms are taken. No transformation is applied to zero values unless they are explicitly included in a Cornell condensed format file (section 2).

The piecewise linear transformation works as in DECORANA (Hill, 1979). The following description of this transformation is copied from the

DECORANA manual. A typical transformation might be

```
 0   1
 2   2
 5   3
10   4
20   5
-1   0  (or merely press RETURN)
```

The negative number -1 serves to terminate the transformation data, and it must be followed by a dummy value such as 0. The meaning of this transformation is that a quantity 0 in the data is transformed to 1, 2 to 2, 5 to 3, etc.
For other numbers the transformation is interpolated linearly. Thus 6.9 is transformed to

3.0 + (6.9-5.0)*(4.0-3.0)/(10.0-5.0) = 3.38.

Non-integer values can be entered in the transformation, so that

```
20.3   5.2
```

would be a perfectly acceptable couplet.

Values outside the range of the transformation are converted to the same values as the extreme values of the transformation. Thus in the example considered above, numbers bigger than 20 would all be transformed to 5. Likewise, if the transformation

```
 1.2    1.2
 2.3    2.3
-1.0    0.0
```

is entered, all numbers less than 1.2 would be transformed to 1.2, all numbers greater than 2.3 would be transformed to 2.3, and numbers between 1.2 and 2.3 would be transformed to themselves (i.e. left unaltered)."
"Three restrictions should be noted:

1. Negative numbers cannot be considered for transformation, as any negative number automatically terminates the transformation data. [But, one may transform to negative numbers in linear methods].
2. Values to be transformed must be entered in ascending order. If this rule is violated, the console message "ENTER TRANSFORMATION ..." is repeated, and the transformation must be entered again from the beginning. This feature can be used to correct mistakes. For example, if instead of the transformation considered above, the user mistakenly types

   ```
   0   1
   1   2
   ```

   then this can be put to rights by typing the couplet

   ```
   0   0
   ```

   which is not in ascending order, and which therefore nullifies the transformation that has been fed in so far.
3. Not more than 46 couplets can be entered to define the transformation. If more are entered, the program will proceed to the next stage regardless."

Table 3.2. Variants of PCA (also available in RDA if Q27 = 1 or 3). See also Noy-Meir et al. (1975) and Prentice (1980). Irrespective of the scaling (Q15) the species and sample points form a biplot which displays approximate abundance values-after-transformation by "innerproducts" (see Jongman et al, 1987) References are given by letters between brackets.

	ANSWERS TO			
	Q26 (samples)	Q27 (species)	Q15 (scaling)	Interpretation of ordination diagram by distances [points] and arrows [inner products or angles]
Ordinary PCA	0	1	1	Euclidean distance between samples [points] (a,c)
			2	covariances between species [arrows] (b)
Standardized PCA	0	3	1	standardized Euclidean distance between samples [points] (c)
			2	correlations between species [arrows] (b)
Double centred PCA	1	1	3	after ln-transformation: appropriate for percentage data (e; see section 7.1) and can fit a unimodal model (d)
PCA standardized by sample norm	2	0	1	cosine theta similarity between samples [arrows] (a, c) = angular separation (f)
PCA standardized by sample norm and centred by species	2	1	1	cosine theta distance (c) between samples [points]
PCA centred and standardized by samples	3	0	1	"correlation coefficient" between samples [arrows] (c, f); controversial!
Noncentred PCA	0	0	1;3	(g, h)
Principal coordinates analysis	1	1	3	dissimilarity between sites when input is -(squared dissimilarity) between samples (section 7.4)

References:
(a) Jongman et al. (1987); (b) Corsten and Gabriel (1976); (c) Prentice (1980); (d) Kooijman (1977); (e) Aitchison (1983); (f) Gordon (1981); (g) Noy-Meir (1973); (h) Ter Braak (1983).

If the minimum abundance value is ≥ 1, the transformation

0	0
1	1
-1	0

transforms abundance to presence/absence.

Q25. TYPE VALUE OF C FOR USE IN LN(Y+C)-TRANSFORMATION
Range of valid answers: xxx [1.0] 999.9

---

xxx = 0.0 in linear methods
xxx = 1.0 in weighted averaging methods

The value of c must be chosen so that y+c (where y is a species value) is strictly positive, otherwise an arithmetic FORTRAN error will occur later on. For weighted averaging methods, c is required to be greater than or equal to 1 because there is likely to be at least one absence (y=0) in the data and ln(y+c) is not allowed to be negative in weighted averaging methods.

Q26. *** CENTRING/STANDARDIZATION BY SAMPLES (IN THE SPECIES DATA) ***
0 = NONE (STANDARD) []
1 = CENTRING (FINE FOR LOGPERCENTAGE DATA)
2 = STANDARDIZATION BY SAMPLE NORM
3 = BOTH 1 AND 2
Range of valid answers: [0] 3

This question is posed for linear methods and defines, in conjunction with the next question which variant of PCA/RDA is chosen (Table 3.2). The default is that neither centring nor standardization by samples is applied. Let y_{ik} be the current value of species k in sample i (k = 1, ..., m; i = 1, ..., n) and let w_k be the weight of species k. Unless requested otherwise in Q28 and Q29: $w_k = 1$. By answering 1, 2 or 3, the value y_{ik} is replaced by the value of

$$\tilde{y}_{ik} = y_{ik} - \sum_{k=1}^{m} w_k y_{ik} / \sum_{k=1}^{m} w_k \qquad \text{(answer = 1)},$$

$$y_{ik} / \left(\sum_{k=1}^{m} w_k y_{ik}^2 \right)^{\frac{1}{2}} \qquad \text{(answer = 2)},$$

$$(\tilde{y}_{ik}) / \left(\sum_{k=1}^{m} w_k \tilde{y}_{ik}^2 \right)^{\frac{1}{2}} \qquad \text{(answer = 3)}.$$

where $\tilde{y}_{ik}$ is the value obtained after centring by samples (see this question, answer = 1).

```
Q27. *** CENTRING/STANDARDIZATION BY SPECIES ***
        0 = NONE          (NON-CENTRED PCA)
        1 = CENTRING      (PCA/RDA ON A COVARIANCE MATRIX)  []
        2 = STANDARDIZATION BY SPECIES NORM
        3 = BOTH 1 AND 2 (PCA/RDA ON A CORRELATION MATRIX)
    Range of valid answers:   0   [1]   3
```

This question is posed for linear methods and defines, in conjunction with previous question, which variant of PCA/RDA is chosen. The default is centring by species only. Let y_{ik} be the value obtained after the centring/standardization by samples and let w_i be the weight of sample i. Unless requested otherwise in Q30 and Q31: $w_i = 1$. By answering 1, 2 or 3 the value of y_{ik} is replaced by the value of

$$\tilde{y}_{ik} = y_{ik} - \sum_{i=1}^{n} w_i y_{ik} / \sum_{i=1}^{n} w_i \qquad \text{(answer = 1)},$$

$$y_{ik} / \left(\sum_{i=1}^{n} w_i y_{ik}^2 \right)^{\frac{1}{2}} \qquad \text{(answer = 2)},$$

$$(\tilde{y}_{ik}) / \left(\sum_{i=1}^{n} w_i \tilde{y}_{ik}^2 \right)^{\frac{1}{2}} \qquad \text{(answer = 3)}.$$

where $\tilde{y}_{ik}$ is the value obtained after centring by species (see this question, answer = 1).

Some of the variants of PCA that can be obtained by answering Q26 and Q27 are listed in Table 3.2. Note that in RDA centring by species is implicit because of the intercept in the regression of the sample scores on the environmental variables.

After centring and standardization by samples and species, CANOCO calculates the Total Sum of Squares of the species data by

$$TSS = \sum_{i} \sum_{k} w_i w_k y_{ik}^2 .$$

Subsequently, all species values are divided by the squareroot of TSS. After division, the total sum of squares of the species data is equal to 1. This has the advantage that the eigenvalues issued by PCA and RDA are fractions of the total sum of squares and that the sum of all eigenvalues in a PCA is equal to 1, except when centring by samples is used in conjunction with standardization by species*.

* Footnote: Use of this exceptional case is discouraged; the sample means are equal to 0 after Q26 but non zero after Q27; the iterative ordination algorithm will nevertheless calculate an analysis centred by samples. The eigenvalues are fractions of TSS as defined in Q27 but do not add to 1.

When multiplied by 100, these fractions are usually refered to as percentages of variance accounted for by the ordination axes. Note that in a partial PCA and in a RDA the sum of all eigenvalues is less than or equal to 1.

```
Q28. TYPE NON-NEGATIVE WEIGHT TO BE GIVEN TO
     *SAMPLES* THAT YOU WILL BE ASKED TO SPECIFY NEXT, OR
     TYPE 0.01 TO GIVE ITEM NEGLIGIBLE WEIGHT
     TYPE 0    TO DELETE ITEM
     Range of valid answers:  0.0  [1.0]  100.0
```

Weights (w) can be assigned to samples in order to give particular samples more ($w > 1$) or less ($w < 1$) emphasis in the analysis, to delete particular samples ($w = 0$) or to make particular samples passive ($w = 0.01$). A "passive" sample has no influence on the extraction of ordination axes, but is added to the ordination afterwards by use of the transition formulae (see section 4.6; see also Jongman et al., 1987; exercises 5.2 and 5.3). Passive samples will appear after the other samples in the output. For $w > 0.01$, the weight of a sample can be interpreted as "the number of times" the sample is included in the analysis. For example, if $w = 2$ for a sample, the same ordination could also have been obtained from an unweighted analysis by including that particular sample twice in the data file(s). This interpretation is, of course, strictly valid only for integer weights ($w = 1, 2, 3, \ldots$), but the mathematics works through equally well for any positive weight (see Section 4).

```
Q29. ENTER NUMBERS (NOT NAMES) OF ITEMS TO BE WEIGHTED ONE PER
     LINE. ENDING LIST WITH A -1.
     OTHER NEGATIVE NUMBERS DENOTE SEQUENCES. FOR EXAMPLE A 4
     FOLLOWED BY A -8 WEIGHTS ITEMS 4 THROUGH 8.
     Range of valid answers:  -n  [-1]  n
```

---

n = highest sample number

This question is posed if Q28 is not answered by a 1. For example, to give sample number 3 and the sample numbers 11, 12, 13, ..., 20, 21 double weight, Q28 should be answered by typing a 2 and giving a RETURN; then Q29 appears and should be answered by typing

```
  3
 11
-21
 -1  (or merely press RETURN)
```

i.e. one number per line. Next, Q28 appears again in order to allow the user to give a different weight to other samples. If a sample is given weight more than once, the last given weight is decisive.

```
Q30. TYPE NON-NEGATIVE WEIGHT TO BE GIVEN TO
     *SPECIES* THAT YOU WILL BE ASKED TO SPECIFY NEXT, OR
     TYPE 1 FOR DEFAULT WEIGHTS (=1) FOR THE (OTHER) ITEMS
     TYPE 0.01 TO GIVE ITEM NEGLIGIBLE WEIGHT
     TYPE 0    TO DELETE ITEM
     Range of valid answers:  0.0  [1.0]  100.0
```

This question is similar to Q28, but with species replacing samples. Passive species are not placed after the other species in the output, but can still be recognized as such by the weight per species shown in the list of species scores displayed at the terminal or on the machine readable copy.

```
Q31. ENTER NUMBER (NOT NAMES) OF ITEMS TO BE WEIGHTED
     ONE PER LINE, ENDING LIST WITH A -1
     OTHER NEGATIVE NUMBERS DENOTE SEQUENCES. FOR EXAMPLE
     A 4 FOLLOWED BY A -8 WEIGHTS ITEMS 4 THROUGH 8
     Range of valid answers:  -m  [-1]  m
```

m = highest species number

This question is the same as Q29, but now applies to species.
How Q26 and Q27 interact with Q28 - Q31, can be deduced from the following example. Suppose Q27 = 3, i.e. the values of each species are standardized to mean 0 and variance 1, so that they have equal weight (in a particular sense). If a species is now given double weight in Q30 and Q31, the weighted analysis gives the same results as an unweighted analysis in which that species is included twice in the data and the same standardization (Q27 = 3) is in force.

```
Q32. IS DOWNWEIGHTING OF RARE SPECIES REQUIRED?
     TYPE 1 IF YES, TYPE Ø IF NO
     Range of valid answers:  [0]  1
```

This question is posed only for weighted averaging methods and is familiar to users of DECORANA (Hill, 1979). Hill (1979) writes: "In some applications individual samples with rare species may distort the analysis. If it is desired to give rare species less weight, while still retaining them in the analysis, then the downweighting parameter can be set to 1. Let AMAX be the frequency of the commonest species. Then the effect of downweighting is to reduce the abundance of species rarer than (AMAX/5) in proportion to their frequency. Species commoner than (AMAX/5) are not downweighted at all." For further details see Hill (1979). Downweights have a similar interpretation to the weights in Q30 - Q31. Their joint effect is

multiplicative. The downweight times the weight given in Q30 - Q31 is listed for all species numbers on the output file in a format of 20 species per line.

Note that rare species can distort the analysis only if they appear in samples with few other, more common species. These are, by definition, deviant samples. The same effect can therefore often be achieved more elegantly by deleting these deviant samples, or by making them passive.

3.7 Questions to specify the output

```
Q33. **** OUTPUT OPTION FOR ****
     CORRELATION MATRIX OF EIGENVECTORS AND ENVIRONMENTAL VARIABLES
     TYPE Ø      FOR NO OUTPUT
          1 (4)      OUTPUT ON FILE .... (name entered at Q1)
     OR THE VALUE BETWEEN BRACKETS FOR OUTPUT TO THE SCREEN AS WELL
     Range of valid answers:   0   [1]   4
```

This question, posed only if there are environmental variables in the analysis, allows the user to see the correlation matrix, means and standard deviations of eigenvectors (ordination axes) and environmental variables at the terminal and to write them to the output file specified in Q1.

If there is not enough data space available to calculate the (full) correlation matrix, a warning is given; some ordination results cannot be computed either in this case.

```
Q34. **** OUTPUT OPTION FOR ****
      ORDINATION RESULTS
     TYPE Ø      FOR NO OUTPUT
          1 (4)      OUTPUT ON FILE .... (name entered at Q1)
          2 (5)      OUTPUT ON FILE .... (name entered at Q17)
          3 (6)      OUTPUT ON BOTH
     OR THE VALUE BETWEEN BRACKETS FOR OUTPUT TO THE SCREEN AS WELL
     ENTER YOUR CHOICE FOR EACH OF THE FOLLOWING ITEMS; ON A SINGLE LINE
     or merely press RETURN for default [1] for all values.
     missing values are replaced by 0
     So one 0 is sufficient for no output
     SPEC-SCOR SAMP-SCOR REGR-COEF T-VALUES INTER-COR ENVI-BIPL CENTROIDS LINEA-COM
```

This question allows the user to see the ordination results of the analysis at the terminal and to write them to the output file and/or to the files for the machine readable copy (see Q17 and Q18). The answers for the items must be entered on a single line, for example,

```
  6  2  1  4  4  6  6  5
```

ending by pressing the return-key. If more numbers are entered on a line than required, the superfluous numbers are ignored. If a machine readable

copy is not asked for in Q16, the lines beginning with 2(5) and 3(6) do not appear; if the answer given is nevertheless 2, or 3 (or 5 or 6), then CANOCO acts as if the answer 1 (or 4) was given. If an error is detected in the answer, the question appears again.

The items listed in Q34 depend on the type of analysis. The abbreviations are (between brackets the symbol used in section 4).

SPEC-SCOR = species scores (u_k)
SAMP-SCOR = sample scores (x_i^*)
REGR-COEF = regression coefficients (c_j) of the environmental variables for an unconstrained ordination axis; canonical coefficients (c_j) for a constrained ordination axis
T-VALUES = t-values associated with the regression coefficients c_j in the multiple regression of x_i^* on x_i
INTER-COR = inter-set correlations between the environmental variables and the ordination scores x_i^*
ENVI-BIPL = scores of environmental variables for drawing a biplot (suitable for quantitative variables)
CENTROIDS = centroids of environmental variables in the ordination diagram (suitable for qualitative (nominal) variables)
LINEA-COM = sample scores which are linear combinations of the environmental variables (x_i)

See section 4 for further explanation.

3.8 Questions to specify additional analyses

The program automatically stops at this point if there is not enough data space available. Otherwise the program asks how to continue.

```
Q35. TYPE
     0 = STOP
     1 = MORE    ANALYSIS WITH CURRENT DATA
     2 = PASSIVE ANALYSIS OF OTHER ENVIRONMENTAL VARIABLES
     3 = AS 2, BUT WITH REGRESSIONS
     Range of valid answers:   [0]   3
```

The user can stop the program by answering 0 or ask for additional analyses using the current species data and covariables. In additional analyses the answers concerning data transformation (Q24 - Q32) and covariables (Q22 - Q23) and the output file (Q1) remain in force and are not posed again.

After answering 1, the user can delete environmental variables (Q36), ask for a statistical test (Q37 - Q40) or modify the type of analysis (Q2, Q8 - Q18). However, the user cannot switch between linear methods and weighted averaging methods and cannot delete samples. If there are no environmental variables, the program continues immediately with Q2, else with Q36.

After answering 2 or 3, CANOCO asks for a (new) set of environmental variables which are used to interpret the current ordination axes and on request later on to extract ordination axes which are linear combinations of them (i.e. to extract new canonical axes). By answering 3 the current sample scores (x_i) which are linear combinations of the previous set of environmental variables, are replaced by fitted values of the regression of the current sample scores (x_i^*) on the newly entered environmental variables. If the answer is a 2, then the current sample scores x_i and x_i^* remain unchanged. In either case, CANOCO will ask for a file name with environmental data (cf. Q5), which environmental variables are to be deleted (Q20), whether interactions are to be included (Q21), which output is required (Q33 and Q34) and how to continue (Q35).

Q36. TYPE 1 IF YOU ONLY WISH TO DELETE ENVIRONMENTAL VARIABLES,
ELSE TYPE Ø
Range of valid answers: [0] 1

If Q36 answered in the affirmative way, CANOCO asks which numbers are to be deleted (Q20), calculates a new ordination according to the current type of analysis using the remaining environmental variables, asks what output is required (Q33 - Q34) and how to continue (Q35). If Q36 is answered negatively, the next question appears in direct or hybrid gradient analyses and Q2 in indirect gradient analyses.

Q37. *** MONTE CARLO PERMUTATION TEST ***
Ø = NO SIGNIFICANCE TEST
1 = TEST OF SIGNIFICANCE OF FIRST CANONICAL AXIS
2 = OVERALL TEST USING THE TRACE STATISTIC
3 = BOTH 1 AND 2
Range of valid answers: [0] 3

This question is posed in direct or hybrid gradient analyses. If the answer is 0, then the program continues with Q2. The other answers lead to Monte Carlo significance tests. The tests are carried out by randomly permuting the sample numbers in the environmental data: the environment data are randomly linked to the species data, giving rise to a "random data set". For each random data set, CANOCO calculates one or two test statistics, namely the first eigenvalue and/or the sum of all eigenvalues (= the trace). (The number of permutations is to be specified in the next question.) If the species react to the current environmental variables, then the test statistic calculated from the data-as-observed will be larger

than most of test statistics calculated from the random data. If the observed value is among the 5% highest values, then the species are significantly related to the environmental variables. Taking the trace as a test statistic gives an overall test of the effect of the environmental variables on the species. Canonical ordination is particular effective if the species-environment relationships can be displayed in one or two dimensions (an ordination diagram). A potentially more powerful test is therefore obtained by using the first eigenvalue rather than the trace as test statistic. In particular, this test shows whether the first canonical axis is significant. To test the significance of the second canonical axis, a separate analysis is needed in which the first ordination axis is entered as a covariable.

The overall test (answer 2) costs less computing time than the test of significance of the first canonical axis (answer 1).

The effect of a particular set of environmental variables can be tested, while taking into account the effect of other variables, by specifying the latter as covariables.

If there is only one environmental variable, then the answers 1, 2 and 3 are equivalent, because the trace is then equal to the first eigenvalue. To save computing time, CANOCO will act in this case as if answer 2 had been given.

Q38. TYPE NUMBER OF RANDOM PERMUTATIONS
Range of valid answers: 1 [99] 999

For a test at the 5%-significance level, minimally 19 permutations are required (the result is then significant if the test statistic for the data is larger than that for any of the 19 permutations, because 1/20 = 0.05). The power of the test can be increased by increasing the number of permutations (Hope, 1968). As each extra permutation costs computer time, taking a number larger than 99 will not usually be worthwhile, whereas 99 permutations is the minimum number of permutation which may result in a significance level of 0.01, because 1/100 = 0.01.

Q39. TYPE TWO INTEGERS (1-30000) AS SEEDS FOR THE RANDOM SEQUENCE,
ON A SINGLE LINE OR PRESS RETURN FOR DEFAULT SEEDS.
Range of valid answers: 1 [23239] 30000, and 1 [945] 30000

A Monte Carlo test needs pseudo-random numbers as input. To start a sequence of pseudo-random numbers, seeds are required. The default seeds are 23239 and 945. To get a different sequence of pseudo-random numbers one needs to specify other values. If more than one test is applied to the same data, the user is advised to specify new seeds for each test.

Q40. IF YOU WANT RESTRICTED PERMUTATION, TYPE THE NUMBER OF COVARIABLES ON TO WHICH THE PERMUTATION MUST BE CONDITIONED, ELSE TYPE Ø
Range of valid answers: [0] p

p = highest number of covariable

This question is posed only if there are covariables in the analysis. An example is given in section 4.11. Restricted permutation is appropriate for analysing data from randomized block experiments (see Cochran and Cox, 1957; Cox, 1958). In such experiments treatments are assigned randomly to samples within each block (= randomization of treatments among samples within each block), instead of being completely randomized among all samples. To test for treatment effects by a permutation test, randomization must be restricted similarly to samples within blocks. In field research (observational research) restricted permutation is called for if samples are taken in a number of different locations and interest focusses on the common variation within the locations. The locations then act as "blocks" in the sense used above.

Data from such experiments can be analysed in CANOCO by taking as covariables a series of dummy variables representing blocks (one for each block) and as environmental variables the treatment variables (commonly also a series of dummy variables). If there are k blocks, k-1 dummy variables are required (delete therefore one of the dummy variables for blocks - which one is completely arbitrary) and answer to the present question with k-1 (see Table 4.13).

CANOCO can only condition the permutation on the first k covariables. If there are more covariables in the analysis, those representing blocks should therefore come first. If the order of the variables in a full format data file happens to violate this requirement of order, the order in which the variables are being read by CANOCO can sometimes be changed by using the so-called T-descriptor in the FORTRAN-format (see any text on FORTRAN, e.g. Metcalf, 1985). If the numbering of variables in a datafile in Cornell condensed format violates this requirement of order, the variables must be renumbered so that the covariables on to which permutation must be restricted receive the lowest numbers.

The user is advised not to condition on a quantitative variable which takes many different values. The test result will then be non-significant anyway.

3.9 Example

As an example of a terminal dialogue, the dune meadow data of Table 2.1 are analysed by canonical correspondence analysis (CCA). Table 3.3 shows the numbers (not the text) of the questions that appear at the terminal and the answers given (the part before the "="-sign, the rest of each line is annotation). Part of the terminal dialogue is shown in Table 3.4. CANOCO starts with reporting how much space for data is available, i.e. how many samples, species, etc., can be analysed. The dimensions

Table 3.3 Annotated copy, as produced by CANOCO (see Table 3.1), of the answers entered at the terminal to obtain a canonical correspondence analysis of the dune meadow vegetation data (Table 2.2) using the environmental data in Table 2.4 as explanatory variables. Sample 20 and species 31, 32 and 33 are made passive. The question numbers are added here for convenience of reference.

```
QUESTION - INPUT AT TERMINAL                          = ANNOTATION

Q1 :     DUNE.OUT                                     = OUTPUT FILE
Q2 :        5  = ANALYSIS NUMBER
Q3 :     DUNEMEAD.SPE                                 = FILE WITH SPECIES DATA
Q5 :     DUNEFUL.ENV                                  = FILE WITH ENVIRONMENTAL DATA
Q6 :       0 = COVARIABLES?
Q14:       1 = SCALING OF SAMPLE AND SPECIES SCORES?
Q16:       0 = MACHINE READABLE COPY OF SOLUTION?
Q19:          0 = SAMPLE NUMBER TO BE OMITTED
Q20:          4 ENVIRONMENTAL VARIABLE TO BE OMITTED
              7 ENVIRONMENTAL VARIABLE TO BE OMITTED
              0 ENVIRONMENTAL VARIABLE TO BE OMITTED
Q21:       -1    0  = PRODUCT OF ENVIRONMENTAL VARIABLES
Q24:        -1.00    0.00 = TRANSFORMATION OF SPECIES DATA
Q28:        0.01000 = WEIGHT ( NOWEIGHT=-1)
Q29:         20  = ITEM GIVEN NONSTANDARD WEIGHT
             -1  = ITEM GIVEN NONSTANDARD WEIGHT
Q28:        1.00000 = WEIGHT ( NOWEIGHT= 1)
Q30:        0.01000 = WEIGHT ( NOWEIGHT= 1)
Q31:         31  = ITEM GIVEN NONSTANDARD WEIGHT
             32  = ITEM GIVEN NONSTANDARD WEIGHT
             33  = ITEM GIVEN NONSTANDARD WEIGHT
             -1  = ITEM GIVEN NONSTANDARD WEIGHT
Q30:        1.00000 = WEIGHT ( NOWEIGHT= 1)
Q32:       0 = DOWNWEIGHTING OF RARE SPECIES?
Q33:       4 = OUTPUT OF CORRELATIONS?
Q34:       4  4  4  4  4  4  4  4 = ORDINATION OUTPUT
Q35:       2 = STOP, MORE ANALYSES, OTHER ENV. DATA?
Q5 :     DUNEFUL.ENV                                      = FILE OF PASSIVE VARIABLES
Q20:          0 ENVIRONMENTAL VARIABLE TO BE OMITTED
Q21:       -1    0  = PRODUCT OF ENVIRONMENTAL VARIABLES
Q33:       0 = OUTPUT OF CORRELATIONS?
Q34:       4  4  4 = ORDINATION OUTPUT
Q35:       0 = STOP, MORE ANALYSES, OTHER ENV. DATA?
```

listed in Table 3.4 can easily be increased when they are too low for a particular data set (see section 9.1). The output is written to the file with name "DUNE.OUT", the analysis number is 5, the species data are on the file "DUNEMEAD.SPE" (Table 2.2) and the environmental data on the file "DUNEFUL.ENV" (Table 2.4), etc. (Table 3.3). There are no covariables. After question 16, CANOCO starts reading the species data. It reports (see Table 3.4) title and data format of the data file, the number of samples (21), the maximum species number encountered (33) and the number of occurrences (219 = total number of couplets in the file). Of the environmental variables on file DUNEFUL.ENV (Table 2.4), the dummy variables 4 and 7 are deleted because they are redundant for specifying the nominal variables agricultural use and management regime, respectively. The species data are not transformed, except that sample 20 and species 31, 32 and 33 are made passive (Table 3.3). There remain 20 active samples, 1 passive sample and 30 active species (Table 3.4). The ordination axes are extracted subsequently. The iteration report monitors the convergence of the iterative ordination algorithm for extracting the ordination axes; the residual usually is not of great interest to the user. Of more interest is the eigenvalue, in the example 0.46121.

After the CCA ordination axes have been extracted the ordination results are displayed at the terminal (Q34). The output of this analysis is taken as example in the next section in which the output of CANOCO is fully explained. The last six lines of Table 3.3 show how to obtain the interset correlations, the biplot scores and the centroids of all environmental variables, including the deleted variables 4 and 7.

Table 3.4 Part of the terminal dialogue of the canonical correspondence analysis specified in Table 3.3

```
 **** CANOCO **** VERSION 2.1 **** MARCH 1987 ****

PROGRAM CANOCO - WRITTEN BY CAJO J.F. TER BRAAK
COPYRIGHT (C) 1987 TNO INSTITUTE OF APPLIED COMPUTER SCIENCE,
BOX 100, 6700 AC WAGENINGEN, THE NETHERLANDS.
CANOCO PERFORMS (PARTIAL) (DETRENDED) (CANONICAL) CORRESPONDENCE ANALYSIS,
PRINCIPAL COMPONENTS ANALYSIS AND REDUNDANCY ANALYSIS.
THE PROGRAM IS AN EXTENSION OF CORNELL ECOLOGY PROGRAM DECORANA (M.O. HILL, 1979)

**** MAXIMUM DIMENSIONS OF PROGRAM AS COMPILED ****
SYMBOL             MAXIMUM NUMBER OF
MMAX   =     750 - SAMPLES
NMAX   =     600 - SPECIES
NZMAX  =      68 - ENVIRONMENTAL VARIABLES
INTMAX =      30 - INTERACTION TERMS
NAMAX  =     100 - COVARIABLES
IDMAX  =   17500 - PRESENCES IN THE SPECIES DATA
NZDAT  =   12000 - ENVIRONMENTAL VALUES
NADAT  =    9000 - VALUES OF COVARIABLES

TYPE 0 FOR INPUT FROM CONSOLE
     1 FOR INPUT FROM FILE
Range of valid answers:       [0]       1
Type your answer(s) or merely press RETURN for default, indicated by []
0
ANSWERS ARE WRITTEN TO FILE CANOCO.CON


TYPE NAME OF OUTPUT FILE
DUNE.OUT

*** TYPE OF ANALYSIS ***
 MODEL              GRADIENT ANALYSIS
                INDIRECT      DIRECT      HYBRID
LINEAR           1=PCA         2= RDA        3
UNIMODAL         4= CA         5= CCA        6
   ,,            7=DCA         8=DCCA        9
               10=NON-STANDARD ANALYSIS
TYPE ANALYSIS NUMBER
Range of valid answers:        1        [5]      10
Type your answer(s) or merely press RETURN for default, indicated by []
 5
ANSWER =  5
```

Table 3.4 (continued/1)

```
TYPE NAME OF FILE WITH SPECIES DATA
DUNEMEAD.SPE

TYPE NAME OF FILE WITH ENVIRONMENTAL DATA
DUNEFUL.ENV

TYPE 1 IF YOU HAVE COVARIABLES, ELSE TYPE 0
*EXPLANATION* COVARIABLES ARE:
              VARIABLES WITH KNOWN OR UNINTERESTING EFFECTS ON THE SPECIES.
              THEIR EFFECTS ARE ELIMINATED WHEN EXTRACTING ORDINATION AXES.
Range of valid answers:      [0]      1
Type your answer(s) or merely press RETURN for default, indicated by []
0
ANSWER =  0

*** SCALING OF ORDINATION SCORES ***
1 = SAMPLE  SCORES ARE WEIGHTED MEAN SPECIES SCORES
2 = SPECIES   ,,    ,, WEIGHTED MEAN SAMPLE    ,,
3 = SYMMETRIC SCALING
Range of valid answers:      [1]      3
Type your answer(s) or merely press RETURN for default, indicated by []
1
ANSWER =  1

TYPE 1 FOR A MACHINE READABLE COPY OF THE SOLUTION
Range of valid answers:      [0]      1
Type your answer(s) or merely press RETURN for default, indicated by []
0
ANSWER =  0

FILE    : DUNEMEAD.SPE
TITLE   :  SPECIES - DUNE MEADOW DATA (M. BATTERINK AND G. WIJFFELS,
1983)
FORMAT :   (I10,X,5(I4,F5.0))

NO. OF COUPLETS OF SPECIES NUMBER AND ABUNDANCE PER LINE :    5


ENTER NUMBERS (NOT NAMES) OF SAMPLES TO BE OMITTED
ONE AT A TIME, ENDING LIST WITH A ZERO
Range of valid answers:      [0]     30
Type your answer(s) or merely press RETURN for default, indicated by []
   0
          0
NUMBER OF SAMPLES            21
NUMBER OF SPECIES            33
NUMBER OF OCCURRENCES       219
```

Table 3.4 (continued/2)

```
FILE  : DUNEFUL.ENV
TITLE :  ENVIRONMENTAL DATA IN FULL FORMAT- DUNE MEADOW DATA

NO. OF ENVIRONMENTAL VARIABLES :    10
FORMAT :
 (I5,F5.0,X,2F3.0,3X,3F2.0/18X,4F2.0)

 ..............................
 ..... Q20 - Q31 not shown .....
 ..............................

IS DOWNWEIGHTING OF RARE SPECIES REQUIRED?
TYPE 1 IF YES, TYPE 0 IF NO
Range of valid answers:      [0]      1
Type your answer(s) or merely press RETURN for default, indicated by []
0
ANSWER =  0

NO. OF ACTIVE  SAMPLES:     20
NO. OF PASSIVE SAMPLES:      1
NO. OF ACTIVE  SPECIES:     30

 ..............................................
 ..... Q33 and correlation matrix not shown .....
 ..............................................

ITERATION REPORT AXIS   1
RESIDUAL  0.073511       AT ITERATION  0
RESIDUAL  0.000315       AT ITERATION  1
RESIDUAL  0.000001       AT ITERATION  2
EIGENVALUE   0.46121

ITERATION REPORT AXIS   2
RESIDUAL  0.045239       AT ITERATION  0

 ..................
 ..... ectetera .....
 ..................
```

4. OUTPUT

4.1 Samples and species in the analysis

The first output of CANOCO is already shown in Table 3.4 and reports, among other things, how many samples and species are in the analysis. The active samples and species jointly determine the ordination. In contrast, passive samples and species do not influence the ordination (their scores on the ordination axes are calculated afterwards). Unless the user specifies otherwise, a sample is active when it occurs (1) with non-zero values for active species in the species data and (2) where relevant, in the environment data and the data for the covariables. A sample is passive when it is made so by the user in Q28 - Q29, or when the sample is not encountered either in the file with environment data or in the file with the covariables. Active species are species that have non-zero values for active samples in the species data and which are not deleted or made passive. The number of active species can be lower than the highest species number encountered in the species data, because some species numbers may be absent in the data.

4.2 Iteration report, eigenvalue and length of gradient

The iteration report (Table 3.4) monitors the convergence of the iterative ordination algorithm for extracting the ordination axes. The residual shown is the root mean square difference between the current sample scores and their predecessor trial scores. Convergence is reached if the residual is less than 0.000050. A maximum of 17 residuals is printed and if the residual is after that still not below 0.000050, a warning is given and the current trial scores are taken as the final sample scores. In general, the algorithm converges quickly, unless the eigenvalues of subsequent axes are nearly equal or very close to 0. When close eigenvalues turn out to cause the problem of non-convergence, the results can nevertheless be trusted; but discard ordination axes with very low eigenvalues (<0.02). The eigenvalues usually are in order of decreasing value, unless the first ordination axes are constrained (canonical) and subsequent axes are unconstrained. Small eigenvalues however sometimes do not appear in the correct order.

The eigenvalue is always a number between 0 and 1; the higher the value, the more important the ordination axis. In linear methods, the eigenvalue is the fraction of the Total Sum of Squares in the species data (after data transformation) extracted by the ordination axis (see Q27; cf. Jongman et al. 1987; section 5.3.2). Multiplied by 100, it is usually referred to as the "percentage variance accounted for" by the axis. In weighted averaging methods, the eigenvalue is a measure of separation of the species' distributions along the ordination axis (Jongman et al., 1987, section 5.2.2). Formally, it is the ratio of the dispersion of the species scores and that of the sample scores, if Q14 = 2, and the inverse of this ratio if Q14 = 1. Eigenvalues of ca. 0.3 and higher are quite common in ecological applications.

Table 4.1 Weighted correlations, means, standard deviations and variance inflation factors of environmental variables in the CCA of the dune meadow data. Note that the dummy variables hayfield and SF are not in the list of variables because they were deleted to avoid multicollinearity (see Q20). The variables manure and NM have the highest variance inflation factors, partly because their correlation is quite high in absolute value ($r = -0.74$).

**** WEIGHTED CORRELATION MATRIX (WEIGHT = SAMPLE TOTAL) ****

A1	1.00						
MOISTURE	0.42	1.00					
MANURE	-0.23	-0.22	1.00				
HAYPASTU	0.16	-0.17	0.48	1.00			
PASTURE	0.02	0.16	0.12	-0.48	1.00		
BF	-0.31	-0.38	-0.18	-0.05	0.03	1.00	
HF	-0.14	-0.18	0.14	-0.26	0.20	-0.30	1.00
NM	0.36	0.36	-0.74	-0.28	-0.11	-0.24	-0.36

	A1	MOISTURE	MANURE	HAYPASTU	PASTURE	BF	HF

VAR	(WEIGHTED) MEAN	STAND. DEV.	INFLATION FACTOR
A1	4.6850	1.8613	1.7814
MOISTURE	2.8015	1.7312	1.8500
MANURE	1.9022	1.3629	8.3034
HAYPASTU	0.4146	0.4927	3.0874
PASTURE	0.2467	0.4311	2.2451
BF	0.1708	0.3763	4.7139
HF	0.3109	0.4629	3.2716
NM	0.2204	0.4145	7.5705

The length of gradient (in SD-units) is reported only if non-linear rescaling of ordination axes is in force, i.e. usually only in DCA with detrending-by-segments. The non-linear rescaling proceeds by iteratively expanding and contracting segments of the ordination axis (see Q11). This process is reported on the output file (LENGTH OF SEGMENTS). The change in length (and the number of segments) is shown by printing the length per segment before and after rescaling (see Hill, 1979).

4.3 Correlation matrix, means, standard deviations and inflation factors

At the terminal two correlation matrices are shown in response to Q33, one for the correlations between environmental variables, displayed before the calculation of ordination axes (Table 4.1) and one displayed thereafter for the correlations between ordination axes (Table 4.2). A single, complete correlation matrix is written to the output file (if enough dataspace is available). For linear methods, one obtains the usual matrix of Pearson correlation coefficients. But if weights are specified for samples in Q28 - Q29, these weights are used in calculating the means, standard deviations and correlation coefficients in the obvious way (Kendall and Stuart, 1973, p. 301). In weighted averaging methods the total abundance in a sample (y_{i+}) acts as a sample weight, even if default weights are used in Q28 - Q29 ($w_i = 1$). In general, weighted means, weighted standard deviations and weighted correlation coefficients are calculated with $w_i^* = w_i \sum_k w_k y_{ik}$ acting as sample weight.

If covariables are present, partial correlations (Kendall and Stuart, 1973, Chapter 27) are displayed on the output file. Partial covariances are displayed at the terminal.

Table 4.1 also shows a column head "INFLATION FACTOR". It is the Variance Inflation Factor (VIF) of a variable in a multiple regression equation (Montgomery and Peck, 1982: section 8.4.2). The name derives from the fact that the variances of estimated regression coefficients $\{c_j\}$ are proportional to their VIF's, namely

$$\text{var}(c_j) = \text{VIF} \times (\text{residual variance})/(n-q-1) \qquad (4.1)$$

where n is the number of samples and q the number of environmental variables in the equation. The VIF is related to the (partial) multiple correlation R_j between environmental variable j and the other environmental variables in the analysis:

$$\text{VIF} = \frac{1}{1 - R_j^2} . \qquad (4.2)$$

If the VIF of a variable is large, say VIF > 20, then the variable is almost perfectly correlated with the other variables and therefore has no unique contribution to the regression equation. As a consequence, its regression coefficient (or its canonical coefficient in canonical ordination) is unstable and does not merit interpretation (Ter Braak, 1986a).

Table 4.2 Weighted correlations between ordination axes, percentages variance accounted for by the species-environment biplot and the trace (= sum of all canonical eigenvalues). The species-environment correlations are 0.96, 0.90, 0.86 and 0.89, respectively.
(SPEC AX = species axis $\{x_i^*\}$; ENVI AX = environmental axis $\{x_i\}$).

**** WEIGHTED CORRELATION MATRIX (WEIGHT = SAMPLE TOTAL) ****

SPEC AX1	1.00						
SPEC AX2	-0.04	1.00					
SPEC AX3	0.08	-0.04	1.00				
SPEC AX4	-0.05	0.13	-0.11	1.00			
ENVI AX1	0.96	0.00	0.00	0.00	1.00		
ENVI AX2	0.00	0.90	0.00	0.00	0.00	1.00	
ENVI AX3	0.00	0.00	0.86	0.00	0.00	0.00	1.00
ENVI AX4	0.00	0.00	0.00	0.89	0.00	0.00	0.00

	SPEC AX1	SPEC AX2	SPEC AX3	SPEC AX4	ENVI AX1	ENVI AX2	ENVI AX3

PERCENTAGE VARIANCE ACCOUNTED FOR BY FIRST S AXES OF SPECIES-ENVIRONMENT BIPLOT

S	PERC
1	37.8
2	62.3
3	75.4
4	86.3

SUM OF ALL CANONICAL EIGENVALUES: TRACE = 1.21967

High VIF's indicate multicollinearity among the environmental variables. If an environmental variable is completely multicollinear, it will automatically be removed. VIF's are always greater than 1.0. For mutually uncorrelated environmental variables all VIF's are equal to 1.0, but this happens only in designed experiments. If all VIF's are given as 1.0000, then CANOCO probably did not calculate them at all.

Table 4.2 shows the correlation matrix of the ordination axes. "SPEC AXk", abbreviation of "species axis number k", stands for the sample scores $\{x_i^*\}$ on k-th ordination axis which are derived from the species scores by weighted averaging or weighted summation (section 4.6) whereas "ENVI AXk",abbreviation of "environmental axis number k", stands for the sample scores $\{x_i\}$ on the k-th ordination axis which are linear combinations of the environmental variables (section 4.7). The correlation between SPEC AXk and ENVI AXk is the species- environment correlation (Ter Braak, 1986a).

In an indirect gradient analysis the species axes are mutually uncorrelated, whereas in a direct gradient analysis the environmental axes are mutually uncorrelated (unless detrending-by segments is in force). In a hybrid analysis the environmental axes of the constrained (canonical) axes have correlation 0 to the species axes of the later unconstrained axes.

If a diagonal element of the correlation matrix is equal to 0.0000, then the corresponding axis was not calculated or had negligible variance.

The means and standard deviation of the ordination axes are reported below the full correlation matrix on the output file. The means usually are equal to 0, except when nonlinear rescaling of axes is in force or, in linear methods, when the data are not centred by species (Q27 = 0 or 2). Each standard deviation is a simple function of the eigenvalue and the species-environment correlation (see sections 4.6 and 4.7). The standard deviation of the environmental axis is always R × (standard deviation of species axis) where R is the species-environment correlation of the corresponding axis.

4.4 Percentage variance accounted for by first s axes of species-environment biplot

Table 4.2 also shows at the bottom the percentage variance accounted for by the first s axes of the species-environment biplot. The species-environment biplot is an ordination diagram in which the environmental variables are represented by arrows (Figs. 4.1 and 4.2). The arrow roughly points in the direction of maximum variation in value of the corresponding variable. The scores from which Fig. 4.2 is prepared are given in the Tables 4.4, 4.5, 4.10 and 4.11.

In linear methods, the species are often also represented by arrows (Fig. 4.1). By looking at the angles between arrows one may get an idea of the correlations between a species' abundance and the environmental variables (Jongman et al., 1987, section 5.5.3; Ter Braak and Prentice, 1987). The plot of arrows of species and environmental variables also allows a qualitative

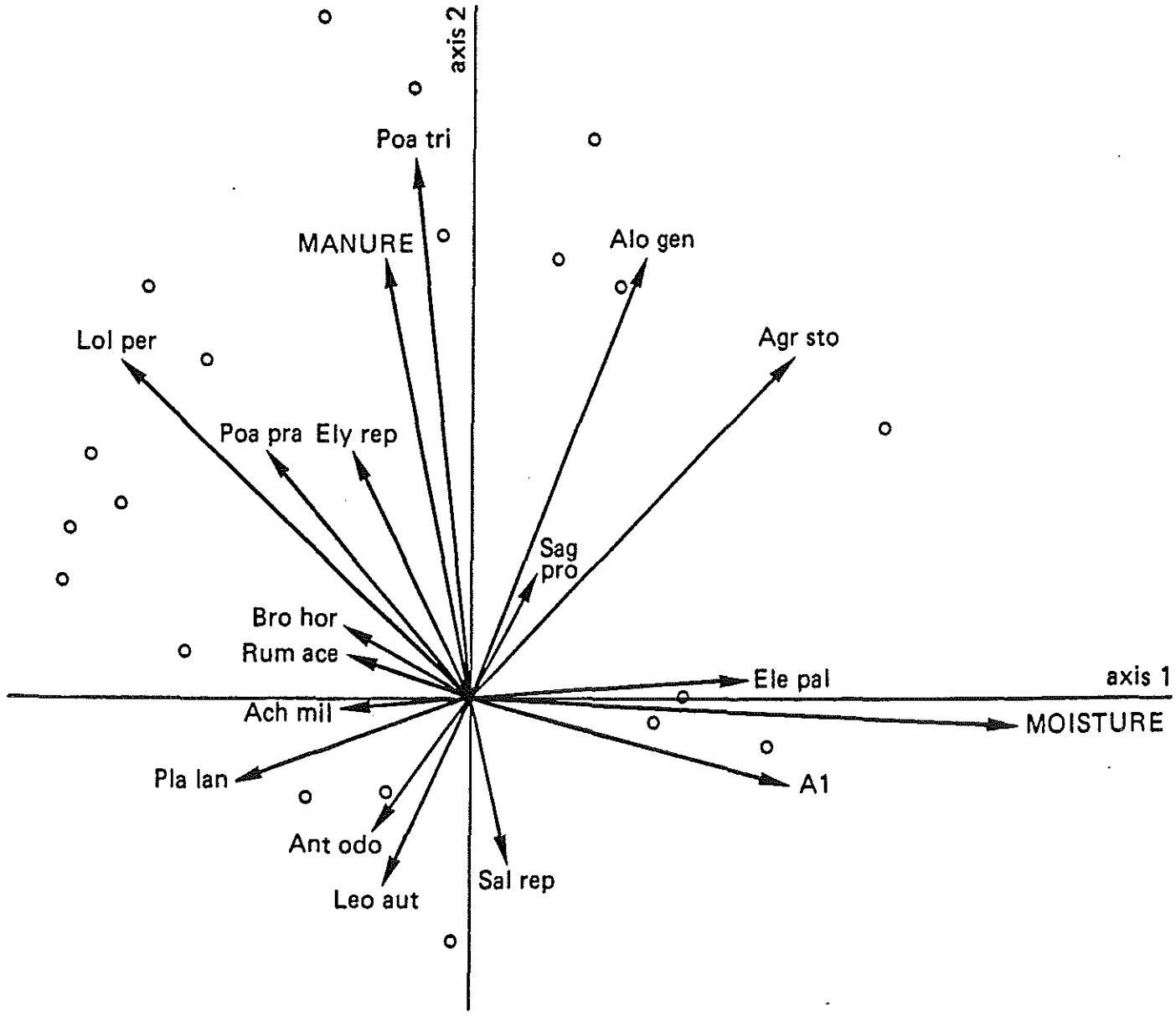

Fig.4.1 Biplot based on redundancy analysis of the dune meadow vegetation data with respect to three environmental variables (arrows: manure, moisture and thickness of the A1 horizon) in dune meadows (o) on the island of Terschelling, The Netherlands. The arrows for plant species and environmental variables display the approximate (linear) correlation coefficients between plant species and the environmental variables. Abbreviations are given as underlying in Table 2.1 (λ_1 = 0.23, λ_2 = 0.15. scaling α = 1).

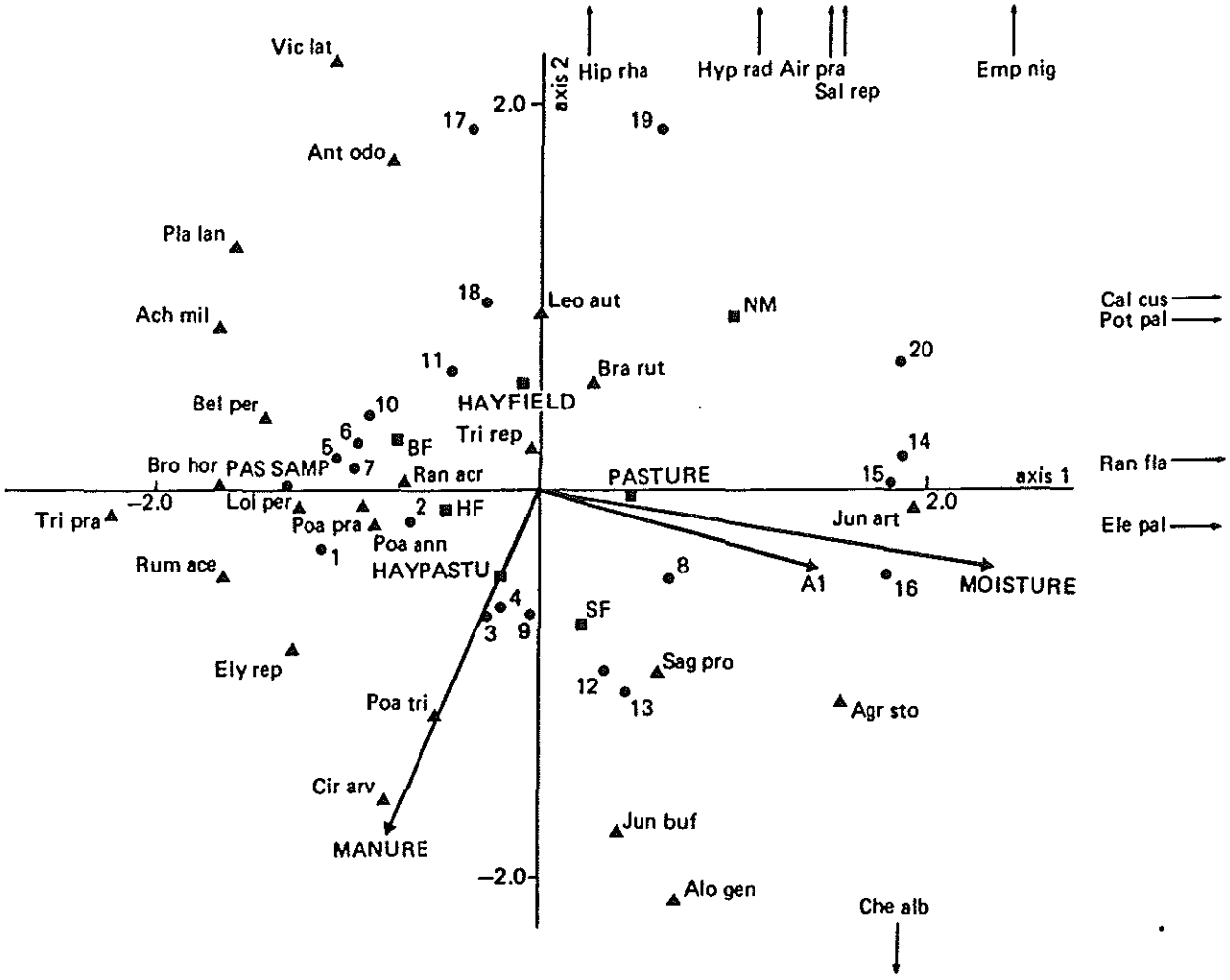

Fig.4.2 Ordination diagram based on canonical correspondence analysis of the dune meadow vegetation data with respect to three quantitative environmental variables (arrows: manure, moisture and A1) and two nominal variables (agricultural use and management regime) of which the classes are shown by their centroids (■) in the diagram. The coordinates of species (▲), sample (●) and class (■) points are given in Tables 4.4, 4.5 and 4.11 respectively. The heads of arrows are at the coordinates given by Table 4.10 (after division by 2.0). Abbreviations of species and environmental variables are given in Tables 2.2 and 2.3, respectively, (λ_1 = 0.46, λ_2 = 0.30, scaling α = 1).

Table 4.3 Species scores as displayed at the terminal. The scores have been multiplied by the multiplier (100) shown at the bottom (N = species number; AX = ordination axis; EIG = eigenvalue; WEIGHT is explained in the text). Species 31, 32 and 33 have weight 0 because they were made passive. The eigenvalues show that the axes 3 and 4 are of minor importance compared to the first two axes.

SPECIES SCORES

N	NAME	AX1	AX2	AX3	AX4		WEIGHT
	EIG	0.46	0.30	0.16	0.13		
1	ACH MIL	-169	83	8	-98	!	16
2	AGR STO	155	-109	-31	-24	!	48
3	AIR PRA	148	391	-294	156	!	5
4	ALO GEN	71	-212	-95	41	!	36
5	ANT ODO	-77	170	-11	66	!	21
6	BEL PER	-143	37	-70	-211	!	13
7	BRO HOR	-167	4	-44	-252	!	15
8	CHE ALB	187	-360	-189	8	!	1
	.....						
	..... species 9 - 26 not shown						
	.....						
27	TRI REP	-5	24	82	-37	!	47
28	VIC LAT	-108	222	87	-200	!	4
29	BRA RUT	27	55	38	66	!	49
30	CAL CUS	332	99	105	-75	!	10
31	HIP RHA	21	319	-111	242	!	0
32	POA ANN	-85	-17	-128	-110	!	0
33	RAN ACR	-71	4	275	207	!	0
	MULTIPLIER	100					

interpretation by the rules of a biplot (Jongman et al., 1987, section 5.3.4 and 5.9.3-4; Gabriel, 1981), leading to approximate values of the covariances between species and environmental variables. The word "approximate" means that the biplot does not display the exact values and, therefore, does not "explain" the total (weighted) variance in all these covariances, but just a fraction thereof. What fraction is given here as a percentage. In RDA, the fraction simply is $(\lambda_1 + \ldots + \lambda_s)$/(sum of all canonical eigenvalues), λ being the eigenvalue of a canonical axis and s being the number of axes of the biplot.

In weighted averaging methods, the species are represented by points in the ordination diagram (Fig. 4.2). The joint plot of species points and environmental arrows forms also a biplot (Ter Braak, 1986a; Jongman et al., 1987, section 5.5.2). This biplot leads to approximate values of the (centred) weighted averages of the species with respect to the (standardized) environmental variables. The word "approximate" again means that the biplot does not display the exact values and, therefore, does not "explain" the total (weighted) variance in all these weighted averages, but just a fraction thereof. What fraction is given as a percentage (see appendix in Ter Braak, 1986a). In CCA, the fraction simply is $(\lambda_1 + \ldots + \lambda_s)$/(sum of all canonical eigenvalues).

The bottom line in Table 4.2 shows the sum of all canonical eigenvalues (without detrending), also termed the trace (Appendix C). In linear methods, the trace is always less than or equal to 1.

In interpreting the percentages of variance accounted for, it must be kept in mind that the goal is not 100%, because part of the total variance is due to noise in the data. Even an ordination that explains only a low percentage may be quite informative. Moreover, the percentage is dependent on the number of variables in the analysis. For example, with only two environmental variables in the analysis, two canonical axes always explain 100%, regardless of whether the result is ecologically meaningful. The numerical importance of an axis is better judged by looking at its eigenvalue or its standard deviation, and its statistical validity by a significance test (Q37).

4.5 Species scores

The species scores are displayed on the terminal in the form shown in Table 4.3. Before display the scores have been multiplied by a multiplier that is chosen in such a way that the scores as displayed lie between -999 and 999. The multiplier is shown at the terminal below the scores. In the example of Table 4.3 the multiplier is 100. The original scores that result from the ordination algorithm are in the example thus a factor 100 lower, e.g. for species 1 (ACH MIL) the original scores are -1.69, 0.83, 0.08 and -0.98. Always note the multiplier when interpreting scores. The column "WEIGHT" gives in linear methods the weights $\{w_k\}$ given to species in Q30 - Q31 and in weighted averaging methods $w_k^* = w_k \sum_i w_i y_{ik}$, the weighted total abundance of a species (where w_i denotes sample weight). Passive species are recognizable by weight 0 (species 31, 32 and 33 in Table 4.3).

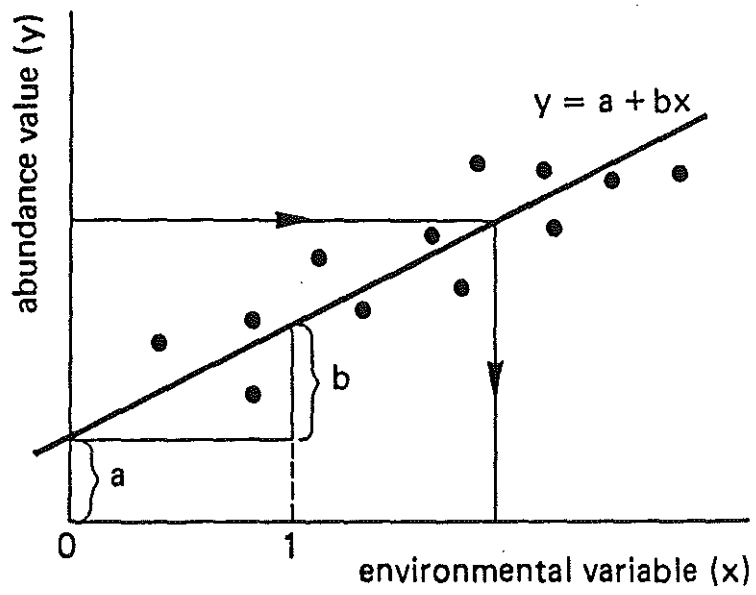

Fig.4.3 A straight line displays the linear relation between the abundance value (y) of a species and an environmental variable (x), fitted to artifical data (o). (a = intercept; b = slope or regression coefficient).

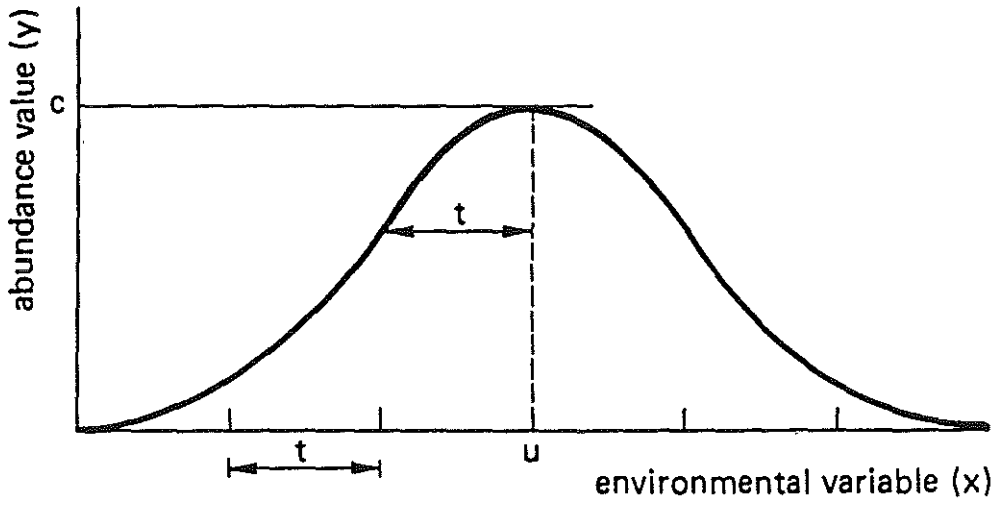

Fig.4.4 A unimodal relation between the abundance value (y) of a species and an environmental variable (x). (u = optimum or mode; t = tolerance; c = maximum.)

They happen in the example to be at the bottom of the table but this is not the general rule. The weights are useful in interpreting ordination diagrams: in weighted averaging methods species at the edge of the diagram often carry low weights; such peripheral species have little influence on the analysis and it is often convenient not to display them at all.

Table 4.4 shows the format in which ordination results are written on the output file. The multiplier used is at the righthand top of the table. The format on the machine readable copy is similar to that shown in Table 4.3.

If a column contains zeroes only, then the column was not calculated. This applies to all ordination results in CANOCO.

The scores of species and samples on the ordination axes depend on each other according to the transition formulae (see e.g. Jongman et al., 1987, exercise 5.1.3 and section 5.9). The precise form of these formulae is governed by the answer to Q14 or Q15. Let $\alpha = 1, 0, \frac{1}{2}$ depending on whether Q14/Q15 is answered by 1, 2 or 3, respectively. In linear methods, the species score (b_k) with respect to an ordination axis is a weighted sum of the sample scores $\{\tilde{x}_i\}$ if $\alpha = 0$; in general:

$$b_k = \lambda^{-\alpha} \sum_i w_i y_{ik} \tilde{x}_i / \sum_i w_i \qquad (4.3)$$

where y_{ik} is value of species k in sample i after data transformation and λ is the eigenvalue of the axis. In weighted averaging methods, the species score (u_k) on the ordination axis is a weighted average of the sample scores $\{\tilde{x}_i\}$ if $\alpha = 0$; in general

$$u_k = \lambda^{-\alpha} \sum_i w_i y_{ik} \tilde{x}_i / \sum_i w_i y_{ik}. \qquad (4.4)$$

Note the subtle difference between the formulae for b_k and u_k, which is partly implicit because in (4.3) y_{ik} is usually centred by species whereas in (4.4) y_{ik} is nonnegative. The difference has a dramatic effect: b_k is (proportional to) the slope of the regression line of a species' abundance values versus the ordination axis (Fig. 4.3), whereas u_k is the centre of the species' distribution along the axis (= approximate optimum; Fig. 4.4); cf. Jongman et al. (1987) and Ter Braak and Looman (1986). This difference is the major reason for using different symbols.

In indirect gradient analysis the sample scores denoted above by $\tilde{x}_i$ are the scores $\{x_i^*\}$ which are "derived from the species scores" (section 4.6) whereas in direct gradient analysis they are the scores $\{x_i\}$ which are linear combinations of environmental variables (section 4.7). The former scores make the species axis, the latter the environmental axis (section 4.3).

The formulae (4.3) and (4.4) also define how to obtain the scores of species excluded from the ordination (passive species).

In linear methods the (weighted) sum of squares of the species scores is set equal to

Table 4.4 Ordination results as listed on the output file. The table shows species scores on four ordination axes (AX1, ..., AX4). The same scores are also arranged in order of their rank order on axis k in the columns with heading "RANKED k". The title of the species data file is shown at the top, followed by the options in force. The options listed are: the ANALYSIS number (Q2) is 5 (= CCA), the number of CANONICAL AXES is 4 (because CCA is a direct gradient analysis technique), the number of COVARIABLES is 0, and the SCALING of the ordination scores (Q14 or Q15) is 1. The centring and standardization (CENT./STAND.) by samples and by species is applicable only in linear methods; it gives the answers to Q26-Q27. The scores have been multiplied by the MULTIPLIER 100. The analysis is without DETRENDING (Q9), DOWNWEIGHTING of rare species (Q32) and nonlinear RESCALING of axes (see Q11, except that 0 means: no rescaling). The number of SEGMENTS (Q10) and the rescaling THRESHOLD (Q12) are therefore 0. The iterative ordination algorithm is not modified by an extra RANKING step in which sample and/or species scores are replaced by rank numbers (see section 5). If the number following "CANONICAL AXES" is negative, then the axes were extracted as linear combinations of a previous set of environmental variables. After TRANSFORMATION the answer to question Q24 is given.

SPECIES - DUNE MEADOW DATA (M. BATTERINK AND G. WIJFFELS, 1983)
CANOCO -- ANALYSIS 5 CANONICAL AXES 4 COVARIABLES 0 SCALING 1 CENT./STAND. BY SAMPLES: 0/0 BY SPECIES: 0/0 MULTIPLIER 100.
DETRENDING 0 DOWNWEIGHTING 0 RESCALING 0 SEGMENTS 0 THRESHOLD 0.00 RANKING BY SAMPLES/SPECIES: 0/0
TRANSFORMATION -1.00 0.00

SPECIES SCORES

N	NAME	AX1	AX2	AX3	AX4	RANKED 1 EIG=0.461			RANKED 2 EIG=0.298			RANKED 3 EIG=0.160			RANKED 4 EIG=0.134		
1	ACH MIL	-169	83	8	-98	21	POT PAL	407	3	AIR PRA	391	21	POT PAL	583	15	JUN BUF	325
2	AGR STO	155	-109	-31	-24	30	CAL CUS	332	25	SAL REP	385	26	TRI PRA	277	23	RUM ACE	283
3	AIR PRA	148	391	-294	156	10	ELE PAL	300	12	EMP NIG	339	33	RAN ACR	275	26	TRI PRA	274
4	ALO GEN	71	-212	-95	41	22	RAN FLA	276	13	HYP RAD	325	23	RUM ACE	215	31	HIP RHA	242
5	ANT ODO	-77	170	-11	66	12	EMP NIG	242	31	HIP RHA	319	18	PLA LAN	160	33	RAN ACR	207
6	BEL PER	-143	37	-70	-211	14	JUN ART	193	28	VIC LAT	222	10	ELE PAL	157	12	EMP NIG	164
7	BRO HOR	-167	4	-44	-252	8	CHE ALB	187	5	ANT ODO	170	30	CAL CUS	105	3	AIR PRA	156
8	CHE ALB	187	-360	-189	8	25	SAL REP	157	18	PLA LAN	125	28	VIC LAT	87	25	SAL REP	153
9	CIR ARV	-80	-185	-269	-196	2	AGR STO	155	30	CAL CUS	99	27	TRI REP	82	14	JUN ART	96
10	ELE PAL	300	-20	157	-113	3	AIR PRA	148	16	LEO TAU	94	29	BRA RUT	38	24	SAG PRO	84
11	ELY REP	-128	-83	-101	-1	13	HYP RAD	112	21	POT PAL	87	22	RAN FLA	32	5	ANT ODO	66
12	EMP NIG	242	339	-396	164	4	ALO GEN	71	1	ACH MIL	83	14	JUN ART	13	29	BRA RUT	66
13	HYP RAD	112	325	-192	64	24	SAG PRO	60	29	BRA RUT	55	16	LEO TAU	9	13	HYP RAD	64
14	JUN ART	193	-10	13	96	15	JUN BUF	41	6	BEL PER	37	1	ACH MIL	8	18	PLA LAN	55
15	JUN BUF	41	-176	-34	325	29	BRA RUT	27	27	TRI REP	24	5	ANT ODO	-11	4	ALO GEN	41
16	LEO TAU	0	94	9	-20	31	HIP RHA	21	22	RAN FLA	15	17	LOL PER	-18	8	CHE ALB	8
17	LOL PER	-125	-9	-18	-83	16	LEO TAU	0	7	BRO HOR	4	2	AGR STO	-31	11	ELY REP	-1
18	PLA LAN	-161	125	160	55	27	TRI REP	-5	33	RAN ACR	4	15	JUN BUF	-34	20	POA TRI	-10
19	POA PRA	-92	-9	-35	-46	20	POA TRI	-53	17	LOL PER	-9	19	POA PRA	-35	16	LEO TAU	-20
20	POA TRI	-53	-116	-35	-10	33	RAN ACR	-71	19	POA PRA	-9	20	POA TRI	-35	2	AGR STO	-24
21	POT PAL	407	87	583	-253	5	ANT ODO	-77	14	JUN ART	-10	7	BRO HOR	-44	27	TRI REP	-37
22	RAN FLA	276	15	32	-46	9	CIR ARV	-80	26	TRI PRA	-13	6	BEL PER	-70	19	POA PRA	-46
23	RUM ACE	-165	-45	215	283	32	POA ANN	-85	32	POA ANN	-17	4	ALO GEN	-95	22	RAN FLA	-46
24	SAG PRO	60	-95	-122	84	19	POA PRA	-92	10	ELE PAL	-20	11	ELY REP	-101	30	CAL CUS	-75
25	SAL REP	157	385	-298	153	28	VIC LAT	-108	23	RUM ACE	-45	31	HIP RHA	-111	17	LOL PER	-83
26	TRI PRA	-223	-13	277	274	17	LOL PER	-125	11	ELY REP	-83	24	SAG PRO	-122	1	ACH MIL	-98
27	TRI REP	-5	24	82	-37	11	ELY REP	-128	24	SAG PRO	-95	32	POA ANN	-128	32	POA ANN	-110
28	VIC LAT	-108	222	87	-200	6	BEL PER	-143	2	AGR STO	-109	8	CHE ALB	-189	10	ELE PAL	-113
29	BRA RUT	27	55	38	66	18	PLA LAN	-161	20	POA TRI	-116	13	HYP RAD	-192	9	CIR ARV	-196
30	CAL CUS	332	99	105	-75	23	RUM ACE	-165	15	JUN BUF	-176	9	CIR ARV	-269	28	VIC LAT	-200
31	HIP RHA	21	319	-111	242	7	BRO HOR	-167	9	CIR ARV	-185	3	AIR PRA	-294	6	BEL PER	-211
32	POA ANN	-85	-17	-128	-110	1	ACH MIL	-169	4	ALO GEN	-212	25	SAL REP	-298	7	BRO HOR	-252
33	RAN ACR	-71	4	275	207	26	TRI PRA	-223	8	CHE ALB	-360	12	EMP NIG	-396	21	POT PAL	-253

$$\sum_k w_k b_k^2 = \lambda^{1-\alpha} \sum_i w_i \qquad (4.5)$$

to obtain valid biplots of species and samples (cf. Laurec et al., 1979). In weighted averaging methods the weighted mean square of the species scores is set equal to

$$\sum_k w_k^* u_k^2 / \sum_k w_k^* = \lambda^{1-\alpha}/(1-\lambda) \qquad (4.6)$$

with w_k^* as defined above. The factor $(1-\lambda)$ in (4.6) together with the definition of $\tilde{x}_i$ ensures that the species and sample scores are in Standard Deviation units (SD). Species scores in SD-units on average have, by definition, unit within-sample variance:

$$\sum_{i,k} w_i w_k y_{ik} (u_k - \tilde{x}_i)^2 / \sum_{i,k} w_i w_k y_{ik} = 1. \qquad (4.7)$$

With this scaling, the length of the ordination axis is, by definition, the range of the sample scores $\{\tilde{x}_i\}$. For the scores as displayed by CANOCO to conform the scaling defined by (4.5)-(4.7), the scores must be divided by the multiplier given with each table of scores (e.g. Tables 4.3 and 4.4).

When nonlinear rescaling of axes is in force (Q11), then u_k is not a simple function of $\{\tilde{x}_i\}$ and (4.6) does not hold, but the axis is in SD-units.

In weighted averaging methods (except when nonlinear rescaling is in force) the species scores have weighted mean 0, whereas in linear methods they have mean 0 only if the species data are centred by samples (Q26 = 1 or 3).

4.6 Sample scores

The sample scores that are derived from the species scores are simply called SAMPLE SCORES in the output, abbreviated to SAMP-SCOR in Q34. They make the species axis in section 4.3. The format is similar to that of the species scores. The column "WEIGHT" (Table 4.5) gives in linear methods the weights $\{w_i\}$ given to samples in Q28 - Q29 and in weighted averaging methods $w_i^* = w_i \sum_k w_k y_{ik}$, the weighted total abundance in a sample (where w_k denotes species weight!). Passive samples are always placed at the bottom of the table (see sample 20, "PAS SAMP", in Table 4.5).

The sample scores are derived from the species scores by the following transition formulae, where α is again 1, 0, $\frac{1}{2}$ depending on whether Q14/Q15 is answered by 1, 2 or 3, respectively. In linear methods, the sample score (x_i^*) on an ordination axis is a weighted sum of the species scores $\{b_k\}$ if $\alpha = 1$; in general

$$x_i^* = \lambda^{\alpha-1} \sum_k w_k y_{ik} b_k / \sum_k w_k . \qquad (4.8)$$

Table 4.5 Sample scores $\{x_i^*\}$ on each of the ordination axes as displayed at the terminal (N = sample number; WEIGHT is explained in the text). Sample 20 (PAS SAMP) is placed at the bottom because it is passive (weight = 0). These sample scores form the species axes.

SAMPLE SCORES

N	NAME	AX1	AX2	AX3	AX4		WEIGHT
	EIG	0.46	0.30	0.16	0.13		
1	1	-113	-32	-41	-49	!	18
2	2	-80	-16	-23	-67	!	42
3	3	-29	-66	-39	-25	!	40
4	4	-22	-61	-55	-41	!	45
5	5	-106	16	40	25	!	43
6	6	-95	24	79	59	!	48
7	7	-96	10	42	25	!	40
8	8	65	-47	-5	-11	!	40
9	9	-5	-65	-21	49	!	42
10	10	-89	39	10	-57	!	43
11	11	-47	61	6	-20	!	32
12	12	33	-94	-23	72	!	35
13	13	44	-105	-40	26	!	33
14	14	187	17	111	-71	!	24
15	15	181	3	95	-34	!	23
16	16	179	-45	35	-21	!	33
17	17	-36	181	-47	36	!	15
28	18	-29	97	-6	-3	!	27
29	19	61	187	-116	68	!	31
30	20	185	65	-12	13	!	31
20	PAS SAMP	-132	1	13	-33	!	0
	MULTIPLIER	100					

In weighted averaging methods, the sample score (x_i^*) is a weighted average of the species scores $\{u_k\}$ if $\alpha = 1$; in general

$$x_i^* = \lambda^{\alpha-1} \sum_k w_k y_{ik} u_k / \sum_k w_k y_{ik}. \qquad (4.9)$$

If there are covariables in the analysis, the scores $\{x_i^*\}$ are made uncorrelated to the covariables before they are printed in order to avoid distortion by the effects of covariables. The sample scores printed are the residuals of a regression of the scores in (4.9) on the covariables. Scores of passive samples for which the values of covariables are available are adjusted by use of the equation of the regression just mentioned.

In weighted averaging methods (except with nonlinear rescaling) the sample scores have weighted mean 0, whereas in linear methods they have mean 0 only in RDA or if the species data are centred by species (Q27 = 1 or 3).

In indirect gradient analyses: $\tilde{x}_i = x_i^*$ with $\tilde{x}_i$ as used in the previous section. The transition formulae consist then of the two formulae (4.3) and (4.8) for linear methods and (4.4) and (4.9) for weighted averaging methods. The (weighted) mean square of the sample scores follows then in linear methods from (4.3) and (4.5), giving

$$\sum_i w_i \tilde{x}_i^2 / \sum_i w_i = \lambda^\alpha \qquad (4.10)$$

and in weighted averaging methods from (4.4) and (4.6), giving

$$\sum_i w_i^* \tilde{x}_i^2 / \sum_i w_i^* = \lambda^\alpha / (1-\lambda) \qquad (4.11)$$

where w_i^* as defined above (except when nonlinear rescaling of axes is in force).

4.7 Regression/canonical coefficients, t-values and linear combinations of environmental variables

The regression/canonical coefficients (Table 4.6) are the coefficients of a weighted multiple regression of the sample scores $\{x_i^*\}$ from the previous section on the standardized environmental variables. There are four columns of coefficients because the regression is calculated for each ordination axis separately. Let z_{ij} be the value of environmental variable j (j = 1, ..., q) in sample i and let $\bar{z}_j$ and s_j be the mean and standard deviation of variable j as defined in section 4.3 (Table 4.1). The environmental variable is standardized to mean 0 and variance 1:

$$\tilde{z}_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j}. \qquad (4.12)$$

The regression/canonical coefficients are now derived from the weighted least squares fit of the multiple regression model

Table 4.6 Regression/canonical coefficients $\{c_j\}$ for each of the ordination axes as displayed at the terminal. They are actually canonical coefficients because the example concerns a direct gradient analysis by CCA.

REGRESSION/CANONICAL COEFFICIENTS FOR STANDARDIZED VARIABLES

N	NAME	AX1	AX2	AX3	AX4
	EIG	0.46	0.30	0.16	0.13
1	A1	115	-172	326	-222
2	MOISTURE	633	-240	-206	-6
3	MANURE	-29	-82	-232	-678
5	HAYPASTU	60	-140	107	171
6	PASTURE	247	11	248	60
8	BF	-87	111	-2	-552
9	HF	-194	72	178	-51
10	NM	188	583	-92	-502
	MULTIPLIER	1000			

Table 4.7 T-values of regression coefficients corresponding to Table 4.6. (N = variable number; FR EXPLAINED = fraction of variance explained, see text). Because Table 4.6 contains canonical coefficients this table has exploratory value only. Tables 4.6 and 4.7 jointly show that moisture is important in defining the first axis: the canonical coefficient of moisture (0.633) is the largest in absolute value
and its t-value (5.57) is appreciably higher than 2.1. On the second axis nature management has a high canonical coefficient but its t-value is only 2.25.

T-VALUES OF REGRESSION COEFFICIENTS

N	NAME	AX1	AX2	AX3	AX4
	FR EXPLAINED	0.38	0.24	0.13	0.11
1	A1	103	-137	306	-273
2	MOISTURE	557	-188	-190	-7
3	MANURE	-12	-30	-101	-385
5	HAYPASTU	41	-84	76	159
6	PASTURE	197	8	207	66
8	BF	-48	54	-1	-416
9	HF	-128	42	123	-46
10	NM	82	225	-42	-299
	MULTIPLIER	100			

$$x_i^* = c_0 + \sum_{j=1}^{q} c_j \tilde{z}_{ij} + \varepsilon_i \qquad (4.13)$$

where c_0 is the intercept, c_j the regression coefficient of environmental variable j and ε_i is the error term with mean 0 and variance inversely proportional to w_i in linear methods and to w_i^* in weighted averaging methods. Because the environmental variables are centred to mean 0, the intercept c_0 is equal to the mean of the species axis, i.e. $c_0 = 0$ except when nonlinear rescaling is in force. The other coefficients are estimated - using matrix notation - by

$$\underline{c} = (\tilde{Z}'W\tilde{Z})^{-1}\tilde{Z}'W\,\underline{x}^* \qquad (4.14)$$

where $\underline{c}$ and $\underline{x}$ are column vectors, $\underline{c} = (c_1, c_2, \ldots, c_q)'$, $\underline{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)'$, $\tilde{Z}$ is a $n \times q$ matrix with elements $\tilde{z}_{ij}$ and W is a $n \times n$ diagonal matrix with as i-th diagonal element w_i in linear methods and w_i^* in weighted averaging methods. The fitted values of the regression are

$$x_i = c_0 + \sum_{j=1}^{q} c_j \tilde{z}_{ij}. \qquad (4.15)$$

Note that the $\{c_j\}$ in (4.15) are estimates whereas in (4.11) they represent the true values, but this difference is not made explicit in the notation; the error term in (4.13) says enough. The fitted values $\{x_i\}$ are termed sample scores which are linear combinations of environmental variables (Table 4.8), abbreviated to LINEA-COM in Q34. They constitute the environmental axis of section 4.3. The correlation between the species axis $\{x_i^*\}$ and the environmental axis $\{x_i\}$ is the multiple correlation between the species axis $\{x_i^*\}$ and the environmental variables (= species-environment correlation).

In indirect gradient analyses, the sample scores $\{x_i^*\}$ are derived from the species data regardless of any environmental variables. The regression is calculated after extraction of the species and sample scores. The coefficients $\{c_j\}$ are therefore regression coefficients and have the well-known statistical properties of regression coefficients (see e.g. Mongomery and Peck, 1982). In contrast, the sample scores $\{x_i^*\}$ in direct gradient analyses also depend on the environmental variables in the analysis. The regression is calculated within the iterative ordination algorithm. The coefficients $\{c_j\}$ have been chosen so as to optimize the fit of the environmental axis $\{x_i\}$ to the species data (and not just to the species axis $\{x_i^*\}$). The coefficients are therefore given a different name: canonical coefficients. They do not have the same statistical properties as regression coefficients. In particular, canonical coefficients have a larger variance than regression coefficients.

Table 4.7 shows t-values of regression coefficients. Because the multiplier is given to be 100, the t-values are actually a factor 100 lower than displayed; the t-value of MOISTURE on the first axis, for example, is

Table 4.8 Sample scores $\{x_i\}$ on each of the ordination axes as displayed at the terminal. These scores scores form the environmental axes. They can be calculated from the regression/canonical coefficients and the standardized environmental variables by using equation (4.15).
Passive samples are not in this table; hence, sample 20 "PAS SAMP" is missing. The column WEIGHT is as in Table 4.5. In the CANOCO output this table appears after the centroid scores (Table 4.11).

SAMPLE SCORES - WHICH ARE LINEAR COMBINATIONS OF ENVIRONMENTAL VARIABLES

N	NAME	AX1	AX2	AX3	AX4		WEIGHT
	EIG	0.46	0.30	0.16	0.13		
1	1	-82	-28	-56	-9	!	18
2	2	-96	7	-10	-65	!	42
3	3	-36	-56	-41	-27	!	40
4	4	-37	-55	-43	-26	!	45
5	5	-110	-5	56	3	!	43
6	6	-110	-15	43	61	!	48
7	7	-77	24	36	9	!	40
8	8	78	-44	13	-9	!	40
9	9	-15	-16	-8	83	!	42
10	10	-71	29	-30	-48	!	43
11	11	-49	44	43	-36	!	32
12	12	50	-86	-5	54	!	35
13	13	86	-107	-30	1	!	33
14	14	203	52	92	-31	!	24
15	15	172	0	95	-37	!	23
16	16	129	-74	0	-16	!	33
17	17	4	140	-22	19	!	15
28	18	-29	148	0	12	!	27
29	19	112	101	-63	22	!	31
30	20	110	103	-67	24	!	31
	MULTIPLIER	100					

thus 5.57. The t-value of a regression coefficient c_j of variable j on an ordination axis is equal to $c_j/se(c_j)$, where $se(c_j)$ - the standard error of the estimate c_j - is the square root of var (c_j) given in (4.1). If the $\{c_j\}$ are regression coefficients, then the t-values can be used in Student t-tests in the usual way (e.g. Montgomery and Peck, 1982; Jongman et al., 1987, sections 3.2.1 and 3.5.2). To test the null hypothesis that the true coefficient of a particular variable on an axis is equal to 0, the t-value of the variable should be compared with the critical value of a Student t-distribution with n-q-1 degrees of freedom (n = number of samples, q = number of environmental variables). A variable is shown to contribute significantly to the regression if its t-value in absolute value exceeds the critical value (the critical value for a t-test at the 5% significance level is ca. 2.1, if n-q-1 > 18).

The Student t-test is not appropriate for tests of significance of canonical coefficients, because they have a larger variance. But the t-values still have an exploratory use. In particular, when the t-value of a variable is less than 2.1 in absolute value, then the variable does not contribute much to the fit of the species data in addition to the contributions of the other variables in the analysis. The variable then does not have an effect that is uniquely attributable to that particular variable (see Jongman et al. 1987, section 3.5.3) and can be deleted without much affecting the canonical eigenvalues. The t-values are therefore of help when one wants to select a subset of environmental variables that explains the species data almost equally well as the full set. The t-values are unimportant, when the only aim of the analysis is to prepare a species-environment biplot.

Note that a table of regression/canonical coefficients of a direct gradient analysis may contain both canonical coefficients and regression coefficients: the first column(s) contain(s) canonical coefficients whereas the later columns may contain regression coefficients. How many columns contain canonical coefficients is indicated by the number of "CANONICAL AXES" given in the heading of a table on the output file. The different columns of the corresponding table of t-values then have different statistical properties!

The fraction of variance that an ordination axis explains in the species-environment biplot is also given in the table of t-values (FR EXPLAINED). It is the same information as given cumulatively in section 4.4 (at the bottom of Table 4.2). In CCA and RDA the fraction explained by axis k is simply λ_k/(sum of all canonical eigenvalues). See section 4.4 for further explanation.

If there are covariables in the analysis, then the values $\tilde{z}_{ij}$ in (4.13) are replaced by the residuals of a multivariate multiple regression of the standardized environmental variables on the covariables (without any further standardization). The regression in (4.13) is then a partial multiple regression (Kendall and Stuart, 1973, sections 27.8 and 27.25; Seber, 1977, Theorem 3.7 (ii); note that there is no need to regress the species data on the covariables). In the variance of the partial regression coefficient c_j (4.1) and in the Student t-test, q must be replaced by q+p, where p is the number of covariables. When the partial regression is performed after

Table 4.9 Inter-set correlations of the environmental variables with the species axes as displayed at the terminal (FR EXTRACTED = fraction of variance in environmental data extracted by axis, see text). Moisture has the largest correlation on the first axis. Both manure and nature management are strongly correlated with the second axis. Note that the importance of manure on the second axis was not obvious from Tables 4.6 and 4.7, presumably because of its correlation with nature management.

INTER SET CORRELATIONS OF ENVIRONMENTAL VARIABLES WITH AXES

N	NAME	AX1	AX2	AX3	AX4
FR EXTRACTED		0.22	0.16	0.07	0.07
1	A1	539	-156	504	-97
2	MOISTURE	883	-153	-120	151
3	MANURE	-296	-690	-169	-160
5	HAYPASTU	-165	-499	-113	-77
6	PASTURE	268	-28	366	-188
8	BF	-349	158	-25	-520
9	HF	-346	-105	376	464
10	NM	546	666	1	38
MULTIPLIER		1000			

Table 4.10 Biplot scores of environmental variables as displayed at the terminal (R(SPEC, ENV) = species-environment correlation, see table 4.2). The scores are used in the construction of the arrows in Fig. 4.2. For explanation see text.

BIPLOT SCORES OF ENVIRONMENTAL VARIABLES

N	NAME	AX1	AX2	AX3	AX4
R(SPEC,ENV)		0.96	0.90	0.86	0.89
1	A1	281	-79	216	-37
2	MOISTURE	460	-78	-51	58
3	MANURE	-154	-350	-72	-61
5	HAYPASTU	-86	-253	-48	-29
6	PASTURE	139	-14	157	-72
8	BF	-182	80	-11	-199
9	HF	-180	-53	161	178
10	NM	284	338	0	14
MULTIPLIER		1000			

extracting the ordination, we obtain an indirect analysis of a partial component analysis, a partial CA or a partial DCA; when it is incorporated within the iterative ordination algorithm, we obtain a direct analysis: partial RDA, partial CCA or partial DCCA.

In section 4.5 it was discussed how the species scores were derived and how they were scaled. In direct gradient analyses the species scores are derived from the sample scores that are linear combination of the environmental variables, i.e. the $\{\tilde{x}_i\}$ in section 4.5 are then equal to the $\{x_i\}$ of the present section. With $\tilde{x}_i = x_i$, four formulae jointly constitute the transition formulae, for linear methods the formulae (4.3), (4.8), (4.14) and (4.15) and for weighted averaging methods the formulae (4.4), (4.9), (4.14) and (4.15). The scaling of the $\{x_i\}$ follows from these formulae and the scaling of the species scores given in (4.5) and (4.6). It can be shown that the (weighted) mean square of the sample scores $x_i = \tilde{x}_i$ satisfies for linear methods (4.10) and for weighted averaging methods (4.11) [except when nonlinear rescaling is in force]. The weighted mean square of the sample scores $\{x_i\}$ is always a factor R^2 smaller than the mean square of the sample scores $\{x_i^*\}$ where R is the species- environment correlation of the corresponding axis (this follows from the regression in (4.13 - 4.15)). For (partial) CCA and DCCA, (4.7) holds with $\tilde{x}_i = x_i$.

For interpretation purposes it is sometimes of interest to obtain the regression/canonical coefficients, $\{c_j^*\}$ say, corresponding to the environmental variables in their original units of measurement, i.e. without the standardization in (4.12). By inserting (4.12) in (4.15) we obtain after some elementary algebraic manipulation

$$x_i = \{c_0 - \sum_{j=1}^{q} \frac{\bar{z}_j}{s_j}\} + \sum_{j=1}^{q} (\frac{c_j}{s_j}) z_{ij}. \qquad (4.16)$$

The desired coefficients are thus $c_j^* = c_j/s_j$. The value of s_j is found in the table of means and standard deviations (section 4.3).

4.8 Inter-set correlations of environmental variables with axes

The inter-set correlations of environmental variables with axes (Table 4.9) are the correlation coefficients between the environmental variables and the species axes consisting of the sample scores $\{x_i^*\}$ (section 4.6). The same correlations can also be found in the full correlation matrix on the output file (section 4.3). If there are covariables in the analysis, the correlations given are partial correlations.

In indirect gradient analyses the species axes do not depend on the environmental variables. The inter-set correlation for a particular variable is then not dependent on which other environmental variables are included in the analysis. But in direct gradient analyses, the species axes may depend on the environmental variables included and therefore the inter-set correlations may also change.

In contrast to regression/canonical coefficients, the inter-set correlations do not become unstable when the environmental variables are strongly correlated with each other, i.e. when the VIF's of section 4.3 are

large. See also section 7.3 and Ter Braak (1986a: canonical coefficients and intra set correlations).

Table 4.9 also shows the fraction of the total variance in the standardized environmental data that is extracted by each species axis (FR EXTRACTED). The fraction extracted is equal to the mean squared inter-set correlation, $\sum_j r_j^2/q$, where r_j is the inter set correlation of variable j (cf. Gittins, 1985, section 3.2.2).

4.9 Biplot scores of environmental variables

In the species-environment biplot (section 4.4) environmental variables are represented by arrows. The biplot scores of environmental variables (Table 4.10) give the coordinates of the heads of the arrows, the coordinates of the species being given in section 4.5. Another way to display nominal environmental variables is described in the next section.

The rules for constructing and interpreting species-environment biplots are the same as those given in Jongman et al. (1987, section 5.3.4) for PCA biplots. Because the scores for species and for environmental variables are often of a different order of magnitude, the biplot is constructed most easily by drawing separate plots of species and of environmental variables on transparent paper, each one with its own scaling. But note that within each plot the scale units of the axes must have equal physical length. The biplot is obtained by superimposing the plots with the axes aligned and the origins of the coordinate systems coinciding.

However, there is an exception on the rule that the origins must coincide: when the mean of a species axis is nonzero (section 4.3), then the origin of the "environmental plot" must coincide with the point in the "species plot" whose coordinates are equal to the means of the species axes (SPEC AXk) given in section 4.3. This exception happens in linear methods when the species data are not centred by species (Q27), and in weighted averaging methods when nonlinear rescaling of axes is in force (Q11). In both cases there is an extra line below the species and sample scores, starting with the word "CENTROID" which specifies the means of the species axes. (Note that in the linear case this centroid is not necessarily equal to the mean of the species scores.) If one does not want to draw separate plots, the head of an environmental arrow can be added to the plot of the species at the point whose coordinates are obtained by the formula

(Y × biplot-score-of-environmental-variable)+(mean-of-species-axis)

where Y is a constant to be chosen by the user such that all heads of arrows fit in the species diagram.

As noted in section 4.4, the species-environment biplot serves to give, in linear methods, a display of approximate values of covariances between species and environmental variables and, in weighted averaging methods, of weighted averages of species with respect to environmental variables. The biplot scores of environmental variables can in principle be obtained by a weighted multivariate multiple regression of these covariances/weighted averages on the species scores (see Ter Braak 1986a). This regression is

actually calculated in CANOCO when detrending-by-segments or nonlinear rescaling of axes or a ranking method (a nonstandard analysis) is in force. But for other methods CANOCO uses a short-cut: as shown in Appendix C the regression gives biplot scores that are a simple function of the inter-set correlations (section 4.8), the standard deviations of the species axes (section 4.3) and the eigenvalues, namely

(inter-set correlation) × (standard deviation of species axis) × a (4.17)

where for linear methods a = 1 and for weighted averaging methods a = $1-\lambda_k$ where λ_k is the eigenvalue of axis k.

If there are covariables in the analysis, (4.17) is multiplied by the residual standard deviation of the regression of each standardized environmental variable on the covariables (= the square root of the diagonal element of the partial covariance matrix displayed at the terminal). In this way, the arrow of an environmental variable becomes shorter, the higher the correlation between this environmental variable and the covariables, (i.e., the more the variation in the environmental variable is already explained by the covariables). The arrow is unaffected when the environmental variable is not correlated to the covariables, i.e. when it contributes entirely new information about the environment. With covariables in the analysis the species-environment biplot approximates in linear methods partial covariances and, in weighted averaging methods, weighted averages with respect to residuals of environment variables (i.e. the environmental variables after eliminating covariable-effects). See Appendix C.

Environmental variables with long arrows are the most important in the analysis; the larger the arrow, the more confident one can be about the inferred covariances (correlations) or weighted averages. Table 4.10 also shows the species-environment correlations, R(SPEC,ENV). A value of 0.00, would indicate that the correlation could not be calculated by CANOCO.

4.10 Centroids of environmental variables in the ordination diagram

Nominal environmental variables can naturally be represented by points in the ordination diagram (Ter Braak, 1986a). Each class of a nominal variable gives one point which is located at the centroid of the sample scores belonging to the class. The centroid on the ordination axis with sample scores $\{x_i^*\}$ is given by

$$\sum_i \tilde{w}_i z_{ij} x_i^* / \sum_i \tilde{w}_i z_{ij} \qquad (4.18)$$

where z_{ij} is (as in section 4.7) the value of environmental variable j in sample i and $\tilde{w}_i = w_i$ (Q30 - Q31) in linear methods and $w_i^* = w_i \sum_k w_k y_{ik}$ in weighted averaging methods. In contrast to biplot scores, the centroids must be plotted in the ordination diagram in the same scale as the sample scores (Fig. 4.2). Representing environmental variables by points is not only useful for classes (dummy variables with $z_{ij} = 0$ or 1) but sometimes also for nonnegative quantitative variables that can be absent, i.e. where the value 0 has a special meaning.

Table 4.11 Centroids of environmental variables (with positive mean values) in the ordination diagram. The scores are used for the points of classes of nominal variables in Fig. 4.2. For explanation see text.

CENTROIDS OF ENVIRONMENTAL VARIABLES (MEAN.GT.0) IN ORDINATION DIAGRAM

N	NAME	AX1	AX2	AX3	AX4
R(SPEC,ENV)		0.96	0.90	0.86	0.89
1	A1	207	-45	102	-17
2	MOISTURE	527	-69	-38	41
3	MANURE	-205	-357	-62	-51
5	HAYPASTU	-189	-429	-68	-40
6	PASTURE	452	-36	326	-145
8	BF	-743	251	-29	-506
9	HF	-497	-113	286	306
10	NM	992	904	1	31
MULTIPLIER		1000			

Table 4.12 Effect of ploughing times on weeds in a barley crop: data file in full format with covariables (4 blocks) and environmental variables (3 treatments = 3 ploughing times) representing a randomized block experiment of 12 sample units laid out in 4 blocks of 3 sample units each. The experiment, which is analyzed further in section 6, is used to illustrate the Monte Carlo permutation test for testing the effect of environmental variables (here: treatments) on species composition. The file has the name "PLOUGH.EXP" in Table 4.13

```
 PLOUGH TIMES IN A 4X3 RANDOMIZED BLOCK EXPERIMENT (DATA B.J.POST)
(I3,X,4F1.0,X,3F1.0)
7
 1 1000 100
 2 1000 010
 3 1000 001
 4 0100 100
 5 0100 010
 6 0100 001
 7 0010 100
 8 0010 010
 9 0010 001
10 0001 100
11 0001 010
12 0001 001
00
BLOC  1 BLOC  2 BLOC  3 BLOC  4 PLTIME 1PLTIME 2PLTIME 3
```

CANOCO calculates the centroids defined by (4.18) for all environmental variables whose mean is positive. The value assigned to the other environmental variables is the mean of the species axis, i.e. the value assigned is 0, unless the species data are not centred or nonlinear rescaling of axes is in force. For variables whose mean is positive but which have some negative values, (4.18) is nonsensical yet given by CANOCO.

CANOCO uses a short-cut to calculate (4.18):

$$\left(\begin{matrix}\text{mean}\\ \text{species axes}\end{matrix}\right) + \frac{\left(\begin{matrix}\text{inter set}\\ \text{correlation}\end{matrix}\right) \times \left(\begin{matrix}\text{standard deviation}\\ \text{species axis}\end{matrix}\right) \times \left(\begin{matrix}\text{standard deviation}\\ \text{environmental variable}\end{matrix}\right)}{\text{mean of environmental variable}} \qquad (4.19)$$

the entries of which are all given in section 4.3. With covariables in the analysis, the standard deviation of the environmental variable in (4.19) is replaced by the residual standard deviation of the regression of the unstandardized environmental variable on the covariables (see Appendix C).

It is of some interest to note that we obtain the same value if we insert x_i for x_i^* in (4.18), i.e. the centroid of sample scores $\{x_i\}$ for a class is equal to the corresponding centroid of the sample scores $\{x_i^*\}$. This is, however, not true in general in additional passive analyses of (other) environmental variables (Q35).

Why nominal variables are naturally represented by points is best seen when there is a single nominal environmental variable, i.e. when there is a single pre-defined classification of samples. CCA with a series of dummy variables reflecting this classification provides an ordination to show maximum separation among the pre-defined groups of samples. This analysis is mathematically equivalent with Feoli and Orlóci's (1979) "analysis of concentration" and also with a simple CA of a two-way table of species-by-groups, the cells of which contain the total abundance of each of the species in each of the groups of samples (Greenacre 1984, section 7.1) In the CA ordination diagram the groups would be represented by points as they take the place of the samples. So why not represent the groups by the same points in a CCA? Similarly, RDA with a series of dummy variables is a variant of canonical variates analysis/multiple discriminant analysis, in which the groups are always represented by group means, i.e., by centroids (4.18).

The species points and the environmental points jointly reflect the relative abundance of species among the environmental classes. The interpretation is the same as that of a diagram of species and sites resulting from a PCA for linear methods and from a CA for weighted averaging methods.

To obtain a direct gradient analysis or a regression analysis subsequent to an indirect analysis, one dummy variable per nominal variable must be removed from the analysis (Q20). In consequence, CANOCO does not give the centroids for the deleted dummy variables. To obtain their centroids, it is most convenient to ask subsequently for an additional passive analysis, (Q35 = 2) in which the same environmental data are entered again, but now without deleting dummies (see Table 3.3). This gives a complete list of centroids and, if wanted, of inter-set correlations and biplot points.

Table 4.13 Annotated copy of the answers entered at the terminal to obtain a Monte Carlo permutation test with covariables in the analysis. The file "PLOUGH.EXP" is shown in Table 4.12. For explanation see text.

```
QUESTION - INPUT AT TERMINAL                          = ANNOTATION

Q1 :      PLOUGH.OUT                                  = OUTPUT FILE
Q2 :        2 =  ANALYSIS NUMBER
Q3 :      PLOUGH83.SPE                                = FILE WITH SPECIES DATA
Q5 :      PLOUGH.EXP                                  = FILE WITH ENVIRONMENTAL DATA
Q6 :        1 = COVARIABLES?
Q7 :      PLOUGH.EXP                                  = FILE WITH COVARIABLES
Q15:        1 = SCALING OF SAMPLE AND SPECIES SCORES?
Q16:        0 = MACHINE READABLE COPY OF SOLUTION?
Q19:           0 = SAMPLE NUMBER TO BE OMITTED
Q20:           1 ENVIRONMENTAL VARIABLE TO BE OMITTED
               2 ENVIRONMENTAL VARIABLE TO BE OMITTED
               3 ENVIRONMENTAL VARIABLE TO BE OMITTED
               4 ENVIRONMENTAL VARIABLE TO BE OMITTED
               7 ENVIRONMENTAL VARIABLE TO BE OMITTED
               0 ENVIRONMENTAL VARIABLE TO BE OMITTED
Q21:        -1    0  = PRODUCT OF ENVIRONMENTAL VARIABLES
Q22:           4             COVARIABLE TO BE OMITTED
               5             COVARIABLE TO BE OMITTED
               6             COVARIABLE TO BE OMITTED
               7             COVARIABLE TO BE OMITTED
               0             COVARIABLE TO BE OMITTED
Q23:        -1    0  = PRODUCT OF             COVARIABLES
Q24:         -1.00    0.00 = TRANSFORMATION OF SPECIES DATA
Q26:        0 = CENTRING/STANDARDIZATION BY SAMPLES
Q27:        1 = CENTRING/STANDARDIZATION BY SPECIES
Q28:         1.00000 = WEIGHT ( NOWEIGHT=1)
Q30:         1.00000 = WEIGHT ( NOWEIGHT=1)
Q33:        1 = OUTPUT OF CORRELATIONS?
Q34:        1  1  1  1  1  1  1  0 = ORDINATION OUTPUT
Q35:        1 = STOP, MORE ANALYSES, OTHER ENV. DATA?
Q36:        0 = ONLY ENV. VARS TO BE DELETED?
Q37:         1 = MONTE CARLO PERMUTATION TEST?
Q38:        99 = NUMBER OF PERMUTATIONS
Q39:      23239   945 = SEEDS FOR RANDOM NUMBERS
Q40:        3 = COVARIABLES TO RESTRICT PERMUTATION
Q35:        0 = STOP, MORE ANALYSES, OTHER ENV. DATA?
```

4.11 Monte Carlo permutation test

The Monte Carlo permutation of samples can either be restricted or unrestricted (Q40). In restricted permutation CANOCO determines classes of samples within which samples are randomly permuted. The classes, termed permutation classes, are constructed from the conditioning covariables: samples belong to the same permutation class if they have identical values on the conditioning covariables.

For example, Table 4.12 shows the covariables and environmental variables of an experiment to test the effect of month of ploughing (three ploughing times) on the subsequent composition of weeds in a barley crop. The experiment was a randomized block experiment of 12 sample units laid out in four blocks of three sample units each. The three treatments (ploughing times) were randomized within each block. The first four dummy variables in Table 4.12 each represent a block, and the last three dummy variables each a ploughing time. We are interested in the effect of ploughing time and want to eliminate possible effects of blocks. The block variables (variables 1-4) should thus be entered in CANOCO as covariables, and the ploughing time variables (variables 5-7) as environmental variables (to avoid collinearity one variable of both sets should be deleted; cf. Q20 and Q22). This is achieved in the terminal dialogue, of which an annotated copy is shown in Table 4.13, by specifying the name of Table 4.12 (PLOUGH.EXP) as both the file with environmental data (Q5) and the file with covariables (Q7). Subsequently, all environmental variables except the variables 5 and 6 (ploughing time 1 and ploughing time 2) are deleted and all covariables except the variables 1, 2 and 3 (block 1, block 2 and block 3) are deleted (Table 4.13). In this way the ploughing times 1 and 2 become the environmental variables and the blocks 1, 2 and 3 become the covariables. It is further asked for (Table 4.13) to carry out 99 permutations and, in response to question Q40, to condition permutations on three covariables, i.e. on the first three covariables which are in this case the variables 1, 2 and 3. From Table 4.12 it is seen, for example, that the samples 1, 2 and 3 have identical values on these three variables (namely 1 0 0) and therefore belong to the same permutation class. No other samples belong to this permutation class. Similarly, the samples 7, 8 and 9 score values 0 0 1 on covariables 1, 2 and 3 (because they belong to block 3) and thus belong to the same permutation class. The user can check in the output of CANOCO (Table 4.14) that these are actually the permutation classes reconstructed by CANOCO.

Samples in a permutation class will be randomly permuted. For example, the species data and covariable data of the samples 7, 8 and 9 will be linked in a particular permutation to, for example, the environmental data of the samples 8, 9 and 7 respectively. In this way sample 7 in the species data is assigned notionally to ploughing time 2 (whereas it was actually ploughed at the first date), etc. If also the samples in the other permutation classes are permuted, a random dataset is obtained. If different ploughing times result in large differences in weed composition, then the differences between ploughing times after permutation, as measured by the trace or the first eigenvalue, are likely to be smaller than in the data observed. This is

Table 4.14 Test of significance of the first canonical ordination axis. Report of permutation classes used in restricted permutation and of the first eigenvalue of the current data (DATA) and of the random data sets generated by Monte Carlo permutation. The values below TRACE are all 0.000 because an overall test of significance was not requested (Q37 = 1). P-value = exact Monte Carlo significance level. Because P = 0.05, the first ordination axis is just significant at the 5% significance level. The effect of ploughing time on weed composition is therefore just significant. For explanation see text.

*** THE PERMUTATIONS ARE CONDITIONED ON THE FIRST 3 COVARIABLE(S) ***

PERMUTATION CLASS 1 CONTAINS THE SAMPLES NUMBERED:
1 2 3

PERMUTATION CLASS 2 CONTAINS THE SAMPLES NUMBERED:
4 5 6

PERMUTATION CLASS 3 CONTAINS THE SAMPLES NUMBERED:
7 8 9

PERMUTATION CLASS 4 CONTAINS THE SAMPLES NUMBERED:
10 11 12

NO	TRACE	FIRST EIGENVALUE
DATA	0.000	0.162
1	0.000	0.081
2	0.000	0.108
3	0.000	0.098
4	0.000	0.115
5	0.000	0.059
.....		
..... simulations 6 - 94 not shown		
.....		
95	0.000	0.109
96	0.000	0.095
97	0.000	0.074
98	0.000	0.090
99	0.000	0.082
P-VALUE	1.00	0.05

indeed what happens in the example (Table 4.14). The first eigenvalue in the data as observed is 0.162 (see the line in Table 4.14 which begins with DATA); the value of trace is set equal to 0, because the test does not concern the trace (Q37=1). The first permutation results in a data set whose first eigenvalue is 0.081. Further permutations result in first eigenvalues 0.108, 0.098, 0.115, 0.059, ...- all smaller than the observed value 0.162. After 99 permutations, four permutations have resulted in a value larger than or equal to 0.162, so the exact Monte Carlo significance value is (4+1)/(99+1) = 0.05 (Table 4.14: P-value; see Hope, 1968).

After a permutation test has been carried out, CANOCO again calculates the eigenvalues and ordination axes of the data as observed. The iteration report and eigenvalues are displayed at the terminal and the output file to enable the user to check that the current data as held by CANOCO are still all right. This is important if there are covariables in the analysis. If there are covariables, the data may have been obstructed during the permutation test by cumulative rounding errors in the calculations. In the limited experience so far with the permutation test (with FORTRAN REALS of 16 bits on a VAX-computer) no such obstruction was detected, but nevertheless the user is warned! If the eigenvalues displayed are identical to those before the permutation test, the analysis can be continued through question Q35.

5. NONSTANDARD ANALYSES

A nonstandard analysis may be obtained by typing the number 10 in response to Q2. The user is warned that the program has been tested with less regard to nonstandard analyses than to standard analyses, and that the nonstandard analyses do not have a secure theoretical basis. In a nonstandard analysis the user is allowed to combine options in a nonstandard way. For example, the question about nonlinear rescaling (Q11) is posed standardly only when detrending-by-segments is in force, but in a nonstandard analysis this question is posed for all weighted averaging methods. In this way the user may specify an analysis in which nonlinear rescaling of axes is used in combination with detrending-by-polynomials or without detrending. (In DECORANA the rescaling question was also posed in basic correspondence analysis, but had no effect.) Nonlinear rescaling of axes is also possible in CCA and DCCA, but its use is somewhat illogical: the optimal linear combinations of environmental variables are searched for, but after these combinations have been determined, they are modified by the nonlinear rescaling of axes, so destroying their optimality property.

The only additional question in a nonstandard analysis, which is posed immediately after Q2 is:

Q41. TYPE VALUES OF ITY, NEIGZ, IORD, JORD

ITY : Allowed values are -100, -4, -3, -2, -1, 0, 1, 2, 3, 4. The sign of ITY discriminates between linear methods (ITY < 0) and weighted averaging methods (ITY $\geq$ 0). The absolute value of ITY determines the order of the polynomial used for detrending. The usual orthogonalization procedures in PCA and CA are thus in force if ITY = -1 and 1, respectively. The values ITY = 0 and -100 have a special meaning: if ITY = 0, then detrending-by-segments is in force; ITY = -100 corresponds to the option in Q27 requiring no centring by species.

NEIGZ : Allowed values are 0, 1, 2, 3, 4. Type 0 for an indirect gradient analysis, 4 for a direct gradient analysis, and 1, 2 or 3 for a hybrid analysis (cf. Q8).

IORD : Allowed values are 0 and 1. Type 1 to replace the samples scores in each step of the iterative ordination algorithm by their rank number, else type 0. For technical reasons, IORD = 1 should not be used in conjunction with detrending-by-segments.

JORD : Allowed values are 0 and 1. Type 1 to replace the species scores in each step of the iterative ordination algorithm by their rank number, else type 0. For technical reasons, JORD = 1 should only be used if the species are numbered consecutively; it should not be used if some species numbers are absent from the data.

The major additional possibility in nonstandard analysis is thus to modify the iterative ordination algorithm so that at each iteration the species scores and/or the site scores are replaced by rank numbers. This modification is described by Ihm and Van Groenewoud (1984: p. 29-30) under the name "reciprocal ranking". This procedure is a heuristic way to circumvent the problem that CA is sensitive to the occurrence of deviant samples and rare species in the data set (Jongman et al. 1987: section 5.2.6). For solving this problem, ranking of either sample scores or species scores will be sufficient in most cases. It may be more appealing to rank species scores than to rank sample scores: ranking of species scores imposes upon the solution a species packing model in which the species'optima are equally spaced (cf. Hill and Gauch, 1980: p. 49).

I do not know whether the reciprocal ranking algorithm gives unique species and samples scores, irrespective of the initial scores. To lessen this possible dependency on initial scores, ranking of scores is performed after two iterations of the iterative ordination algorithm starting from the usual initial scores. (Note that one iteration of this algorithm as implemented in CANOCO involves 4 passes of the data.) For technical reasons, the reciprocal ranking algorithm usually does not converge in 15 iterations; nevertheless the final scores are precise enough for most practical purposes. The final iteration is performed without ranking. If IORD or JORD is equal to 1, then it is implied that samples scores are derived from the species scores (Q14/15 = 1). By consequence, the final scores satisfy the equations (4.8), (4.9), (4.14) and (4.15) with $\alpha = 1$. However, the species scores are not a simple function of the samples scores; the equations (4.3) and (4.4) do not hold. In linear methods the mean square of the sample scores is set to 1 (cf. (4.10)). In weighted averaging methods the sample and species scores are scaled to SD-units, either by nonlinear rescaling of axes or by using equation (4.7).

Question 41 also allows detrending-by-polynomials to be used in linear methods. This use is, however, not supported by theory and therefore not recommended: linear methods applied to data arising from unimodal models produce a "horseshoe" which scrambles the order of samples along the first axis. In contrast, the "arch" produced by weighted averaging methods does not scramble the order of samples along the first axis.

6. EXAMPLES

6.1 Dune meadow data

The dune meadow data (section 2.1) were collected to investigate the differences in vegetation among dune meadows that have been subjected to different management regimes, namely standard farming, biodynamic farming, hobby farming and nature management. To obtain a display of the differences, we applied canonical correspondence analysis to the vegetation data (Table 2.2) with three of the four dummy management variables as environmental variables (variables 7-10 of Table 2.3). Fig. 6.1 displays the ordination diagram of the species scores, sample scores and centroids of the management variables on axes 1 and 2. The first axis ($\lambda_1 = 0.32$) is seen to separate the meadows receiving nature management from the remaining meadows and the second axis ($\lambda_2 = 0.18$) separates the meadows of standard farms from those of hobby farms and biodynamical farms, although the separations are not perfect. The species displayed on the right-hand side of the diagram occur mainly in the meadows receiving nature management, and those on the upper-left in the meadows of standard farms, and so on. To investigate whether the observed differences could be accounted for by pure chance, we used the Monte Carlo permutation test with the first eigenvalue as test statistic. The 99 random data sets generated by random permutation of meadows all yielded a lower eigenvalue. It is therefore concluded that there are significant differences in vegetation among the management regimes ($P \leq 0.01$).

A further question of interest is whether the differences in vegetation can be fully accounted for by three environmental variables related to soil characteristics; (1) thickness of the A1 horizon, (2) moisture and (3) quantity of manuring, whose effects are clear from Fig. 4.2, or whether the variation that remains after fitting these three environmental variables is systematically related to management regime. To answer this question a partial canonical correspondence analysis was carried out with the three environmental variables as covariables and the dummy management variables as the variables-of-interest (= environmental variables in CANOCO). The first eigenvalue of this analysis ($\lambda_1 = 0.15$) was subjected to the Monte Carlo permutation test (99 unrestricted permutations) leading to a P-value of 0.20. The variation in vegetation that remains after fitting thickness of the A1 horizon, moisture and manure is therefore not significantly related to management regime. In conclusion, the three soil characteristics are sufficient to explain the vegetation differences between management regimes.

6.2 Weeds in summer barley

Post (1986) carried out a randomized block experiment to investigate, among other things, the effect of time of ploughing on the subsequent weed vegetation composition in summer barley. The experimental design has already been described in section 4.11, the plot size was 3×2 m^2. The ploughing times were in 1983 March 9, March 23 and April 6. The log-

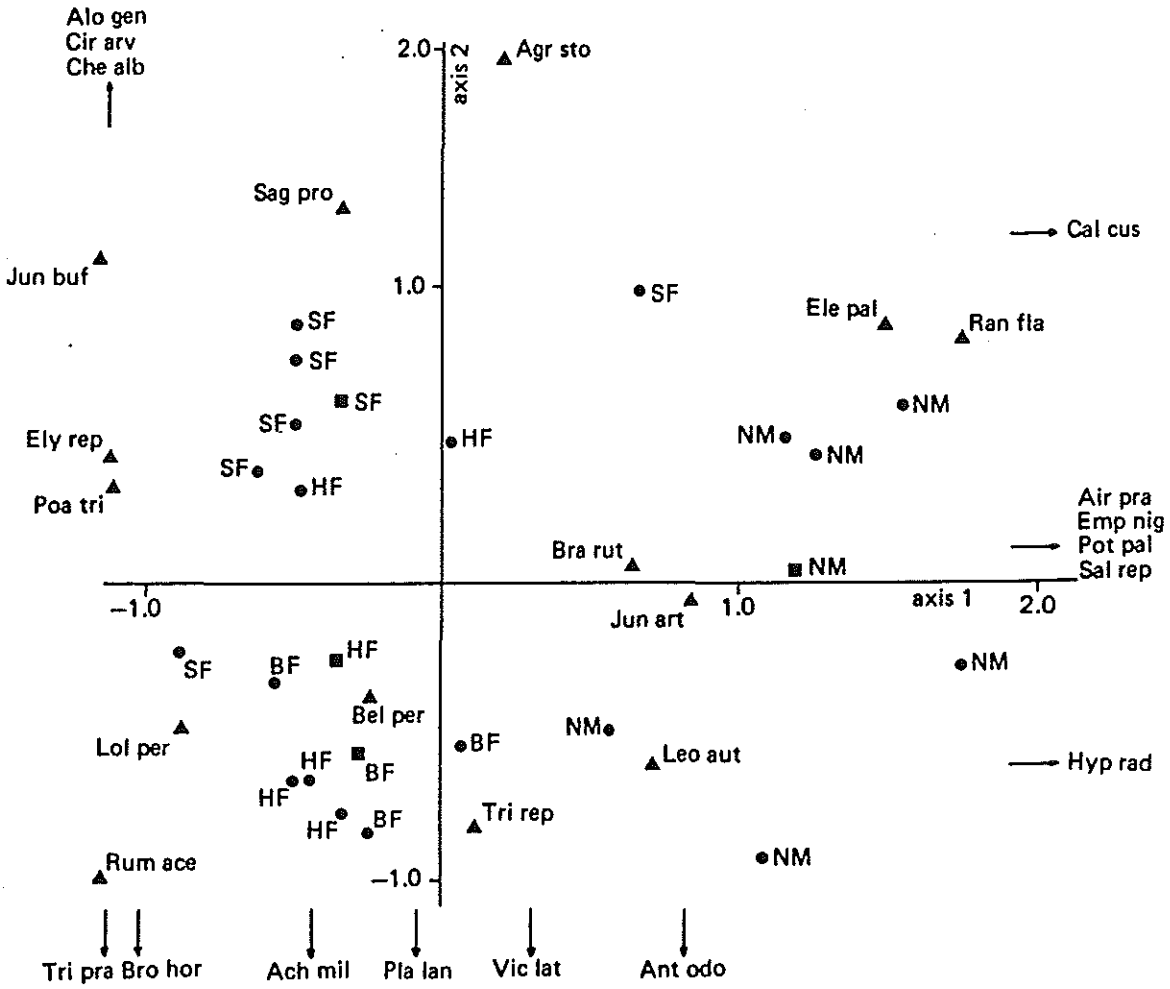

Fig.6.1 Ordination diagram based on canonical correspondence analysis of the dune meadow vegetation data with respect to management regime. The diagram optimally displays the differences in species composition among different types of management. The types of management (■) are located in the diagram at the centroids of the samples (●) belonging to that type (see section 4.10). For abbreviations of species (▲) and management types see Table 2.1 and 2.3 (λ_1 = 0.32, λ_2 = 0.18, scaling α = 1).

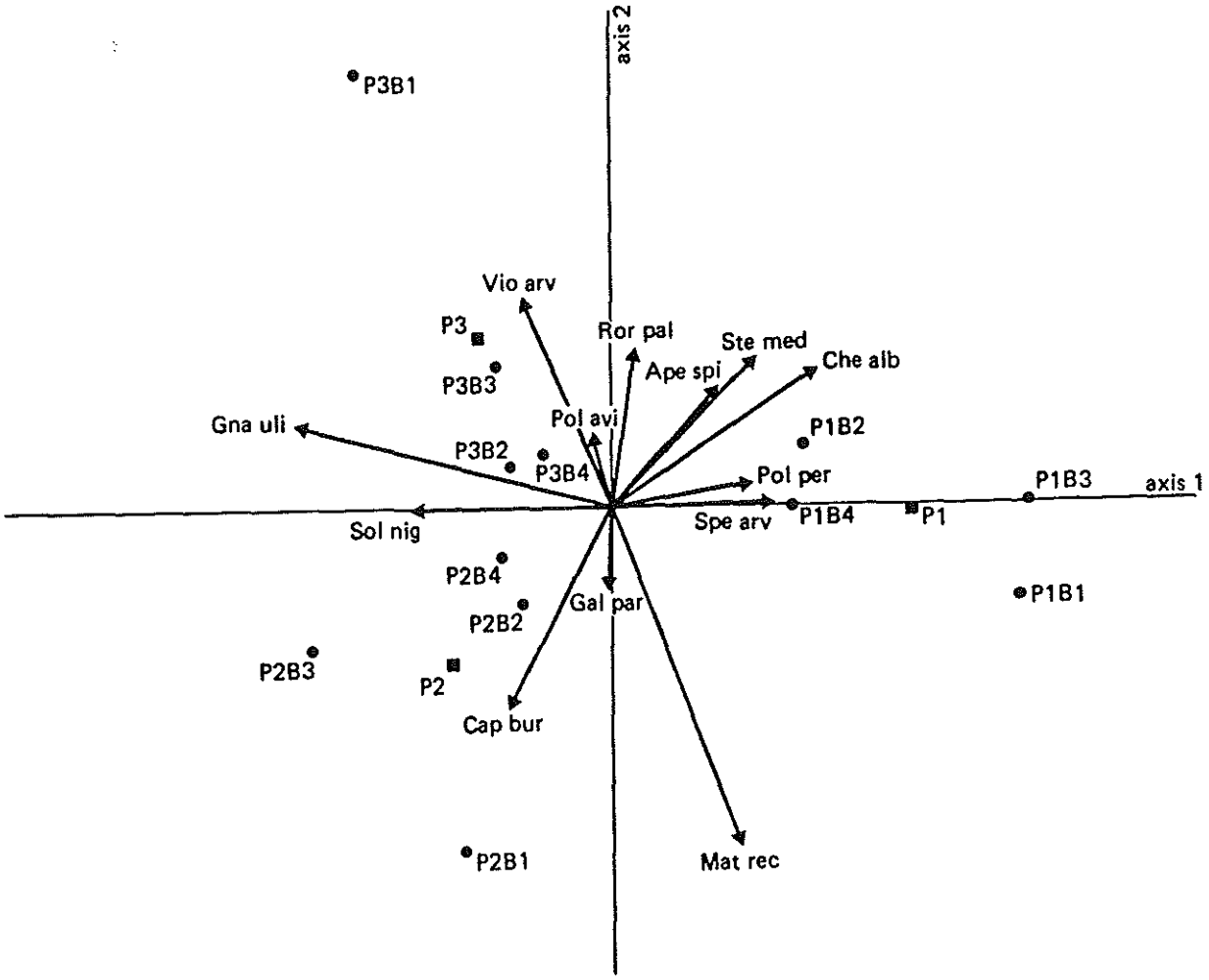

Fig.6.2 Ordination diagram based on partial redundancy analysis of the 1983 weed composition data with respect to ploughing times (P1 = March 9, P2 = March 23, P3 = April 6; data from Post, 1986), the covariables being formed by the blocks of the experiment (B1, B2, B3, B4). The diagram shows species (arrows), plots (●) labeled by ploughing time and block number, and the centroids (■) of the ploughing times. Abbreviations of species names are:
Ape spi = Apera spica-venti; Cap bur = Capsella bursa pastoris; Che alb = Chenopodium album; Gal par = Galinsoga parviflora; Gna uli = Gnaphalium uliginosum; Mat rec = Matricaria recutita; Pol avi = Polygonum aviculare; Pol per = Polygonum persicaria; Ror pal = Rorippa palustris; Sol nig = Solanum nigrum; Spe arv = Spergula arvensis; Ste med = Stellaria media; Vio arv = Viola arvensis (λ_1 = 0.16, λ_2 = 0.06, scaling α = 1).

transformed counts (ln(y+1)) of 13 weed species in May 1983 were subjected to partial canonical correspondence analysis, as specified in section 4.11. The length of the first ordination axis was as low as 0.3 SD, so that use of a linear method appeared more appropriate. Fig. 6.2 shows the ordination diagram resulting from a partial redundancy analysis and visualizes, for example, that Chenopodium album and Spergula arvensis are most abundant after ploughing on March 9 (P1) whereas Capsella bursa pastoris is most abundant after ploughing on March 23 (P2).

A restricted Monte Carlo permutation test (Table 4.14) showed that the effect of ploughing time on weed composition was on the margin of significance (P = 0.05).

In spring 1984 the plots were all cultivated by rotary tillage on a single day (to obtain a more even distribution of seeds in the seed bank). The counts made thereafter in May 1984 were subjected to the same analysis as the 1983 counts (λ_1 = 0.069) but failed to show significant differences (P = 0.45, test on first eigenvalue). Apparently the one single date of rotary tillage cancelled the effects of the previous treatments and/or recruted the same seedlings from the field seed bank.

Experimental data are standardly analyzed by the analysis of variance. Because the interest was not directed to one particular weed species, multivariate analysis of variance is called for, yet cannot be used, because the number of response variables (13 species) is larger than the number of experimental units (12). Partial redundancy analysis combined with Monte Carlo permutation tests is an attractive alternative escaping the restriction on the number of response variables.

6.3 Gene frequency data

McKecknie et al. (1975) studied the genetic variability in 21 colonies of the butterfly Euphydryas editha in relation to 11 environmental variables (Fig. 6.3). The genetic data consist of the frequencies of different alleles at various loci. Manly (1985) used the Mantel test (Sokal, 1969) and multiple regression to analyse data from three loci (Pgm 6 alleles; Pgi, 8 alleles; Hk, 3 alleles). We use the same data as Manly (1985) in this example. The data are percentages; for example, the frequencies of the three alleles of Hk in the colony named AF are 37, 59 and 4% and in the colony SL 16, 84 and 0%. As the data contain many zero values, CCA is an appropriate technique for analyzing these data (see section 7.1). The ordination diagram based on CCA (Fig. 6.3) shows how the allele frequencies vary along the environmental variables. For example, Hk2 has its highest percentages at higher altitudes and lower latitudes than Hk1, and Pgi5 has its peak frequency at higher levels of precipitation that Pgi4 whereas Pgi6 is intermediate. The first eigenvalue (λ_1 = 0.16) is significant (P = 0.01, Monte Carlo test) and there is thus evidence for a statistical relation between allele frequency and the environmental variables. The effects of different environmental variables cannot be separated out, however, because of their high multicollinearity: is it altitude, the annual minimum or maximum temperature or the minimum temperature during the post-diapause that is responsible for the variation along their common dimension in the diagram? The example shows that CCA is not just a way of testing statistical significance as is done in the Mantel test but that it also gives a neat display of the relation being put to test.

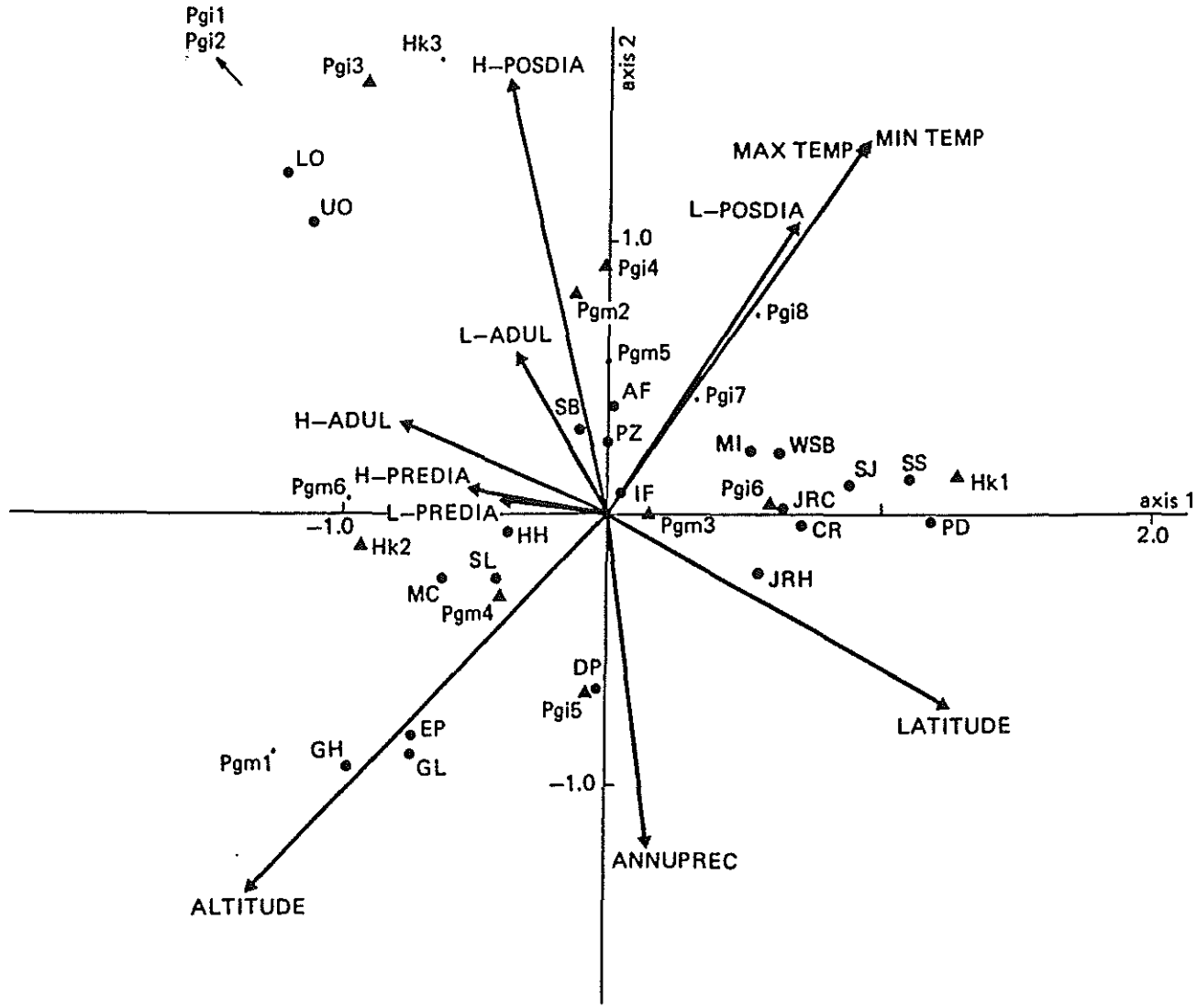

Fig.6.3 Ordination diagram based on canonical correspondence analysis of allele frequencies in the butterfly Euphydryas editha in 21 colonies (●) with respect to 11 environmental variables (data from McKecknie et al., 1975). The alleles (▲) from the three loci Pgm, Pgi and Hk, are numbered in order of increasing mobility class. Infrequent alleles are displayed by a dot. The environmental variables (arrows) are: altitude; latitude; annual precipitation (Annuprec), annual maximum and minimum temperature (Max-temp, Min-temp); highest and lowest temperature in the post-diapause (H-Posdia, L-Posdia), the adult phase (H-Adul, L-Adul) and the pre-diapause (H-Predia, L-Predia). The abbreviations of colony names follow Manly (1985). (λ_1 = 0.16, λ_2 = 0.07, scaling α = 1/2).

7. MISCELLANEOUS TOPICS

7.1 Percentage data/compositional data

Percentage data are obtained, for example, when for each sampling unit a fixed number of individuals is counted and each individual is identified to belong to one of m species. This sampling method is common in palynology and diatom research. Information from such a sample resides in the fraction of individuals belonging to each of the species. Percentage data also frequently arise in chemistry and geology where a sample is analyzed into its constituents (compositional data). In this section we present two alternative methods of analysis of such data. The first method, based on a series of papers by Aitchison (1982-4), amounts to applying linear methods to log-percentage data which are centred both by samples and by species (see also Aitchison, 1986). Because the logarithm of the percentages is analyzed, the method is attractive only when the data contain few zero values. The second method derives from a generalized linear model and amounts to applying weighted averaging methods to the untransformed percentage data. It is appropriate when the data contain many zeroes.

Percentage data containing no zeroes

Let p_{ik} be the fraction of species k in sample i ($\sum_{k=1}^{m} p_{ik} = 1$; $p_{ik} > 0$).

Because the fractions are positive, it is not acceptable to model them by a linear model such as

$$\eta_{ik} = a_k + b_k x_i + \varepsilon_{ik} \qquad (7.1)$$

because there is nothing to prevent the righthand side from resulting in a negative value. Equation (7.1) is the familiar straight line regression model with a_k the intercept, b_k the slope parameter or regression coefficient, x_i the value of an explanatory variable x and ε_{ik} an error term with mean zero and variance σ_k^2. The problem that the equation can result in negative values could be solved by modeling the fractions by $e^{\eta_{ik}}$, but then the model values still do not need to sum to 1. This problem is solved by dividing the $e^{\eta_{ik}}$ by their sum, yielding

$$p_{ik} = \frac{e^{\eta_{ik}}}{\sum_{j=1}^{m} e^{\eta_{ij}}} . \qquad (7.2)$$

The fractions $\{p_{ik}\}$ are said to follow a logistic normal distribution if the error ε_{ik} in (7.1) follows a normal distribution with mean 0 and covariance matrix $\sum$ (Aitchison, 1982: p. 162).

Now we have posed a model for fractions, we derive a method of analysis. Retracing the steps of the preceding argument, we take logarithms of fractions and obtain from (7.1) and (7.2):

$$\log p_{ik} = \gamma_i + a_k + b_k x_i + \varepsilon_{ik} \qquad (7.3)$$

where $\gamma_i = -\log(\Sigma_j e^{\eta_{ij}})$ is an incidental parameter. Interestingly, the incidental parameters $\{\gamma_i\}$ can be removed by centring the log-fractions by samples. When the data are also centred by species we obtain quantities y_{ik}, say,

$$y_{ik} = \tilde{p}_{ik} - \tilde{p}_{i.} - \tilde{p}_{.k} + \tilde{p}_{..} \qquad (7.4)$$

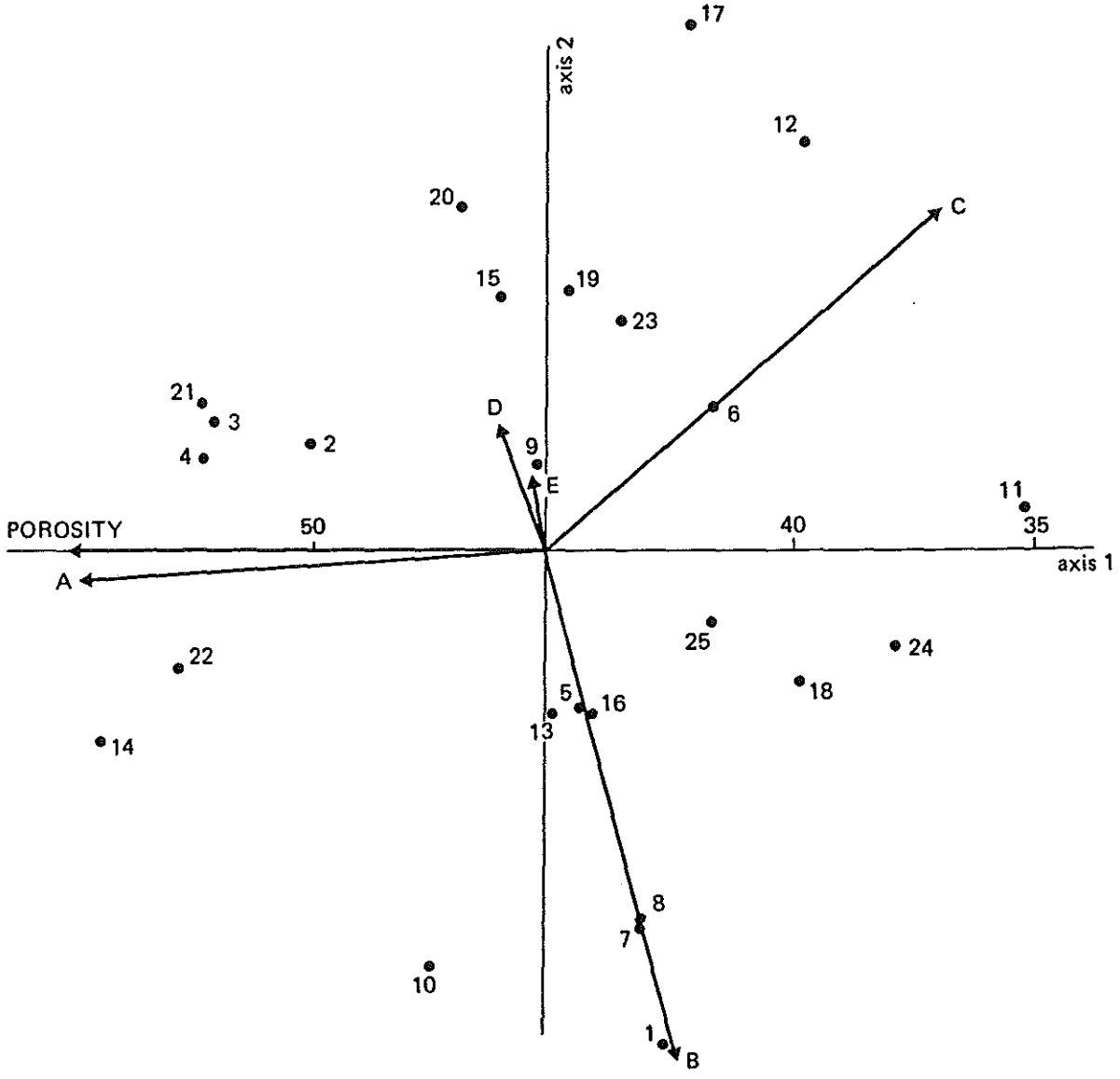

Fig.7.1 Ordination diagram based on a hybrid of redundancy analysis and principal components analysis on the geochemical composition of 25 samples (o) of "coxite" (artificial data from Aitchison, 1984a). The analysis is on log-percentages which are double centred and thereby overcomes the constant sum problem of percentage data. The first axis is constrained by porosity (the scale values shown are porosity values); the second axis displays residual variation in the composition of coxite, i.e. variation not explained by porosity. The five constituents of coxite (A, B, C, D, E,) are shown by arrows ($\lambda_1 = 0.39$, $\lambda_2 = 0.48$, scaling $\alpha = 1$).

where $\tilde{p}_{ik} = \log p_{ik}$ and a dot replacing an index denotes that the average is taken over the index. On inserting (7.3) in (7.4) we obtain a linear model for the quantities $\{y_{ik}\}$,

$$y_{ik} = b_k^* x_i^* + \varepsilon_{ik}^* \qquad (7.5)$$

where $b_k^* = b_k - b_.$, $x_i^* = x_i - x_.$ and $\varepsilon_{ik}^* = \varepsilon_{ik} - \varepsilon_{i.} - \varepsilon_{.k} + \varepsilon_{..}$ is still an error term with mean 0. Note that there is nothing in the above derivation which prevents us from using more than one explanatory variable in (7.1). Percentage data without zero values can thus by analyzed with the linear methods available in CANOCO by using the logtransformation and centring both by samples and by species. When using principal components analysis (PCA), one obtains what Aitchison (1984b: p. 622) calls loglinear-contrast principal components. Use of redundancy analysis (RDA) opens up the possibility of applying regression analysis to percentage data (cf. section 7.3). The Monte Carlo permutation test is then useful to test the effect of particular environmental variables.

As an example we use the boxite and coxite data sets presented by Aitchison (1984a: pp. 535-536) which each consist of the percentages of five chemical constituents in 25 samples of rock taken at different depths. We tested the hypothesis whether the chemical composition of the rock samples depends on depth. For the boxite and coxite data we obtain P-values of 0.59 and 0.21 respectively. Using a different test statistic, Aitchison (1984: p. 553) obtained P-values of ca. 0.35 and 0.001, respectively. There is a discrepancy for the coxite data which is caused by the large residual correlations among the constituents in these data. The type I error of both tests is the same, but the type II error of the test in CANOCO is larger than of Aitchison's test statistic. Aitchison's test which is based on the standard multivariate linear hypothesis is more powerful when there are large residual correlations. Although Aitchison's test detects that the composition of coxite is significantly related to depth, depth explains only 6% of the variance (because $\lambda_1 = 0.06$). The other variable given by Aitchison, porosity of the rock, is much more strongely related to composition: it explains 39% of the variability and is significant (P = 0.01 in 99 Monte Carlo permutations). Fig. 7.1 show the ordination diagram of RDA on porosity; the first axis ($\lambda_1 = 0.39$) displays the relation of composition with porosity; the second axis displays residual variation ($\lambda_2 = 0.48$). Porous rocks lie on the left hand side of the diagram and contain the largest percentages of constituent A. The least porous rocks lie on the right hand and contain the largest percentages of constituent C.

Finally we can note that Kooijman (1977) proposed transformation (7.4) to perform a Gaussian ordination of abundance data (not percentages) via linear methods.

Percentage data containing zeroes

Zero values present a problem in the preceding approach because the logarithm of 0 is $-\infty$. When the data contain few zeroes the problem may be

circumvented by replacing zeroes by an arbitrary small value, but this is unattractive when there are many zeroes because the result may depend considerably on the choice of the value replacing zero.

An alternative approach can be based on a generalized linear model (McCullagh and Nelder, 1983) for percentage data. Instead of defining a model for observed fractions as is done in (7.2) we define a model for expected fractions. Let y_{ik} from now on be the fraction of species k in sample i and Ey_{ik} the expected fraction. In analogy, with (7.1) and (7.2) we define the multinomial logit model (e.g. McCullagh and Nelder, 1983: p. 106; Anderson, 1984: p. 5)

$$Ey_{ik} = \frac{e^{\eta_{ik}}}{\sum_{j=1}^{m} e^{\eta_{ij}}} \qquad (7.6)$$

where η_{ik} is a linear predictor, e.g.

$$\eta_{ik} = a_k + b_k x_i . \qquad (7.7)$$

In comparison with (7.1), the error term has been dropped in (7.7). As McCullagh and Nelder (1983: p. 142) note, the regression coefficients in (7.7) can be estimated from data $\{y_{ik}\}$ and known $\{x_i\}$ by using standard computer packages by transforming (7.6) to a loglinear model (see below: (7.11)).

When both the $\{b_k\}$ and the $\{x_i\}$ are unknown, Eqs. (7.6) and (7.7) define an ordination model for percentage data. There are then two routes which show that approximate estimates of the unkown parameters can be obtained by applying correspondence analysis to the fractions $\{y_{ik}\}$.

The first route begins by rewriting (7.6) and (7.7) and using a first order Taylor approximation (Ihm and van Groenewoud, 1984: p. 49)

$$Ey_{ik} = \gamma_i^* \alpha_i^* e^{b_k x_i} \approx \gamma_i^* \alpha_k^* (1 + b_k x_i) \qquad (7.8)$$

where $\gamma_i^* = [\Sigma_j e^{\eta_{ij}}]^{-1}$ and $\alpha_k^* = e^{a_k}$. When for $\gamma_i^* a_k^*$ the simple estimate $y_{i+}y_{+k}/y_{++}$ is inserted we obtain (with y_{ik} replacing Ey_{ik})

$$y_{ik} \approx \frac{y_{i+}\, y_{+k}}{y_{++}} \; (1 + b_k x_i) \; . \qquad (7.9)$$

This is the reconstitution formula of correspondence analysis (Greenacre, 1984: p. 93; Ter Braak, 1985: p. 861). This similarity was noted also by Goodman (1981); the equality in (7.8) defines Goodman's RC-model.

The second route begins by noting that (7.6) does not change when (7.7) is replaced by (Fig. 7.2)

$$\eta_{ik} = a_k^* - \tfrac{1}{2} (x_i - u_k)^2 \qquad (7.10)$$

with $a_k^* = a_k + \frac{1}{2} u_k^2$ and $u_k = b_k$; the missing term in (7.7), $\frac{1}{2} x_i^2$, cancels

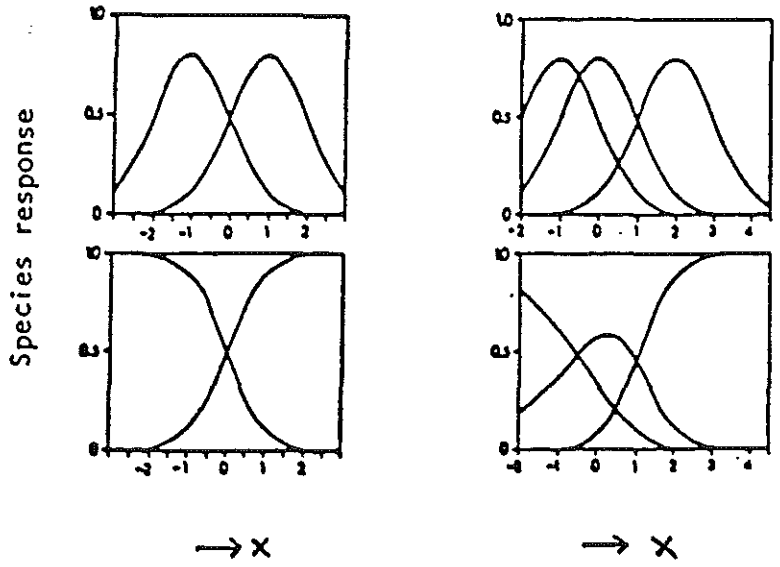

Fig.7.2 The top figures display the Gaussian model $Ey_{ik} = \exp(\eta_{ik})$ for the abundance of two and three species along a gradient x (m = 2, left; m = 3, right) and the bottom figures display the corresponding model for percentage data, i.e. the multiple logit model (7.11). (After Ihm and Van Groenewoud, 1984.)

out because it occurs in both the numerator and denominator of (7.6). Further, (7.6) and (7.10) can be written as the general loglinear model (McCullagh and Nelder, 1983: p. 106, p. 142)

$$\log Ey_{ik} = \gamma_i + a_k^* - \tfrac{1}{2}(x_i - u_k)^2 \qquad (7.11)$$

where $\gamma_i = -\log\left(\Sigma_j e^{\eta_{ij}}\right)$ is an incidental parameter (c.f. (7.3)). (We can take the logarithm of Ey_{ik} because Ey_{ik} is a positive value even if some of the $\{y_{ik}\}$ are 0). Except for the incidental parameter, model (7.11) was taken as the starting point by Ter Braak (1985) in showing that correspondence analysis gives under particular conditions an approximate solution to the fitting of a unimodal model by maximum likelihood. It turns out that the derivation carries through also with incidental parameter included. In calibration context (with known values of $\{a_k^*\}$ and $\{u_k\}$), the weighted averages (4.9) with $\alpha = 1$ are under the same conditions efficient estimators of the sample scores $\{x_i\}$; cf. Ter Braak and Barendregt (1986). Moreover, the derivation of canonical correspondence analysis given in Ter Braak (1986a) carries through equally for percentage data using model (7.6)

and (7.10) and the constraint that x_i is a linear combination of environmental variables. Note that (7.6) and (7.10) together define Ihm and Van Groenewoud's model B.

The model can be extended to two-dimensions by taking

$$\eta_{ik} = a_k^* - \tfrac{1}{2}[(x_{i1} - u_{k1})^2 + (x_{i2} - u_{k2})^2] \ . \qquad (7.11)$$

Again correspondence analysis can be used to obtain approximate estimates, except that detrending may be required to remove the arch effect when it occurs (section 3.4).

Paradox

It is rather paradoxical that the generalized linear model (7.6) with (7.7) is identical to the unimodal model (7.6) with (7.10). This apparent paradox was already noted to occur in correspondence analysis in the discussion of Ter Braak (1985). It entails that for percentage data an ordination diagram of correspondence analysis can equally well be interpreted by the rules of a biplot as by the rules of a joint plot (Jongman et al. 1987, sections 5.2.5 and 5.3.4). The problem is of course that to infer from the plot the percentage value of a particular species in a sample, the values of all remaining species data are required also.

In section 3.4 (question Q14) it was noted that there is some arbitrariness in correspondence analysis of how to scale the sample scores $\{x_i\}$ with respect to the species scores $\{u_i\}$. The scaling is governed by the value of α in the sections 4.5 and 4.6. For incidence and abundance data there exists a best fitting value of α (which is unfortunately unknown in general). But for percentage data α is completely arbitrary: in (7.6) with (7.10) the model does not change if we take βu_k, x_i/β and $a_k^* + \tfrac{1}{2}(\beta-1)u_k^2$ instead of u_k, x_i and a_k^*, respectively. Moreover, the optima $\{u_k\}$ may be shifted arbitrarily with respect to the sample scores $\{x_i\}$; with a shift from u_k to $u_k + d$ for a constant d just change also the value of a_k^* to $a_k^* + du_k$.

These puzzling properties appear to be inherent to percentage data; they are shared by both the linear approach and the weighted averaging approach presented in this section. In contrast, the location and scale of the optima $\{a_k\}$ are well determined for abundance data and incidence data (cf. Kooijman, 1977).

7.2 Nominal response data

Outside ecology, correspondence analysis is most frequently applied to nominal data (Gifi, 1981, Greenacre, 1984). Nominal data arise when each response variable consists of a series of mutually exclusive categories or classes. For example, vegetation type and soil type are nominal variables. Correspondence analysis can be applied to nominal variables to investigate their interrelations. It is termed multiple correspondence analysis (Greenacre, 1984) or homogeneity analysis (Gifi, 1981) if there are more than

two such variables. When interest focusses on how the nominal response variables depend on external explanatory variables, one can use canonical correspondence analysis. Applied to such data, it is equivalent to redundancy analysis of qualitative variables (Israëls, 1984). Torgerson (1958: p. 338) already described the types of data which can be analysed by what is now called correspondence analysis. In biology, nominal data are encountered frequently in numerical taxonomy and genetics.

To analyze nominal response variables with CANOCO, each nominal variable must be represented by a series of dummy variables each representing a category: y_{ik} = 1 or 0 depending on whether sampling unit i belongs or does not belong to category k (cf. section 2.1). Each category is thus a species in the terminology of CANOCO and each individual a sample. For the analysis by CANOCO the categories of different nominal variables must be assigned different numbers. One can number them consecutively from 1 to m with m the total number of categories. Alternatively, if the maximum number of categories per variable is less than 10, one can reserve the number 11-19 to the categories of nominal variable 1, the numbers 21-29 to the categories of nominal variable 2, etcetera. Nominal response data are best supplied to CANOCO in Cornell condensed format (sections 2.1 and 2.3). If one has, as in Table 2.5, 3 nominal variables one needs to specify 3 couplets per sampling unit (= individual). With nominal data, the species scores are category quantifications in the sense of Gifi (1981). For each nominal variable, the weighted mean of its category scores is equal to 0.

The theory of section 7.1 can be applied to nominal data. With such data, (7.6) models the probability that sample i belongs to category k. In applying multiple correspondence analysis conditional independence is assumed, i.e. the joint probability that a sample belongs to the categories k_1, k_2, k_3, ... of nominal variables 1, 2, 3, ... is simply the product of each of the category probabilities given by (7.6). The logarithm of the joint probability can therefore be expressed as $\Sigma_l \phi_{il} + \eta_{ik(l)}$ where l indexes the nominal variables. By modeling $\eta_{ik(l)}$ by (7.10) and using the approach of Ter Braak (1985) we obtain an alternative derivation of multiple correspondence analysis. This approach shows that the category quantification can equally well be considered optima of response curves (Fig. 7.2) with respect to the ordination axes.

In some applications a sampling unit may belong partly to one category and partly to another one. For such data, fuzzy coding has been proposed: for example, the sample is assigned the value 0.5 for both categories (see Greenacre, 1984: p. 159, codage flou in French). Obviously fuzzy coding is allowed in CANOCO. Percentage data are of this form (sections 6.3 and 7.1).

If one has aggregated data on nominal response variables, the data become the number of individuals belonging to each category. This type of data presents no problem for CANOCO. It is similar to abundance data which list, for example, the number of organisms belonging to each of m species.

If the Guttman effect (= the arch effect) crops up, detrending-by-polynomials is appropriate to remove it. The simple explanation of the Guttman effect given by Jongman et al. (1987: section 5.2.3) applies equally well to nominal data.

7.3 Multiple regression, redundancy analysis, principal components analysis and canonical correlation analysis

Redundancy analysis (RDA) can be expressed as a constrained form of multiple regression of the species' responses on the explanatory (environmental) variables. This type of regression is called reduced rank regression (Davies and Tso, 1982). Estimates of the regression coefficients can be obtained from the CANOCO output.

Let y_{ik} be the response value of species k in sample i (i = 1, ..., n; k = 1, ..., m), let $\tilde{z}_{ij}$ be the value of environmental variable j in sample i (j = 1, ..., q) and assume for convenience of notation that the environmental variables are standardized to mean 0 and variance 1 as in equation (4.12). The usual model of a multiple regession of the responses of species k on the q environmental variables is

$$y_{ik} = a_{k0} + \sum_{j=1}^{q} a_{kj} \tilde{z}_{ij} + \varepsilon_{ik} \qquad (7.12)$$

where a_{k0} is the intercept, a_{kj} the regression coefficient of environmental variable j of the regression for species k and ε_{ik} an error term with mean 0. Model (7.12) represents m separate multiple regressions, one for each species, on the q environmental variables. We want to estimate the coefficients $\{a_{kj}\}$ by way of redundancy analysis. Because the environmental variables are standardized to zero mean, the estimate of a_{k0} is equal to $y_{.k}$, the mean value of the k-th species variable.

The rank of reduced rank regression is the number of ordination axes used to estimate the regression coefficients. For convenience of notation, we take rank = 2; the general case follows without problems. With two ordination axes, the model of redundancy analysis can be written as

$$y_{ik} = y_{.k} + \tau(b_{k1}x_{i1} + b_{k2}x_{i2}) + \varepsilon_{ik} \qquad (7.13)$$

$$x_{i1} = \sum_{j=1}^{q} c_{j1}\tilde{z}_{ij} \text{ and } x_{i2} = \sum_{j=1}^{q} c_{q2}\tilde{z}_{ij} \qquad (7.14)$$

where b_{k1} and b_{k2} are the scores of species k on axis 1 and axis 2 (section 4.5), x_{i1} and x_{i2} are the scores of sample i which are linear combinations of the environmental variables (section 4.7), c_{j1} and c_{j2} are the canonical coefficient of environmental variable j for axis 1 and axis 2 (section 4.7, confer (4.15) where $c_0 = 0$ in RDA) and τ is a constant to be defined below. On inserting (7.14) in (7.13) we obtain after rearranging terms

$$y_{ik} = y_{.k} + \tau \sum_{j=1}^{q} (b_{k1}c_{j1} + b_{k2}c_{j2}) \tilde{z}_{ij} + \varepsilon_{ik}. \qquad (7.15)$$

The redundancy analysis model implies therefore a model for the regression coefficients a_{kj} in (7.12), namely the "rank 2 model"

$$a_{kj} = \tau(b_{k1}c_{j1} + b_{k2}c_{j2}). \qquad (7.16)$$

Estimates of the regression coefficients $\{a_{kj}\}$ can thus be obtained from estimates of the species scores $\{b_{k1}, b_{k2}\}$ and the canonical coefficients $\{c_{j1}, c_{j2}\}$ given by CANOCO (sections 4.5 and 4.7). It are "rank 2 estimates". The constant τ is $n^{-1}\sqrt{TSS}$ where TSS is the Total Sum of Squares in the species data given in the output (see section 3.6, question Q27). When samples and species have equal weight (in Q27: $w_i = 1$, $w_k = 1$), then, in the present notation, $\tau = n^{-1}\{\Sigma_{i,k} (y_{ik}-y_{.k})^2\}^{\frac{1}{2}}$. [The constant τ arises from the particular scaling of the species data (see Q27), of the species scores (4.5) and of the sample scores (4.10)].

Users who wish to calculate the regression coefficients for unstandardized environmental variables, can use the procedure explained in section 4.7 above equation (4.16). If there is just a single response variable (m = 1) and the scaling Q15 = 1 is chosen, then the canonical coefficients are actually the regression coefficients of the regression of a standardized response variable on standardized explanatory variables. The eigenvalue is then simply R^2, the squared multiple correlation between the response variable and the explanatory variables.

The above formulae give insight in the properties of the ordination diagram of RDA. The sample and species scores in (7.13) can be used to construct a biplot (Gabriel, 1971, Jongman et al. 1987: section 5.3.4). As follows from (7.13), the fitted values of the regression can be inferred from the biplot (of course, up to the proportionality constant τ and the means $\{y_{.k}\}$). The RDA biplot gives a least-squares approximation of the fitted values of (7.12) (Davies and Tso, 1982). By plotting the sample score $\{x_i^*\}$ (section 4.6) instead of $\{x_i\}$ (section 4.7) one obtains a biplot which attempts to approximate the observed values rather than the fitted values of model (7.12). A biplot which gives a still better approximation to the observed values can be obtained by principal components analysis (PCA). The PCA model with two ordination axes is equation (7.13) with x_i^* replacing x_i. PCA is thus RDA without the constraint (7.14) on the sample scores. The PCA biplot gives a least squares approximation of the observed values $\{y_{ik}\}$.

The standard RDA biplot contains arrows for environmental variables based on the biplot scores of environmental variables (section 4.9). Together with the species points they allow inference of the covariances between species and environmental variables (section 4.4). Equation (7.16) suggests another biplot. By plotting the canonical coefficients and the species scores, one obtains a biplot which approximates the regression coefficients of the environmental variables for each of the species. This plot displays the partial effects of each environmental variable taking into account the effect of the other variables, whereas the standard RDA biplot displays "marginal" effects. The partial effect is the effect of the variable with the other variables being held constant, whereas the marginal effect is the effect of the variable with the other variables covarying in the particular way they do in the dataset. This distinction parallels the distinction between canonical coefficients and inter-set correlations (Ter Braak, 1986a). The biplot of regression coefficients is useless when the regression

coefficients are unstable due to multicollinearity between the environmental variables. The standard RDA biplot is not hampered by multicollinearity between the environmental variables.

Tso (1981) showed that canonical correlation analysis is also a technique for reduced rank regression. The difference with redundancy analysis lies in the assumptions about the error term in (7.13). If the errors follow a multivariate normal distribution, both techniques are maximum likelihood techniques but under different assumptions. In redundancy analysis, being a least-squares technique, it is assumed that the errors are uncorrelated and have the same variance, i.e. that the covariance matrix of the errors is of the form $\sigma^2 I$. In canonical correlation analysis no assumptions are made about the covariance matrix of the errors, except that the matrix is nonsingular.

In canonical correlation analysis the covariance matrix of the errors must therefore be estimated from the data. Because the estimated covariance matrix must be nonsingular, the number of samples in canonical correlation analysis must be greater than m+q+1 where m is the number of species and q the number of environmental variables. This sets a restriction on the number of species compared to the number of samples that can be analysed by canonical correlation analysis. In redundancy analysis there is no such restriction because there is only one parameter to estimate in the covariance matrix, namely σ^2. In most studies in community ecology the number of samples is small compared to the number of species. This makes canonical correlation analysis unattractive for such studies.

7.4 Principal coordinates analysis (PCO)

Principal coordinates analysis, alias classical scaling (Gower, 1966; Torgerson, 1958: p. 254-259; Jongman et al., 1987: section 5.6) is a simple method for multidimensional scaling. It takes as input a table of dissimilarities or similarities between samples and derives from it a sample ordination. In the ordination diagram the sample points are arranged such a way that sample points which are close together correspond to samples that are similar, and samples which are far apart correspond to samples that are dissimilar.

A principal coordinate analysis (PCO) can be obtained with CANOCO by taking as "species data" a square table of similarities or a square table with elements $-\delta_{ij}$ where δ_{ij} is the dissimilarity between sample i and sample j ($i = 1, \ldots, n$; $j = 1, \ldots, n$). In the "species data" there are thus as many species as there are samples, the j-th species corresponding to the j-th sample. The further specifications are:

Q2 = 1 - principal components analysis (PCA)
Q15 = 3 - symmetric scaling
Q26 = 1 - centring by samples
Q27 = 1 - centring by species

If one has a data file with the values of δ_{ij}, one can use the transformation of Q24 to obtain the values of $-\delta_{ij}$ by specifying 0 0

followed on the next lines: by 100 -100 and -1 0 (assuming that all δ_{ij} are smaller than 100).

If the input dissimilarities $\{\delta_{ij}\}$ are in fact computed as squared Euclidean distances from species data, the resulting sample scores are identical to a PCA applied to the species data using centring by species and the scaling of a Euclidean distance biplot (Q15 = 1). Principal coordinate analysis is based on this similarity to PCA, but is more general, because one can use other measures of (dis)similarity than Euclidean distance.

Unfortunately CANOCO cannot be used to obtain the solution of a constrained PCO. When choosing the RDA-option (Q2 = 2) in the above setting, CANOCO solves the wrong eigenvalue equation (Appendix B). Partial PCO is not available either in CANOCO.

7.5 Interchanging species and samples; weighted averaging ordination

Frequently ecologists wish to interpret ordination diagrams by using external data on the species, for example indicator values which characterize their habitat requirements (Persson, 1981). Such data can be used in CANOCO either by deriving sample values from such data by calibration (Jongman et al., 1985; chapter 4) or by interchanging species and samples in the input data files. Because weighted averaging methods treat species and samples in a symmetric way (unless detrending is in force) there is nothing against an interchange. In linear methods this is possible too, when the user takes care in choosing the options for centring and standardization (Q26 and Q27) so as to leave the analysis unaffected. In this way one can also obtain ordinations which are constrained by linear combinations of species properties.*) With one or two such properties, this is a way to obtain weighted averaging ordinations of samples (Gauch, 1982; Jongman et al., 1987: section 4.3) by using CCA.

Weighted averaging ordinations of species can be obtained by a CCA with just one or two environmental variables (Ter Braak, 1986a).

7.6 Weighting samples and species

If environmental data are to hand, the standard way to "focus" the ordination on particular gradients or particular research questions is, of course, to include such data in a direct gradient analysis. If no such data are to hand, the ecologist can still focus the analysis on particular gradients by assigning large weights to samples and species that are

---

*) Footnote: In constrained linear analyses the species scores will be centred automatically because it are CANOCO-sample-scores (Table 3.2 and section 4.6). This may be undesirable as it implies centring by samples of the data in the original unconstrained analysis (without interchanging species and samples).

considered as extremes on the gradients of interest. This option can have a similar effect as an expert choice of end points of ordination axes in a polar ordination (Gauch, 1982).

7.7 Calibration by CANOCO

CANOCO is not designed for calibration, yet can be used for it. The disadvantage is that the output requires post-processing. Calibration in CANOCO can proceed either by inverse regression or by constrained ordination. In both cases one needs a training set of species data with corresponding values of the variable to be calibrated. With inverse regression CANOCO can be used to obtain the transfer function, but not to obtain inferred values.

In inverse regression, the calibration variable is taken as the response variable and the species are taken as explanatory variables. The resulting regression equation is the transfer function.
The name "inverse regression" typifies quite well how to obtain the transfer function in CANOCO: the variable to be calibrated must be entered as species data (Q3) and the species data as environmental data (Q5)! By asking for a RDA and Q15 = 1, canonical coefficients are obtained which are the coefficients of the regression of the standardized calibration variable on standardized species variables. Unstandardized coefficients can be obtained in a similar way as in (4.16). Unthinking use of inverse regression is particularly dangerous when the species variables show high multicollinearity (section 4.3). In general, the multiple correlation coefficient (square root of first eigenvalue) gives an overoptimistic impression of the precision when using the transfer function.

Calibration by constrained ordination proceeds by taking the calibration variable as the only environmental variable. All samples with known species composition are included in the species data, and the samples with known value of the calibration variable are included in the environmental data. Samples of which the value of the calibration variable is to be inferred, will therefore be passive samples in CANOCO. One can choose between RDA and CCA. We limit the discussion here to the CCA case. See Ter Braak and Prentice (1987) for the RDA case and Naes et al. (1986) for a related linear method.

The CCA case reduces to two-way weighted averaging followed by simple linear regression. On using the notation of section 4.7, let z_{i1} be the value of the calibration variable of sample i of the training set. A simple method to estimate the optimum of species k with respect to the calibration variable is by taking the weighted average

$$u_k = \sum_{i=1}^{n} y_{ik} z_{i1} / \sum_{i=1}^{n} y_{ik} . \qquad (7.17)$$

Similarly, the value of the calibration variable can be inferred by taking the weighted average, i.e. the (preliminary) estimate is

$$z_{i1}^{*} = \sum_{k=1}^{m} y_{ik} u_k / \sum_{k=1}^{m} y_{ik}. \qquad (7.18)$$

Because averages are taking twice, the range of the calibration variable is shrunken. To undo this and to obtain a best fitting equation we regress the inferred values for the training set on the observed values using the model

$$z_{i1}^{*} = c_0^{*} + c_1^{*} z_{i1} + \varepsilon_i^{*} . \qquad (7.19)$$

Inverting (7.19), we obtain the estimate

$$\text{est}(z_{i1}) = \frac{z_{i1}^{*} - c_0^{*}}{c_1^{*}} . \qquad (7.20)$$

As follows from the similarity between (4.4), (4.9) and (4.13) with (7.17), (7.18) and (7.19), the estimate in (7.20) can be obtained from the CANOCO output by the formula

$$\text{est}(z_{i1}) = \bar{z}_1 + \frac{s_1 x_i^{*}}{c_1} \qquad (7.21)$$

where $\bar{z}_1$ and s_1 are the mean and standard deviation of the calibration variable (given by CANOCO; see Table 4.1), c_1 is the canonical coefficient for the standardized calibration variable (see Table 4.6) and x_i^{*} is the sample score on the first axis. The score x_i^{*} is given by CANOCO for both active and passive samples (section 4.6). Equation (7.21) can thus be used to infer missing values of the calibration variable.

The correlation between $\text{est}(z_{i1})$ and z_{i1} in the training set is given by the species-environment correlation of the first axis. For the calibration to be useful, the first constrained eigenvalue should be large compared to the other eigenvalues, because only then the calibration variable is the dominant variable determining the species composition.

Gasse and Tekaia (1983) proposed a similar procedure, the main difference being that the calibration variable is first divided in classes and CCA is applied with respect to the resulting nominal variable. They obtain the transfer function by regressing the calibration variable on the sample scores $\{x_i^{*}\}$ on a number of axes (not only the first axis).

Ter Braak and Van Dam (in prep.) applied these procedures to infer past-pH values in soft water lakes and pools from diatom assemblages and compared them with a more formal statistical procedure based on maximum likelihood.

7.8 Canonical variates analysis (CVA)

A canonical variates analysis (CVA), alias Fisher's linear discriminant analysis, can be obtained with CANOCO, because CCA is a generalization of CVA (Chessel et al. 1987, Lebreton et al. 1988). Suppose you want a CVA to see which linear combinations of environmental variables discriminate best between clusters of samples, e.g. obtained by a cluster analysis on species data. For this, specify the clusters as dummy variables in a file, for example CLUSTERS.DAT. This is easily done in condensed format (see Table 2.5).

A CVA is obtained by

- asking for a CCA at Q2
- entering CLUSTERS.DAT as species data at Q3
- entering the environmental data at Q5
- asking for "species scores which are weighted mean sample scores" at Q14 (Q14=2).

The species scores are the cluster means in the CVA ordination diagram.

The sample scores that are linear combinations of environmental variables are the individual points in the diagram.

The biplot scores for the environmental variables form with the species scores a biplot of the cluster means of each of the environmental variables (a weighted least-squares approximation) and form with the individual points a biplot of the environmental data (a least-squares approximation with the individual points given a priori). The percentage variance accounted for by the species-environment biplot reported by CANOCO is, however, not the usual percentage reported for a CVA. See section 9.5.

The sample scores that are linear combinations of environmental variables are scaled so that the within-cluster variance equals 1 (in this variance the divisor is n and not n-g with g the number of clusters). See equation (4.7).

The eigenvalues reported by CANOCO are those of the eigenvalue equation:

$$(B - \lambda T)\, \underline{c} = \underline{0} \qquad (7.22)$$

where λ is de eigenvalue, $\underline{c}$ the vector of canonical coefficients (loadings), B the matrix of between-cluster sums of squares and products and T the matrix of total sums of squares and products.

The permutation test can be used to see whether the difference between clusters are statistically significant. This test has the advantage over the usual tests in CVA in that it does not require the assumption that the environmental variables are normally distributed.

By specifying covariables a partial CVA is obtained. Partial CVA is also known as one-way Multivariate ANalysis of COvariance (MANOCO). This tests for discrimination between clusters in addition to the discrimination obtainable with the covariables.

8. ITERATIVE ORDINATION ALGORITHM

In CANOCO a general iterative algorithm is used to solve the transition formulae of the linear and weighted-averaging methods described in the sections 3 and 4. Its essential features are described here. In the description, user-defined weights for sites and species (section 3.6: Q28-Q32) are excluded for clarity of exposition; they are set equal to 1. The algorithm operates on response variables, each recording the value, abundance or presence/absence of a species at various sites, and on two types of explanatory variables: environmental variables and covariables. By environmental variables we mean here explanatory variables of prime interest, in contrast with covariables which are "concomitant" variables whose effect is to be removed. When all three types of variables are present, the algorithm describes how to obtain a partial constrained ordination. The other linear and weighted averaging techniques are all special cases, obtained by omitting various irrelevant steps.

Let $Y = [y_{ik}]$ ($i = 1, \ldots, n$; $k = 1, \ldots, m$) be a site-by-species matrix containing the observations of m species at n sites and let $Z_1 = [z_{1il}]$ ($i = 1, \ldots, n$; $l = 0, \ldots, p$) and $Z_2 = [z_{2ij}]$ ($i = 1, \ldots, n$; $j = 1, \ldots, q$) be site-by-covariable and site-by-environmental variable matrices containing the observations of p covariables and q environmental covariables at the same n sites, respectively. The observations z_{1il} and z_{2ij} may take any real value. The observation y_{ik} may take any real value in linear methods but must be greater than or equal to 0 in weighted averaging methods. Further, denote the species and site scores on the s-th ordination by $\underline{u} = [u_k]$ ($k = 1, \ldots, m$) and $\underline{x} = [x_i]$ ($i = 1, \ldots, n$), the canonical coefficients of the environmental variables by $\underline{c} = [c_j]$ ($j = 1, \ldots, q$) and collect the site scores on the $(s - 1)$ previous ordination axes as columns of the matrix A. If detrending-by-polynomials is in force (Step A10), then the number of columns of A, s_A say, is greater than s-1. In the algorithm we use the assign statement " := ", for example a := b means "a is assigned the value b". If the left hand side of the assignment is indexed by a subscript, it is assumed that the assignment is made for all permitted subscript values: the subscript k will refer to species ($k = 1, \ldots, m$), the subscript i to sites ($i = 1, \ldots, n$) and the subscript j to environmental variables ($j = 1, \ldots, q$).

Preliminary calculations

P1. Calculate species totals $\{y_{+k}\}$, site totals $\{y_{i+}\}$ and the grand total y_{++}. If a linear method is required, set

$$r_k := 1, \; w_i := 1, \; w_i^* := \frac{1}{n} \tag{8.1}$$

and if a weighted averaging method is required, set

$$r_k := y_{+k}, \; w_i := y_{i+}, \; w_i^* := y_{i+}/y_{++} \tag{8.2}$$

P2. Standardize the environmental variables and covariables to zero mean and unit variance, e.g. for environmental variable j calculate its mean $\bar{z}$ and variance v

$$\bar{z} := \sum_i w_i^* z_{2ij}, \quad v := \sum_i w_i^* (z_{2ij} - \bar{z})^2 \tag{8.3}$$

and set $z_{2ij} := (z_{2ij} - \bar{z})/\sqrt{v}$

P3. Calculate for each environmental variable j the residuals of the multiple regression of the environmental variable on the covariables, i.e.

$$\underline{c}_j^* := (Z_1' W Z_1)^{-1} Z_1' W \underline{z}_{2j} \tag{8.4}$$

$$\tilde{\underline{z}}_{2j} := \underline{z}_{2j} - Z_1 \underline{c}_j^* \tag{8.5}$$

where $\underline{z}_{2j} = (z_{21j}, \ldots, z_{2nj})'$, $W = \text{diag}\,(w_1, \ldots, w_n)$ and $\underline{c}_j^*$ is the p-vector of the coefficients of the regression of $\underline{z}_{2j}$ on Z_1. Now define $\tilde{Z}_2 = [\tilde{z}_{2ij}]$ $(i = 1, \ldots, n,\ i = j, \ldots, q)$.

Iteration algorithm

Step A0 Start with arbitrary, but unequal site scores $\underline{x} = [x_i]$. Set $x_i^0 = x_i$.

Step A1 Derive new species scores from the site scores by

$$u_k := \sum_i y_{ik} x_i / r_k. \tag{8.6}$$

Step A2 Derive new site scores $\underline{x}^* = [x_i^*]$ from the species scores

$$x_i^* := \sum_k y_{ik} u_k / w_i. \tag{8.7}$$

Step A3 Make $\underline{x}^* = [x_i^*]$ uncorrelated with the covariables by calculating the residuals of the multiple regression of $\underline{x}^*$ on Z_1:

$$\underline{x}^* := \underline{x}^* - Z_1 (Z_1' W Z_1)^{-1} Z_1' W \underline{x}^*. \tag{8.8}$$

Step A4 If $q \leq s_A$, set $x_i := x_i^*$ and skip Step A5.

Step A5 If $q > s_A$, calculate a multiple regression of $\underline{x}^*$ on $\tilde{Z}_2$

$$\underline{c} := (\tilde{Z}_2' W \tilde{Z}_2)^{-1} \tilde{Z}_2' W \underline{x}^*, \tag{8.9}$$

and take as new site scores the fitted values:

$$\underline{x} := \tilde{Z}_2 \underline{\varrho}. \tag{8.10}$$

Step A6 If $s > 0$, make $\underline{x} = [x_i]$ uncorrelated with previous axes by calculating the residuals of the multiple regression of $\underline{x}$ on A:

$$\underline{x} := \underline{x} - A\,(A'WA)^{-1}A'W\underline{x}. \tag{8.11}$$

Step A7 Standardize $\underline{x} = [x_i]$ to zero mean and unit variance by

$$\bar{x} := \sum_i w_i^* x_i, \quad s^2 := \sum_i w_i^* (x_i - \bar{x})^2, \tag{8.12}$$

$$x_i := (x_i - \bar{x})/s.$$

Step A8 Check convergence, i.e. if

$$\sum_i w_i^* (x_i^0 - x_i)^2 < 10^{-10} \tag{8.13}$$

goto Step A9, else set $x_i^0 := x_i$ and goto Step A1.

Step A9 Set the eigenvalue λ equal to s in (8.12) and add $\underline{x} = [x_i]$ as a new row to the matrix A.

Step A10 If detrending-by-polynomials is required, calculate polynomials of $\underline{x}$ up to order 4 and first-order polynomials of $\underline{x}$ with the previous ordination axes,

$$x_{i2} := x_i^2, \; x_{i3} := x_i^3, \; x_{i4} := x_i^4, \; x_{i(b)} := x_i a_{ib} \tag{8.14}$$

where a_{ib} are the site scores of a previous ordination axis ($b = 1, \ldots, s-1$). Now perform for each of the (s+2)-variables in (8.14) the Steps A3-A6 and add the resulting variables as new variables to the matrix A.

Step A11 Set $s := s+1$ and goto Step A0 if required and if further ordination axes can be extracted, else stop.

At convergence, the algorithm gives the solution with the greatest real value of λ to the following transition formulae [where $R = \mathrm{diag}(r_1, \ldots, r_m)$ and $W = \mathrm{diag}(w_1, \ldots, w_n)$ and where the notation B^0 is used to denote $B(B'WB)^{-1}B'W$, the projection operator on the column space of a matrix B in the metric defined by the matrix W]

$$\underline{u} = \lambda^{-\alpha} R^{-1} Y' \underline{x} \tag{8.15}$$

$$\underline{x}^* = \lambda^{\alpha-1} (I - Z_1^0) W^{-1} Y \underline{u} \tag{8.16}$$

$$\underline{c} = (\tilde{Z}_2' W \tilde{Z}_2)^{-1} \tilde{Z}_2' W \underline{x}^* \tag{8.17}$$

$$\underline{x} = (I - A^0) \tilde{Z}_2 \underline{c}. \tag{8.18}$$

(It follows from Appendix equation (A.5) that there is some freedom of choice how to distribute the eigenvalue λ over these equations, e.g. Ter Braak and Prentice (1987) assigned the eigenvalue to (8.18).) The wiggle above Z_2 is there as a reminder that the original environmental variables were replaced by residuals of a regression on Z_1 in (8.5) i.e. in terms of the original variables

$$\tilde{Z}_2 = (I - Z_1^0) Z_2 \tag{8.19}$$

Remarks

1. Note that u_k in the algorithm takes the place of b_k in section 4.

2. Special cases of the algorithm are: constrained ordination: p = 0; partial ordination: q = 0; (unconstrained) ordination: p = 0, q = 0; linear calibration and weighted averaging: p = 0, q = 1; (partial) multiple regression: m = 1. The corresponding transition formulae follow from (8.15) - (8.18) with the proviso that, if q = 0, Z_2 in (8.19) must be replaced by the n×n identity matrix and generalized matrix inverses are used.

3. The centring of $\underline{x}$ in step A7 and the centring of Z_1 in step P2 make redundant the centring-by-species of the species data in the linear methods. If centring-by-species is not required (for example in noncentred PCA), then the sample scores $\underline{x}$ in step A7 are not centred, i.e. $\bar{x}$ in (8.12) is set equal to 0. If centring-by-samples is required, then this is accomplished by centring the new species scores obtained in step A1. This centring makes redundant the centring-by-samples of the species data. In CANOCO the centred values $\tilde{y}_{ik}$ in question Q26-Q27 (section 3.6) are computed for the calculation of the total sum of squares (TSS), but they are not stored. This is the crux of the efficiency of CANOCO for community ordination: only the non-zero values in the species data Y need to be stored.

4. The standardization in P2 removes the arbitrariness in the units of measurement of the environmental variables, and makes the canonical coefficients comparable among each other, but does not influence the values of λ, $\underline{u}$ and $\underline{x}$ to be obtained in the algorithm.

5. The Step A3 and A6 simplify considerably if the covariables and the columns of A are to W-orthonormal variables, i.e. when they are made uncorrelated. This is done in CANOCO. The steps A3-A6 form a single projection of $\underline{x}^*$ on the column space of $(I - A^0)\bar{Z}_2$ if and only if A defines a subspace of $\bar{Z}_2$. As each ordination axis defines a subspace of $\bar{Z}_2$, this is trivially so without detrending. The method of detrending-by-polynomials as defined in step A10, ensures that A defines also subspace of $\bar{Z}_2$ if detrending is in force. The transition formulae (8.15) - (8.18) define an eigenvalue equation of which all eigenvalues are nonnegative real values. For a proof see Appendix A.

6. The algorithm is a kind of power algorithm to obtain eigenvalues (Gourlay and Watson, 1973: chapter 4). The convergence of power algorithms is slow. To speed up the convergence, the algorithm developed by Hill (1979) is used (subroutine EIGY). Each iteration of this algorithm consists of four iterations of Steps A1 - Step A7 leading consecutively to four trial vectors $\underline{x}_1$, $\underline{x}_2$, $\underline{x}_3$ and $\underline{x}_4$. With X the n×4 matrix containing these trial vectors, the algorithm calculates the dominant eigenvector of the 4×4 matrix X'X and the dominant eigenvector is taken as the vector of trial scores in the next iteration. The eigenvalues of X'X are determined by reducing X'X to a symmetric tridiagonal form of which the dominant eigenvector can be calculated explicitly (Gourlay and Watson, 1973). The steps A1-A7 are performed in subroutine TRANS.

   For the calculation of subdominant eigenvectors (axes 2, 3 and 4), step A6 in which trial vectors are made uncorrelated with previous axes, is essential. It replaces the "deflation" process described in many textbooks (e.g. Gourlay and Watson, 1973: chapter 5). By deflation, a new matrix is derived from the original matrix in such a way that the dominant eigenvector of the new matrix is the second eigenvector of the original matrix. The disadvantage of deflation is that although the original matrix is sparse (i.e. contains few nonzero elements), the deflated matrix is not sparse in general. Replacing the deflation process by step A6 has the advantage that the species data can remain in condensed format when extracting the ordination axes 2, 3 and 4.

7. If a particular scaling of the biplot or the joint plot is wanted, the ordination axes may require linear rescaling. With linear methods one can choose between a Euclidean distance biplot and a covariance biplot, which focus on the approximate Euclidean distances between sites and correlations among species, respectively (Ter Braak, 1983). With weighted averaging methods it is customary to use the site scores $\underline{x}^*$ (8.16) and the species scores $\underline{u}$ (8.15) to prepare an ordination diagram after a linear rescaling so that the average within-site variance of the species scores is equal to 1 (see Eq. (4.9)). The linear rescaling involves multiplication of the species and sites scores of an ordination axis by a function of its eigenvalue.

9. TECHNICAL DETAILS

9.1 Dimensioning

The capacity of the program can be changed by changing the values in the PARAMETER statement on line 72-73 of CANOCO (see also Table 3.4). Note that NZMAX should always be 8 plus the maximum number of environmental variables one wants to analyze. The number of environmental variables and of covariables that can be 'read from file' is allowed to be 4 times larger. The number of variables analyzed is determined after deletion and after the definition of interaction terms. The maximum number of covariables includes the variables used in detrending by polynomials (maximum 12). Similarly, the maximum number of values of covariables include the polynomials used for detrending (maximum 12 × (number of samples)). The number of presences in the species data is equal to the number of non-zero values in a full format file or to the number of couplets specified in a Cornell condensed format file.

9.2 Structure of the main program

1. Input parameters are entered by subroutine AANVAN (Q1-Q18).
2. The species data matrix is read by subroutine QUIKIN (Q19).
3. The environmental data matrix is read by subroutine ENVIN (Q20-Q21).
4. The covariable data matrix is read by subroutine ENVIN (Q22-Q23). The names of the variables are printed in the output file.
5. The species data are transformed by subroutine SWEIGH (Q24-Q32). This routine determines which samples and species are active and which are passive.
6. Passive and deleted samples are removed from the environmental data (subroutine SAMDEL). Deleted samples are removed from the covariable data (subroutine SLUITA).
7. For linear methods, the total sum of squares (Q27) is calculated in subroutine SPEVAR.
8. The weights for species and samples (step P1, section 8) are calculated. (WEIGHX contains the sample weights of section 4.6; WEIGHY contains the species weights of section 4.5, both divided by their total).
9. The covariables are made uncorrelated by the Gram-Smidt orthogonalization process (subroutine ORTHO).
10. The covariance matrix and correlation matrix of the environmental data are calculated in the subroutines SSQPRO and SSPCOR (adapted from Herraman, 1968).
11. The environmental variables are standardized to mean 0 and variance 1 (step P2, section 8).
12. With covariables in the analysis, the environmental variables are made uncorrelated to the covariables by subroutine PROJCT. This amounts to step P3 (section 8). The new environmental values are residuals of a multiple regression on the covariables. A new covariance matrix of environmental variables is calculated.

13. The correlation matrix of the environmental variables (or, with covariables, the new covariance matrix) is displayed at the terminal by the subroutine PRIMAT, if desired (Q33).
14. The inverse of the correlation matrix is calculated by the subroutine GSWEEP (Clarke, 1982). Multicollinearity between environmental variables is determined in this routine. The subroutines SSQPRO, SSSPCOR and GSWEEP are the only routines that use double precision.
15. The trace (= sum of all canonical eigenvalues) is calculated in routine TRACES.
16. The means, standard deviations and variance inflation factors of environmental variables are displayed at the terminal. The means and standard deviations refer to the original variables before standardization (step 10).
17. The eigenvectors and eigenvalues are calculated by subroutines EIGY and TRANS (see section 8). The eigenvectors are in XENV$\underline{s}$ and YEIG$\underline{s}$ for samples and species, respectively ($\underline{s}$ = 1,2,3,4). The canonical coefficients are in BEIG$\underline{s}$. In contrast to DECORANA, iteration is on the sample scores and not on the species scores. This is advantageous for linear methods as the species scores often have a nonzero mean. Moreover, the number of samples is usually less than the number of species. In contrast to samples, passive species and deleted species are not removed from the calculations (but their influence is made negligible by the species weights). The scalar NEIGZ determines the number of constrained/-canonical axes.
18. If detrending-by-polynomials is required, step A10 (section 8) is carried out by subroutine POLXTZ. If a polynomial is linearly dependent to the covariables or the previous columns of the matrix A of section 8, this polynomial is not added to A. The condition $q \leq s_A$ in step A4 and A5 is checked by adapting the value of NEIGZ.
19. If detrending-by-segments is required, the segments used in the detrending process are determined in the subroutine CUTUP (Hill, 1979).
20. After the eigenvectors have been calculated, the sample scores x_i^* are determined for canonical ordination axes by weighted summation or weighted averaging of species scores (see section 4.6). They are centred, if required, and with covariables they are made uncorrelated to the covariables (subroutine ORTHO). For ordination axes that are not canonical, the sample scores x_i are determined by regressing the eigenvectors on the environmental variables (in subroutine NOTCAN). The scores are linearly rescaled in subroutine POSMUL (section 9.3).
21. The full correlation matrix of ordination axes and environmental variables is calculated, and if desired, printed on the output file (subroutine PRIMAT). It are partial correlations if there are covariables.
22. The biplot scores and centroids of environmental variables are calculated as described in section 4.9.
23. The percentage variance accounted for is calculated in subroutine TRACET (section 4.4) and is printed.

24. The ordination results are displayed at the terminal, printed on the output file and written to the machine readable files (Q34) by subroutine EIGOUT. This routine determines the multipliers (section 4.5).
25. Further analyses are specified in subroutine MANIP (Q35-Q37).
26. The Monte Carlo permutation test is directed by subroutine PERMUT (Q38-Q40). See section 9.4.

9.3 Scaling of the axes

In CANOCO the species and sample scores are scaled according to Eqs. (4.5)-(4.7) and (4.10)-(4.11). This is achieved by subroutine POSMUL. Users who wish another scaling may change the definition of the arrays FX and FY of this subroutine. The scalar IBI contains the answer to Q14/Q15; α = 1, 0, $\frac{1}{2}$ for IBI = 1, 2 and 3, respectively. On entry the eigenvector sample scores $\tilde{x}_i$ (section 4.5) have a weighted mean square of 1 (cf. Eqs. (4.10) and (4.11)) and are derived from the species scores, i.e. the left hand sides of (4.5) and (4.6) are equal to λ_s^{-1}. The arrays YEIG$\underline{s}$, XEIG$\underline{s}$, XENV$\underline{s}$ and BEIG$\underline{s}$ contain the scores u_k (or b_k), x_i^*, x_i and c_k as defined in section 4. In POSMUL XEIG$\underline{s}$, XENV$\underline{s}$ and BEIG$\underline{s}$ are multiplied by FX$\underline{s}$, and YEIG$\underline{s}$ by FY$\underline{s}$.

If nonlinear rescaling of axes is in force (usually in DCA with detrending-by-segments), then the minimum of the scores x_i^* is determined for each axis and used to make the minimum of x_i^* equal to 0 in the output. For details of the nonlinear rescaling see Hill (1979).

9.4 Monte Carlo permutation test

CANOCO uses the pseudo-random generator described by Wichman and Hill (1982). It is a multiplicative congruential generator (Zeisel, 1986): $X_{n+1} = \alpha X_n$ modulo m with α = 1655 54252 64690 and m = 2781 71856 04309. Its period is 6.95×10^{12}. The generator uses three seeds, two of which can be specified by the user in response to Q39. The third seed is set to 15357 and cannot be modified by the user.

In contrast to the description of the permutation test in section 4.11, the samples in the environmental data are kept fixed and the samples in the species data and covariable data are permuted. The permutation is applied to the covariables via the array INPERM which contains sample indices and to the species data by the arrays IBEGIN and IEND which specify the beginning and end of the data of each sample in the condensed arrays IDAT and QIDAT (which contain the species numbers and species values, respectively).

Subroutine CONDPR determines the permutation classes (section 4.11). It takes a linear combination of the conditioning covariables and determines which samples have the same value for this linear combination. A random permutation is determined by drawing n pseudo random values between 0 and 1 and ordering the sample numbers in parallel with the values obtained within each permutation class (this is achieved by a single ranking by adding the permutation class numbers to the pseudo random numbers and some housekeeping of sample indices).

With covariables, in the analysis, the environmental variables must be regressed on the covariables (step P3, section 8). The residuals of this regression replace the environmental variables. This regression must be

carried out in each permutation. Because no copy is being hold of the original environmental variables, they are reconstituted from the residuals, the regression coefficients and the covariables. The latter two determine the fitted values. Adding to these the residuals restores the original values (subroutine RESTOZ). Step P3 (section 8) is carried out by the steps 9 and 12 of section 9.2.

In each permutation, the inverse of covariance matrix of the (residual) environmental variables is determined (steps 12 and 14 of section 9.2). If Q37 = 1, then the first eigenvalue of the data after permutation is calculated (subroutine EIGY and TRANS). If Q37 = 2, then the trace statistic is calculated (subroutine TRACES). The printing of results and the calculation of the P-value is done in subroutine PERMUT.

9.5 Some points concerning CVA

Most computer programs calculate the eigenvalue of

$$(B - \theta W)\, \underline{c} = \underline{0} \qquad (9.1)$$

where W is the matrix of within-cluster sums of squares and products. Because T=B+W it can be shown that:

$$\theta = \lambda / (1 - \lambda). \qquad (9.2)$$

θ is closely related to an F-ratio. CVA can be defined as the technique that chooses the linear combination of environmental variables that gives the highest F-ratio in a one-way analysis of variance (with clusters as 'treatments'). It can be shown that the maximized F-ratio is equal to

$$F = \left[(n-g)/(g-1)\right] \theta \qquad (9.3)$$

with n the number of samples and g the number of clusters. Note however that this F-ratio does not follow an F-distribution. Use the permutation test instead.

The percentage variance accounted for in CVA is, for a two-dimentsional ordination diagram, usually taken to be

$$V = (\theta_1 + \theta_2) \,/\, (\text{sum of all } \theta\text{'s}) \qquad (9.4)$$

The percentage variance accounted for by the species-environment biplot as given by CANOCO is however

$$C = (\lambda_1 + \lambda_2) \,/\, (\text{sum of all } \lambda\text{'s}) \qquad (9.5)$$

V and C are both percentages of weighted variances, but the weights differ. With V, the inverse of the within-cluster matrix W is used for weights, whereas with C the inverse of the total matrix T is used.

From the CANOCO output, V can be calculated when there are less than six clusters by calculating all θ's from the canonical eigenvalues λ given by CANOCO. With six or more clusters, a lower and upper bound for V can be derived as follows (for two dimensions):

$$\text{lower bound for } V = (\theta_1 + \theta_2)/(a+b) \qquad (9.6)$$
$$\text{upper bound for } V = (\theta_1 + \theta_2)/(a+d) \qquad (9.7)$$

where $a = \theta_1 + \theta_2 + \theta_3 + \theta_4$
$b = (\text{trace} - a)/(1 - \lambda_4)$
$d = \text{trace} - a$

with trace the trace reported by CANOCO (= sum of all λ's) and θ is calculated from λ by formula (9.2). With six clusters the actual V is precisely equal to the lower bound given by (9.6).

10. INSTALLATION NOTES

CANOCO consists of ca. 5000 lines of code written in standard FORTRAN77. You may need to adapt the following features to let the program run on your computer system.

Input from the terminal is assumed to come from unit number SYSIN. Output to the terminal is assumed to be written to unit number SYSOUT. The values of SYSIN and SYSOUT can be changed in the main program (line 160-161). For VAX-computers and IBM-mainframe one should set SYSIN to 5 and SYSOUT to 6. For IBM-PC's with MS-FORTRAN, SYSIN and SYSOUT should both be set to 0.

The input and output files of CANOCO are described in section 3.2. Unit numbers of input files are:

INPENV = 2 (file of environmental data and of covariables; see Q5, Q7 and Q35)
SPECIN = 3 (file of species data; see Q3)

Unit numbers of output files are:

ICONS = 1 (annotated copy of terminal dialogue; see sections 3.2 and 3.3 and Table 3.3)
IPRN = 4 (output file for line printer; see Q1)
IPUN = 7 (file with machine readable copy; see Q17)
IPUN2 = 8 (file with machine readable copy; see Q18)

The unit numbers can be changed in the main program (after line 168). The names of the files are asked for in the terminal dialogue, except the name of the annotated copy. This file is given a name in the main program (line 155) by the statement

```
CANOIN  =  'CANOCO.CON'.
```

You must change the name CANOCO.CON, if this name is not a valid file name on your computer system. You have then also to make a change in subroutine ERRORF where CANOIN is assigned the name 'CANOC2.CON' on line 14. File are actually opened in subroutine OPENF. If your system has an option for READONLY files, you can add such option in subroutine OPENF.

Some systems require that SYSIN (the console) is rewinded each time a question is answered by RETURN (see section 3.3). If so, change line 153 of the main program to

```
IBMMFR = 1.
```

If this does not help, adapt the subroutine ERRORF which is called from the subroutines READI, READF and READC (these subroutines read the answers to questions from the terminal).

11. ACKNOWLEDGEMENTS

I would like to thank Dr. I.C. Prentice for discussions that stimulated me to develop CANOCO and to extend the program to include partial constrained ordination. I am grateful to Drs. M.O. Hill and H.G. Gauch for permission to use the code of DECORANA, which is still one fifth of the code of CANOCO. Drs. O.F.R. van Tongeren helped me to make the program more user-friendly and adapted the program for IBM-PC's. Several researchers of the Research Institute for Nature Management, in particular Drs. P.F.M. Verdonschot and A.G.M. Schotman, helped to improve CANOCO by detecting errors in earlier versions. Drs. H.J.B. Birks, J.J. de Gruijter, A.A.M. Jansen, B.J. Post, I.C. Prentice, H.C. Prentice and O.F.R. van Tongeren commented on (parts of) the manual. Miss J. van de Peppel prepared the figures and Mrs. M. Mijling typed the manuscript.

12. REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). J.R. Statist. Soc. B 44: 139-177.

Aitchison, J. (1983). Principal component analysis of compositional data. Biometrika 70: 57-65.

Aitchison, J. (1984a). The statistical analysis of geochemical compositions. Math. Geol. 16: 531-564.

Aitchison, J. (1984b). Reducing the dimensionality of compositional data sets. Math. Geol. 16: 617-635.

Aitchison, J. (1986). The statistical analysis of compositional data. Chapman and Hall, London.

Anderson, J.A. (1984). Regression and ordered categorical variables. J.R. Statist. Soc B. 46: 1-30.

Chessel, D., Lebreton, J.D. et Yoccoz, N. (1987). Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie. Revue de Statistique Appliquée 1987 (4): 55-72.

Clarke, M.R.B. (1982). The Gauss-Jordan sweep operator with detection of collinearity, (AS 178). Appl. Statist. 31: 166-168.

Cochran, W.G. and Cox, G.M. (1957). Experimental designs. Wiley, New York.

Cox, D.R. (1958). Planning of experiments. Wiley, New York.

Corsten, L.C.A. and Gabriel, K.R. (1976). Graphical exploration in comparing variance matrices. Biometrics 32: 851-863.

Davies, P.T. and Tso, M.K.S. (1982). Procedures for reduced-rank regression. Appl. Statist. 31: 244-255.

Feoli, E. and Orlóci, L. (1979). Analysis of concentration and detection of underlying factors in structured tables. Vegetatio 40: 49-54.

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

Gabriel, K.R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. In: V. Barnett (ed.): Interpreting multivariate data. Wiley, Chicester, p. 147-173.

Gauch, H.G. (1982). Multivariate analysis in community ecology. Cambridge University Press, Cambridge.

Gifi, A. (1981). Nonlinear multivariate analysis. Department of Data Theory, University of Leiden, Leiden.

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

Goodman, L.A. (1981). Association models and canonical correlation in the analysis of cross-calssifications having ordered categories. J. Amer. Statist. Ass. 76: 320-334.

Gordon, A.D. (1981). Classification. Methods for exploratory analysis of multivariate data. Chapman and Hall, London.

Gourlay, A.R. & G.A. Watson (1973). Computational methods for matrix eigen problems. Wiley, New York.

Gower, J.C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53: 325-338.

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

Heiser, W.J. (1986). Joint ordination of species and sites: the unfolding technique. In : New developments in numerical ecology, (P. Legendre and L. Legendre, eds.), Springer-Verlag, Berlin, in press.

Herraman, C. (1968). Normalizing a symmetric matrix (AS11). Sums of squares and products matrix (AS12). Appl. Statist. 17: 287-292.

Hill, M.O. (1973). Reciprocal averaging: an eigenvector method of ordination. J. Ecol. 61: 237-249.

Hill, M.O. (1979). DECORANA: a FORTRAN program for detrended correspondence analysis and reciprocal averaging. Section of Ecology and Systematics, Cornell University, Ithaca, New York.

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

Hope, A.C.A. (1968). A simplified Monte Carlo significance test procedure. J. Roy. Statist. Soc. Series B 30: 582-598.

Ihm, P. & H. van Groenewoud, 1984. Correspondence analysis and Gaussian ordination. COMPSTAT lectures 3: 5-60.

Israëls, A.Z. (1984). Redundancy analysis for qualitative variables. Psychometrika, 49: 331-346.

Jongman, R.H.G., Ter Braak, C.J.F. and Van Tongeren, O.F.R. (1987). Data analysis in community and landscape ecology. Pudoc, Wageningen.

Kendall, M.G., and Stuart, A. (1973). The advanced theory of statistics. Vol. II. Inference and relationship. Griffin, London.

Kenkel, N.C., and Orlóci, L. (1986). Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. Ecology 67: 919-928.

Kooijman, S.A.L.M., 1977. Species abundance with optimum relations to environmental factors. Annals of System Research 6: 123-138.

Laurec, A., Chardy, P., de la Salle, P., and Rickaert, M. (1979). Use of dual structures in inertia analysis: ecological implications. In: Multivariate Methods in Ecological Work, (L. Orlóci, C.R. Rao, and W.M. Stiteler, eds.), pp 127-174. Intern. Co-operative Publ. House, Fairland, Maryland.

Lebreton, J.D., Chessel, D., Prodon, R. et Yoccoz, N. (1988). L'analyse des relations especes-milieu par l'analyse canonique des correspondances. I Variables de milieu quantitatives. II Variables de milieu qualitatives. Acta Oecologia, Oecologia generalis 9: 53-67 and 9: 137-151.

Manly, B.F.J. (1985). The statistics of natural selection on animal populations. Chapman and Hall, London.

McCullagh, P. and Nelder, J.A. (1983). Generalized linear models. Chapman and Hall, London.

McKechnie, S.W., Ehrlich, P.R., and White, R.R. (1975). Population genetics of Euphydryas butterflies. I Genetic variation and the neutrality hypothesis. Genetics 81: 571-594.

Metcalf, M. (1985). Effective FORTRAN 77. Clarendon Press, Oxford.

Meulman, J. (1986). A distance approach to nonlinear multivariate analysis. DSWO Press, Leiden.

Minchin, P. (1986). An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69.

Montgomery, D.C. & E.A. Peck (1982). Introduction to linear regression analysis. Wiley, New York, 504 pp.

Naes, T., Irgens, C., and Martens, H. (1986). Comparision of linear statistical methods for calibration of NIR instruments. Appl. Statist. 35: 195-206.

Noy-Meir, I. (1973). Data transformation in ecological ordination. I Some advantages of non-centering. J. Ecol. 61: 329-341.

Noy-Meir, I., Walker, D. and Williams, W.T. (1975). Data transformations in ecological ordination. II. On the meaning of data standardization. J. Ecol. 63: 779-800.

Persson, S. (1981). Ecological indicator values as an aid in the interpretation of ordination diagrams. J. Ecol. 69: 71-84.

Post, B.J. (1986). Factors of influence on the development of an arable weed vegetation. Proc. EWRS Symposium 1986, Economic Weed Control: 317-325.

Prentice, I.C. (1980). Multidimensional scaling as a research tool in quaternary palynology: a review of theory and methods. Rev. Palaeobot. Palynol. 31: 71-104.

Rao, C.R. (1973). Linear statistical inference and its application. 2nd ed. Wiley, New York.

Robert, P., and Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: the RV-coefficient. Appl. Statist. 25: 257-265.

Seber, G.A.F. (1977). Linear regression analysis. Wiley, New York.

Sokal, R.R. (1969). Testing the statistical significance of geographical variation patterns. Syst. Zool. 28: 227-232.

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

Ter Braak, C.J.F. (1985). Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics 41: 859-873.

Ter Braak, C.J.F. (1986a). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. Ecology, 67: 1167-1179.

Ter Braak, C.J.F. (1986b). The analysis of vegetation-environment relationships by canonical correspondence analysis. Vegetatio 69: 69-77.

Ter Braak, C.J.F. (1987). Ordination. Pages 91-173 in: Data analysis in community and landscape ecology (R.H.G. Jongman, C.J.F. Ter Braak and O.F.R. Van Tongeren, eds.). Pudoc, Wageningen.

Ter Braak, C.J.F., and L.G. Barendregt (1986). Weighted averaging of species indicator values: its efficiency in environmental calibration. Math. Biosci. 78: 57-72.

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

Ter Braak, C.J.F. and Prentice, I.C. (1988). A theory of gradient analysis. Advances in ecological research, 18:271-317.

Tso, M. K-S. (1981). Reduced-rank regression and canonical analysis. J.R.Statist. Soc. B 43: 183-189.

Van den Wollenberg, A.L. (1977). Redundancy analysis. An alternative for canonical correlation analysis. Psychometrika 42, 207-219.

Wichman, B.A. and Hill, I.D. (1982). An efficient and portable pseudo- random number generator (AS183). Appl. Statist. 31: 188-190 (correction Appl. Statist. 33: 123).

Wold, H. (1982). Soft modeling. The basic design and some extensions. Pages 1-54 in: Systems under indirect observation. Causality-structure-prediction, Vol. II (Jöreskog, K.G. and Wold, H. eds.) North-Holland Publishers, Amsterdam.

Zeisel, H. (1986). A remark on Algorithm AS 183. Appl. Statist. 35: 89.

APPENDIX A

Theorem : The equations (8.15) - (8.18) define an eigenvalue equation whose eigenvalues are nonnegative real values.

Proof: Note the following relations for the projection operator B°, with W a positive definite diagonal matrix:

$$B^\circ = B(B'WB)^{-1}B'W \tag{A.1}$$

$$B^\circ B^\circ = B^\circ \qquad \text{(idempotent)} \tag{A.2}$$

$$WB^\circ = (B^\circ)'W \tag{A.3}$$

$$B^\circ W^{-1} = (B^\circ W^{-1})'W(B^\circ W^{-1}) \tag{A.4}$$

(A.1) is the definition of B°. (A.2) is a consequence of B° being a projector. (A.3) follows (A.1) and $W = W'$. (A.4) follows by algebraic manipulation using (A.2) and (A.3).

Starting from (8.15) we obtain by successive insertions:

$$\underset{\sim}{u} = \lambda^{-\alpha}R^{-1}Y'\underset{\sim}{x} = \lambda^{-\alpha}\, R^{-1}Y'(I-A^\circ)\tilde{Z}_2\underset{\sim}{c} = \tag{A.5}$$

$$= \lambda^{-\alpha}R^{-1}Y'(I-A^\circ)\tilde{Z}_2(\tilde{Z}_2'W\tilde{Z}_2)^{-1}\tilde{Z}_2'Wx^* =$$

$$= \lambda^{-\alpha}R^{-1}Y'(I-A^\circ)\tilde{Z}_2^\circ x^* = \lambda^{-1}R^{-1}Y'(I-A^\circ)\tilde{Z}_2^\circ(I-Z_1^\circ)W^{-1}Y\underset{\sim}{u}\ .$$

Because of (8.19), $\tilde{Z}_2$ is a subspace of $I-Z_1^\circ$ so that $\tilde{Z}_2^\circ(I-Z_1^\circ) = \tilde{Z}_2^\circ$. Further, because of (8.10) and step A10, A is a subspace of $\tilde{Z}_2$, so that $(I-A^\circ)\tilde{Z}_2^\circ = \left[(I-A^\circ)\tilde{Z}_2\right]^\circ$.

On using these equations and (A.4) with $B = (I-A^\circ)\tilde{Z}_2$, we obtain

$$\lambda R\underset{\sim}{u} = Y'B^\circ W^{-1}Y\underset{\sim}{u} = Y'(B^\circ W^{-1})'W(B^\circ W^{-1})Y\underset{\sim}{u} \tag{A.6}$$

$$= C'Cu, \text{ say}$$

(A.6) defines a generalized eigenvalue problem. Because R is symmetric and positive definite and C'C is symmetric and semi-positive definite, its eigenvalues are real and nonnegative.

APPENDIX B

Constrained principal coordinates analysis

We develop a principal coordinates analysis in which the ordination axes are constrained to be linear combination of external variables. The form of constrained principal coordinates analysis that we define relates to principal coordinates analysis as principal components analysis relates to reduced-rank regression (Davies and Tso, 1982) which is better known as redundancy analysis (Van den Wollenberg, 1977; Israëls, 1984). Other approaches to constrained principal coordinates analysis are given by Meulman (1986).

Let $D = \{\delta_{ij}\}$ be a n×n matrix of dissimilarities between sample i and sample j ($\delta_{ij} \geq 0$; $\delta_{ii} = 0$) and Z a n×q matrix of n observations of q explanatory variables ($q < n$). Let A be a n×n matrix whose elements $\{a_{ij}\}$ are derived from D by

$$a_{ij} = -\tfrac{1}{2}\{\delta_{ij} - \delta_{i.} - \delta_{.j} + \delta_{..}\} \qquad (B.1)$$

Definition: Constrained principal coordinates analysis of the dissimilarity matrix D with respect to the q explanatory variables in Z is the eigen analysis of the matrix $Z(Z'Z)^{-1}Z'A$ with A defined in (B.1).

Theorem: If there exists a n×m matrix Y for some integer $m \geq 0$, such that D is the matrix which elements contain the squared Eucledean distances between the rows of Y, then constrained principal coordinates analysis of D with respect to Z is identical to redundancy analysis of Y with respect to Z.

Proof: The crux of principal coordinates analysis is that under the assumptions of the theorem $A = YY'$, where it is assumed that the column means of Y have been substracted already (Gower, 1966). This can be shown from the relation [with $y'_{(i)}$ the i-th row of Y]

$$\delta_{ij} = \sum_{k=1}^{m} (y_{ik} - y_{jk})^2 = y'_{(i)}y_{(i)} + y'_{(j)}y_{(j)} - 2\, y'_{(i)}y_{(j)} . \qquad (B.2)$$

Because $a_{ij} = y'_{(i)}y_{(j)}$, we have from (B.2)

$$a_{ij} = -\tfrac{1}{2}\,\delta_{ij} + \tfrac{1}{2}\,a_{ii} + \tfrac{1}{2}\,a_{jj} . \qquad (B.3)$$

Moreover, when Y is centred by columns, $a_{.j} = 0$ and $a_{i.} = 0$, so that a_{ij} can be obtained by double centring the matrix with elements $-\frac{1}{2}\delta_{ij}$, whence (B.1).

Let Z° denote the symmetric idempotent projection operator on the columns of Z, i.e. $Z^\circ = Z(Z'Z)^{-1}Z'$. Redundancy analysis of Y with respect to Z is identical to principal components analysis of the matrix $Z^\circ Y$ which contains the fitted values of multiple regressions of each column of Y on the columns of Z (Davies and Tso, 1982). The sample scores are thus the eigenvectors of $Z^\circ YY'Z^\circ$, i.e. the sample scores satisfy the eigenvector equation (with λ the eigenvalue)

$$Z^\circ YY'Z^\circ \underset{\sim}{x} = \lambda \underset{\sim}{x} . \qquad (B.4)$$

Now note that $\underset{\sim}{x}$ lies in the column space of Z because of the left most Z° in (B.4). Therefore $\underset{\sim}{x} = Z^\circ \underset{\sim}{x}$, so that

$$Z^\circ YY' \underset{\sim}{x} = \lambda \underset{\sim}{x} \qquad (B.5)$$

Under the assumptions of the theorem

$$Z^\circ YY' \underset{\sim}{x} = Z^\circ A \underset{\sim}{x} = \lambda \underset{\sim}{x} \qquad (B.6)$$

Hence, constrained principal coordinates analysis of D with respect to Z results in the same sample scores and eigenvalues as redundancy analysis of Y with respect to Z.

Important: Unfortunately CANOCO cannot solve the eigenvalue equation (B.6). When a dissimilarity matrix is analysed with CANOCO using RDA and double centring, then CANOCO solves the eigenvalue equation

$$Z^\circ AA' \underset{\sim}{x} = \lambda \underset{\sim}{x} \qquad (B.7)$$

In general, (B.6) an (B.7) have different solutions. It is therefore strongly recommended to use neither RDA nor the associated Monte Carlo permutation test when analysing a dissimilarity matrix with CANOCO.

APPENDIX C

Trace and short-cut formulae (4.17) and (4.19)

The notation of section 8 is used in this Appendix. The trace calculated by CANOCO is in this notation

$$\text{trace } (R^{-1}Y'\tilde{Z}_2(\tilde{Z}_2'W\tilde{Z}_2)^{-1}\tilde{Z}_2'Y). \qquad (C.1)$$

The trace is the sum of the canonical eigenvalues (without detrending) as follows from the eigenvalue equation in (A.5) with $I-A^{\circ}=I$.

To obtain (4.17), let U and X^* be the m×s and n×s matrices whose columns contain the species scores and sample scores $\{x_i^*\}$, respectively, on s ordination axes and let Λ be the s×s diagonal matrix $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_s)$. The weighted averages of the species with respect to the environmental variables in $\tilde{Z}_2$ are given by the m×q matrix

$$S_{12} = R^{-1}Y'\tilde{Z}_2 \ . \qquad (C.2)$$

In linear methods S_{12} contains (partial) covariances multiplied by n. The biplot scores of the environmental variables are obtained by fitting the model

$$S_{12} = UC_e' + \text{error} \qquad (C.3)$$

by a weighted multiple regression of S_{12} on U. The estimator for the q×s matrix C_e of biplot scores is

$$C_e = S_{12}'RU(U'RU)^{-1} = \tilde{Z}_2'YU(U'RU)^{-1} \qquad (C.4)$$

where R is the weight matrix. Unless detrending-by-segments or nonlinear rescaling of axes is in force, we have [with $\tilde{\Lambda} = I$ in linear methods and $\tilde{\Lambda} = (I-\Lambda)^{-1}$ in weighted averaging methods]

$$U'RU = \Lambda^{1-\alpha}\Sigma_i w_i\tilde{\Lambda} \qquad (C.5)$$

because U contains orthogonal eigenvectors scaled according to (4.5) and (4.6) [in section 4.5: $\Sigma_k w_k^* = \Sigma_i w_i^*$]. Further note that on using (8.16) and (A.3)

$$\tilde{Z}_2'WX^*\Lambda^{1-\alpha} = \tilde{Z}_2'W(I-Z_1^\circ)W^{-1}YU = \tag{C.6}$$

$$\tilde{Z}_2'W\,W^{-1}(I-Z_1^\circ)'YU = \left((I-Z_1^\circ)\tilde{Z}_2\right)'YU = \tilde{Z}_2'YU$$

so that (C.4) can be simplified to

$$C_e = \left[\tilde{Z}_2'WX^*(\Sigma_i w_i)^{-1}\right]\tilde{\Lambda}^{-1} \tag{C.7}$$

whence (4.17) when $\tilde{Z}_2$ is standardized. If there are covariables, $\tilde{Z}_2$ is not standardized after the regression on the covariables and (4.17) must therefore be multiplied by the residual standard deviation. These formulae carry through in passive analysis of environmental variables (section 3.8: Q35 = 2 or 3), in contrast to formulae that use the environmental axes and the intra set correlations.

We now derive (4.19) from (4.18). Let Z_2 contain the values of the environmental variables before standardization. In matrix notation the numerator in (4.18) is then $Z_2'WX^*$. From (8.16) we know that the columns of X^* are orthogonal to the covariables; therefore

$$Z_2'WX^* = Z_2'W(I-Z_1^\circ)X^* = \left[(I-Z_1^\circ)Z_2\right]'WX^*. \tag{C.8}$$

With (C.8), (4.19) follows from (4.18) by noting that if the species axis has a nonzero mean, the mean value carries through additively in (4.18).