UPDATE NOTES: CANOCO VERSION 3.10

Cajo J.F. ter Braak

1 New features in CANOCO 3.10 compared to CANOCO 2.x:

- Forward selection of environmental variables (with testing if desired) to investigate which variables best explain the species data.

- Regression/ordination diagnostics for samples and species to check which species and samples fit well and which do not fit into the ordination diagram. In (detrended) correspondence analysis, a measure for niche width of species is given and a measure for sample heterogeneity.

- Redundancy analysis with a statistical weighting of species. This option makes the analysis invariant to linear rescaling of species (an attractive property shared with canonical correlation analysis).

- Percentages variance accounted for in CA and CCA. Formerly, only eigenvalues were calculated which were often misunderstood as percentages (they were in PCA/RDA).

- Collinear environmental variables (e.g. the K-th dummy variable of a nominal variable with K classes) are kept in the analysis. The default analysis thus yields CENTROIDS for all classes instead of for K-1 classes. In CANOCO 2.x, this required an extra analysis with passive environmental variables.

- Data input in free format, without the need to add sample numbers in front of each line. The error messages when reading all types of input are improved.

- Principal component scores of species adjusted or unadjusted for species variance. The new default is 'adjusted', which means that species scores are exactly the correlations of species with axes if the scaling of a correlation biplot is chosen (the former covariance biplot).

- Hill's scaling is no longer the default in CA/CCA. Instead, the default is the biplot scaling, which is easier to interpret (scaling to 1 instead of to 1/(1-lambda) and to lambda instead of to lambda/(1-lambda)).

- Choice between short and long dialogue. In the short dialogue, one can delete samples, covariables, environmental variables and choose a transformation for the species data. The analysis can be continued in the usual way with passive analyses, Monte Carlo statistical tests, etc.

- You may specify your own defaults in a file CANOCO.INI, including names of files to be analyzed. This is especially convenient when using the short dialogue.

- Screenwise output. CANOCO can present screenwise output if an ANSI screen code driver is installed on the computer (e.g. DEVICE=ANSI.SYS in the CONFIG.SYS of a MS-DOS computer). After the message at the bottom of the screen

'Press RETURN for more, S to Skip, Q to Quit page-mode'
the user can press <RETURN> to continue, S <RETURN> for skipping the page-mode for the current output item, or Q <RETURN> to quit the page-mode for the entire CANOCO session. Subsequent output starts from the top of the screen.

- More than 4 ordination axes may be extracted by asking for More analyses with current data (instead of Stop).

- Environmental variables can be turned into covariables after the question which asks for more analyses with current data.

- The methods used for the Monte Carlo test have been changed. The advantages of the version 3.x method are:
  -- interactions effects can be validly tested.
  -- the correlation structure of the explanatory variables (covariables and environmental variables) is not changed during permutation.
  -- more power by permutation of residuals under the full model.
  The old and new method yield equivalent results in simple situations such as overall tests without covariables and tests with a conditioning on all covariables (randomized block designs).

- In Monte Carlo testing, special permutation schemes have been added: for time series, for samples on a line transect, on a grid and for repeated measurements. Examples are given how to analyze the important Before-After-Control-Impact design.

- The solution file (the former machine readable copy) is, by default, in a new format in which the ordination scores are given as decimal values. The former format reported the scores in whole numbers with the original scores multiplied by the MULTIPLIER (1, 10, 100 or 1000). With the decimal format, the user can specify, in the CANOCO.INI file, symbols to separate numbers, to enclose names and to signal the end of a set of scores. These options make it easier to input the file into a spreadsheet. CANOPLOT does accept both types of format of the solution file. CANODRAW accepts the decimal format only.

- The MS-DOS redirection symbols > and < work now properly with the answer file CANOCO.con. Therefore, one can run CANOCO 'in batch', without output to the screen, or integrate CANOCO into the environment one is accustomed to, for example, Desqview, Windows or Pcshell.

- CANOCO can give an environment-by-species table. This table contains:
  in CA/CCA/DCA: weighted averages of species with respect to standardized environmental variables,
  in PCA/RDA: correlations between species and each environmental variable, when species are centred and standardized, and similar covariances, when species are just centred.

- More flexible memory management, whence more data can be analyzed. On an MS-DOS PC with 405 Kb free memory, CANOCO 3.10 can analyze 500 samples, 500 species, 100 covariables and 58 environmental variables, but with some additional limitations on the number of species occurrences and environmental and covariable values (total data space 45000 reals) A 80x87 coprocessor is optional. On a PC with math coprocessor or on a standard Apple Macintosh, CANOCO 3.10 can analyze at most 1000 samples, 700 species, 75 environmental variables and 100 covariables, (total data size <80000).

CANOCO is written in standard FORTRAN 77. Ready-to-use versions are available for MS-DOS computers and the Apple Macintosh. The ready-to-use version does not include the source code. CANOCO has been successfully implemented on various mainframe computers. For implementation on mainframes and workstations, the source code of CANOCO is available on a DOS or Apple diskette, together with compilation notes and a demo version of the program for DOS or Apple Macintosh.

Default CANOCO.INI for version 3.10

```
*CANOCO (values start in position 2)
 1  = range [0,1]   = (01) decimal output in file for canoplot
   = char           = (02) separator between decimal values in file for canoplot
   = char           = (03) character by which to enclose names „  ”   ”
   = char           = (04) character to close the scores of each item „  „
 1  = range [0,1]   = (05) pagemode of screen
 25 = range [10,100]= (06) number of lines on a screen
 2  = range [-3,3]  = (07) scaling ordination scores pca/rda
 2  = range [-3,3]  = (08) scaling ordination scores ca/dca/cca/dcca
 2  = range [1,4]   = (09) dimension of biplot in DCA and t-value biplot in PCA/RDA
 26 = range [10,46] = (10) number of segments in detrending process in DCA
 4  = range [0,20]  = (11) number of times for nonlinear rescaling
 0  = range [0,100] = (12) 100 TIMES rescaling threshold
 0  = range [0,1]   = (13) downweighting of rare species in ca/cca/dca
 1  = range [0,4]   = (14) centring/standardization by species in pca/rda
 0  = range [0,3]   = (15) centring/standardization by samples in pca/rda
 0  = range [0,1]   = (16) long dialogue
 0  = range [0,1]   = (17) forward selection of environmental variables
 3  = range [0,3]   = (18) ordination diagnostics
 4  = range [0,4]   = (19) output of correlation matrix
 1  = range [0,1]   = (20) spec-envi table on file canoco.pun
 0  = range [0,1]   = (21) symmetric autocovariance function in grid permutations
 0  = range [0,3]   = (22) transformation of species data
 1  = range [0,1]   = (23) value of c in log( y + c) transformation
 7  = range [1,9]   = (24) default analysis number (1=PCA 2=RDA, etc.)
 ANSWERS.CON                          = (25) answer file (input from file)
 CANOCO.SPE                           = (26) file with species data
                                      = (27) file with covariables
                                      = (28) file with environmental data
 CANOCO.OUT                           = (29) print file
 CANOCO.SOL                           = (30) solution file for CANOPLOT or other prog
 CANOCO.PUN                           = (31) output file for spec-envi table
 2 2 0 0 1 2 2 2  = 8 values in range [0,6]  = (32) output ordination results
*ENDCANOCO
```

2 Summary of the ordination

2.1 Summary of the ordination without covariables in the analysis

An example summary from a CCA of the Dunemeadow data is:

```
**** Summary ****

Axes                                   1      2      3      4  Total inertia

Eigenvalues                      :   .461   .298   .160   .134          2.115
Species-environment correlations  :   .958   .902   .855   .889
Cumulative percentage variance
   of species data               :   21.8   35.9   43.5   49.8
   of species-environment relation:   37.8   62.3   75.4   86.3

Sum of all unconstrained eigenvalues                                    2.115
Sum of all canonical     eigenvalues                                    1.220
```

Which items are displayed depends on the particular analysis, e.g. in DCA (segments) the gradient lengths of the axes are given also.

The eigenvalues measure the importance of an axis (values between 0 and 1).

The total inertia is the total variance in the species data as measured by the chi-square of the sample-by-species table divided by the table's total (Greenacre, 1984). Note that, for abundance data or presence-absence data, chi-square does not have its usual statistical meaning; in particular, it does not follow the chi-square distribution. In PCA/RDA, the total variance is always set to 1, because the species data are scaled in this way (CAN21, p. 30).

The species-environment correlation measures the strength of the relation between species and environment for a particular axis. It is akin to the canonical correlation in canonical correlation analysis. It is the correlation between the sample scores for an axis derived from the species data and the sample scores that are linear combinations of the environmental variables. Note that a high correlation does not mean that an appreciable amount of the species data is explained by the environmental variables. The amount explained is given by the eigenvalue in constrained analyses (RDA/CCA) and by r^2 * eigenvalue in unconstrained analyses (PCA/CA). These amounts are given in the next item.

The percentage of variance of the species data explained by the axes is given cumulatively. Except in DCA (segments), these percentages can easily be derived from the eigenvalues and the sum of all unconstrained eigenvalues, e.g., for axis 2, 100 * $(\lambda_1+\lambda_2)$/(sum of all unconstrained eigenvalues). For abundance data or presence-absence data, these percentages are usually quite low, in particular when analyzed with CA/CCA, but this is nothing to worry about. Species data are often very noisy. An ordination diagram that explains only a low percentage may be quite informative (cf. Gauch, 1982).

With environmental variables in the analysis, CANOCO uses these to explain the species data. This yields fitted values for the species. In PCA/RDA, the fitted values can be obtained by a multiple regression for each species on the environmental variables. In

CA/CCA, this is a weighted regression (see Escoufier, 1985 and the Section Ordination diagnostics). The total variance of the fitted values is precisely the sum of all constrained eigenvalues. Each axis explains a part of this variance. This information is given cumulatively in the line 'percentage variance of species-environment relation'. In RDA/CCA, the percentages can easily be calculated from the eigenvalues and the sum of all constrained eigenvalues, e.g., for axis 2, 100 * $(\lambda_1+\lambda_2)/$ (sum of all constrained eigenvalues). In PCA/CA, the formula is a bit more difficult (namely, the eigenvalues in the nominator must be multiplied by the square of the species-environment correlation). The fitted values with two axes can be displayed in a two-dimensional biplot of the species scores and the sample scores that are linear combinations of the environmental variables.

There exists another interpretation of the percentage variance of the species-environment relation. In PCA/RDA, the relationships between the species data and the environmental data can be summarized in a table of species by environmental variables with as entry the correlation between each particular species and each particular environmental variable. The total weighted variance in this table (CAN21, p.94, (C.1)) is precisely the sum of all constrained eigenvalues. Each axis, again, explains part of this table and this information is reported cumulatively as the percentage variance of the species-environment relation. The correlations in the table, as approximated by two axes, can displayed in a two-dimensional biplot of species scores (adjusted for species variance) and the biplot scores of environmental variables. If the species scores are not adjusted for species variance (scaling of ordination scores < 0), the biplot displays covariance instead of correlation. (With covariables in the analysis, partial covariances are displayed.) In CA/CCA/DCA, the entries in the table are weighted averages of species with respect to environmental variables instead of correlations, but for the rest the interpretation is the same.

2.2 Summary of the ordination with covariables in the analysis

An example summary for CCA of the Dunemeadow data with as covariables thickness of the A1 horizon, moisture and quantity of manuring, is as follows

```
**** Summary ****

Axes                                  1      2      3      4  Total inertia

Eigenvalues                     :   .166   .096   .093   .070          2.115
Species-environment correlations :  .940   .793   .803   .771
Cumulative percentage variance
   of species data               :  12.3   19.5   26.4   31.5
   of species-environment relation: 36.9   58.3   78.8   94.3

Sum of all unconstrained eigenvalues (after fitting covariables)       1.346
Sum of all canonical     eigenvalues (after fitting covariables)        .450

Percentages are taken with respect to residual variances
                 i.e. variances after fitting covariables
```

We see from the table that the sum of all unconstrained eigenvalues is no longer equal to the total inertia, because the covariables have already explained of some inertia in species data, namely 2.115 - 1.346 = 0.769. In a CCA with thickness of the A1, moisture and manuring as only environmental variables, the sum of all constrained eigenvalues is indeed 0.769. The additional inertia explained by the other variables, i.e. the environmental variables in the above summary, is 0.450. Note that the sum of 0.769 + 0.450 = 1.219, which is, apart from rounding error, equal to the sum of all constrained eigenvalues in our first CCA on all environmental variables. It is thus possible to decompose the total inertia as is usually done in the analysis of variance and regression analysis. The covariables explain 100 * 0.769/2.115 = 36% of the inertia and our current environmental variables (eliminating covariables) 100 * 0.450/2.115 = 21%. The remaining 43% of the total inertia is unexplained. The theory of decomposing variance is given in full by Whittaker (1984).

The inertia in the species data after fitting the covariables is 1.346. Of this residual inertia, the first axis explains 0.166, i.e. 100 * 0.166/1.346 = 12.3%. That is 100 * 0.166/0.450 = 36.9% of what can in total be explained by the current environmental variables. One finds these percentages in the summary table.

3 Scaling of ordination scores

3.1 Introduction

CANOCO has six ways to scale ordination scores. This section gives guidelines to help making this choice. If you find the choice difficult, it may be a comfort that the different choices of scaling all yield the same ordering of ordination scores and the same Summary of the ordination. Although the scalings do not affect the main aspects of the ordination, they do affect the amount of scatter among axes of an ordination diagram. The scaling also influences some aspects of the interpretation of ordination diagrams. The differences in interpretation are minor if the ratios of eigenvalues are close to 1.

In the long dialogue, the user is asked which scaling CANOCO must use. In the short dialogue, the default scaling is used; this default can be changed by using a CANOCO.INI file (see Initialization file). For the novice, it is probably best to stick to the default scaling.

The default scaling (without Initialization file) has the following properties:

In PCA/RDA, the species scores and biplot scores of environmental variables are correlations with the ordination axes (the 'eigenvector sample scores', see Glossary) and the eigenvector sample scores have variance 1. All three items can be displayed jointly. All pairs of these items (species + environmental variables, species + samples, samples + environmental variables) can be interpreted according to the rules of a biplot (Jongman et al, 1987: section 5.3.4). The biplot of species + environmental variables yields approximate correlation coefficients between species and environmental variables, and between species among themselves and between environmental variables among themselves. The biplot of species + samples yields approximates abundance values (standardized by species) and the biplot of samples + environmental variables yields approximate environmental values (standardized by environmental variables). However, the plot is not correctly scaled to look at inter-sample distances. If inferring inter-sample distances is the main use of the plot, use scaling 1, instead of the default scaling 2. (In analyses with covariables, the word correlation in the above description should formally be replaced by covariance, but the values still lie between -1 and +1).

In CA/CCA, the species scores are weighted averages of eigenvector sample scores, the biplot scores of environmental variables are correlations with the ordination axes and the eigenvector sample scores have variance 1. The species and environmental variables form a biplot which yields approximate weighted averages of species with respect to environmental variables, and the samples and environmental variables form a biplot of approximate environmental values. Thus, projection of samples on an environmental variable yield environmental values and the projection of a species on this variable gives a weighted average. Indeed, if we plot the sample scores on the same scale as the species, then the projection point of the species lies approximately at the centroid of the sample projection points. The weighted averaging principle is thus optimally present in this scaling. Moreover, distances between species are approximate chi-square distances and, from the environmental arrows, one obtains a biplot approximation of correlation among environmental variables (as in PCA/RDA). In this scaling it is also possible to interpret the joint plot of species and samples as a biplot; the values that are approximated for a species are proportional to its relative abundance y_{ik}/y_{i+} (see Ordination diagnostics).

In the default DCA/DCCA (i.e. with detrending-by-segments), Hill's scaling is used, as in DECORANA, in which samples are weighted averages of species. It allows to make a joint plot of species and samples (Jongman et al, 1987: section 5.2.5) and to make a biplot of species scores with biplot scores for environmental variables.

In the above descriptions, the eigenvector sample scores were plotted, i.e. the sample scores derived from the species in indirect techniques (PCA/CA/DCA) and the samples scores derived from the environmental variables in direct techniques (RDA/CCA/DCCA). In direct techniques the species-sample plot approximates the *fitted* abundance values. Plotting the sample scores derived from the species will give a better fit of the original abundance values. (The two sets of sample scores will not differ much if the species-environment correlation is high).

We discuss the scaling for linear methods (PCA/RDA) and unimodal methods (CA/CCA) separately. We assume that the analysis is without covariables. In a final section, it is described what happens with covariables.

3.2 PCA/RDA

In the long dialogue, the user is asked the question:

```
*** Scaling of ordination scores ***
1 = Euclidean distance biplot
2 = correlation biplot
3 = symmetric scaling
Type corresponding negative number for covariance-based scores
Range of valid answers:      -3       [2]       3
```

The six choices are grouped in two sets of three. We first discuss whether to use a negative value or a positive value for the scaling.

The negative values correspond to traditional scalings (the only ones possible in CANOCO v2.x). They use the raw species data, in which the species usually have different variances. The resulting species scores are covariances between species and

eigenvectors if the scaling = -2, and proportional to covariances if the scaling = 1 or 3 (CAN21, p. 44, eq (4.3)). By contrast, the positive values adjust the species scores for species variance. They use data standardized by species. The resulting species scores are correlations between species and eigenvectors if the scaling = 2 and
proportional to correlations if the scaling = 1 or 3. A biplot of species scores and samples displays the raw (centred) species data, for a scaling < 0, and the standardized (centred) species data for a scaling > 0.

The new options are to counteract the effect of total abundance of a species on its score (e.g. Hill, 1973). The following considerations may clarify the situation. The more abundant species often have the larger variance (even after log-transformation) and thus tend to get the largest scores (when not adjusted). In the ordination diagram, the most abundant species thus usually have the longest arrows (they lie far from the origin). That conspicuous position often only says that the species is a dominant and does not say that the species correlates well with the axes. Less abundant species may correlate much better. Adjustment for species variance solves this problem. Another solution is to perform the PCA/RDA on standardized species data (through the option for centring/standardization by species in the long dialogue). But this solution has the disadvantage that it tends to give rare species unduly large weight (see Jongman et al, 1987, section 5.3.5).

We now discuss the choice between scaling 1, 2, and 3 (or -1, -2, -3, which is analogous). See also Jongman et al (1987, section 5.3.4) and the section on biplots.

Scaling 1 (Euclidean distance biplot) is the most appropriate when you are going to focus on the sample ordination, in particular, on the mutual position of samples in the ordination diagram. With this scaling, distances between samples in the diagram approximate Euclidean distances in species-space.

Scaling 2 or -2 (correlation biplot) is the most appropriate when you are going to focus on the species ordination. Mutual correlations between species are best inferred from a diagram in this scaling. Species with arrows that make a sharp angle are inferred to be positively correlated, with a higher correlation the longer the arrows are. By contrast, obtuse angles reflect negative correlations. By projecting all species points on the arrow for a particular species, the order of the projection points orders their correlation with that particular species (if the scaling is 2; if the scaling is -2, an ordering of covariances is obtained). Scaling 2 also facilitates the quantitative interpretation of species-environment biplots. Both species scores and biplot scores for the environmental variables are correlations with the ordination axes in scaling 2. Thus, one can use the same scale unit in the diagram for both species and environmental variables. By adding a circle with radius 1 to the diagram, the diagram allows easy quantitative evaluation of the correlations between species and environmental variables. Scaling 2 or -2 is also appropriate when one wishes to study how the sample values of an environmental variable change over the ordination diagram (Jongman et al, 1987, Fig. 5.16, equation 5.12). The biplot scores of the environmental variable gives the correct direction of maximum change across the diagram. If you wish this type of diagram with scaling 1, you must divide the biplot scores for environmental variables of that analysis by the corresponding eigenvalue. Although, in scaling 1, the biplot scores themselves are optimal in conjunction with the species, they need modification for use with samples. This unpleasant feature of scaling 1 is the reason why it is not any longer the default in

CANOCO.

Scaling 3 is intermediate between 1 and 2. It does not have any extra mathematical optimality, but may be convenient as a compromise.

If the analysis is centred and standardized by species (PCA/RDA on a correlation matrix), then one may conjecture that positive scalings give the same result as the corresponding negative scalings. However, that is not presently the case, because with this centring/standardization CANOCO sets the species variance internally to 1/m instead of to 1.

3.3 CA/CCA

In the long dialogue for CA and CCA, the user is posed the question:

```
*** Scaling of ordination scores ***
1 = sample  scores are weighted mean species scores
2 = species   ,,     ,, weighted mean sample    ,,
3 = symmetric scaling
Type corresponding negative number for Hill's scaling
Range of valid answers:       -3      [2]      3
```

The positive scalings standardize the ordination scores to λ^{α}, whereas the negative values standardize the ordination scores to $\lambda^{\alpha}/(1-\lambda)$, with $\alpha = 0$, 0.5 or 1. The negative scalings were the only ones possible in CANOCO v2.x. Scaling -1 has the advantage that it equalizes the root mean square species tolerance among axes (Hill, 1979, CAN21 p.46 and Jongman et al, 1987, p. 103). Lebreton et al (1987) use the positive scaling values, scaling 2 in particular. With the positive scalings, the ordination diagrams of CA and CCA can be interpreted as biplots that give approximate values of (transformed) species abundance values (e.g. ter Braak, 1985, eq. 2.4, see Section Ordination diagnostics) and also, more informally, as joint plots. By contrast, with the negative scaling the ordination diagrams can only be interpreted as a joint plot in which mutual distances yield approximate orderings of abundance values (Jongman et al. 1987, section 5.2.5). The difference in appearance of the ordination diagram is small if the eigenvalues are low. The paradox noted in CAN21 (p. 68) is relevant here. A clear advantage of scaling 2 is that environmental biplot scores are correlations of environmental variables with the axes (Lebreton et al, 1987). The resulting ordination diagram thus allows both a more easy intuitive and a more quantitative interpretation.

We now discuss the choice between scaling 1, 2 and 3. Scaling 1 or -1 is more appropriate when the focus is on the configuration of the samples in the ordination: with scaling 1, inter-sample distance approximates their chi-square distance (Jongman et al. 1987, equation 5.15). Scaling 2 or -2 is more appropriate if the emphasis is on the species configuration: with scaling 2, inter-species distance approximates their chi square-distance. With both species points and biplot points for environmental variables in the diagram, scaling 2 or -2 fits well with the idea that this biplot approximates weighted averages of species with respect to environmental variables: the species points are namely at the centroid of the samples in which they occur and are thus also exactly the weighted average of the projection points of the samples on the environmental arrow. As in

PCA/RDA in scaling 2 or -2, the projection points approximate the sample values of the environmental variable. Scaling 2 has, again, the advantage that it yields correlations for environmental biplot scores.

3.4 DCA (segments)

In DCA (segments), there is only one scaling: the original scaling used in DECORANA (Hill, 1979) and also described in Jongman et al (1987, p 106). This scaling is most akin to the above scaling -1 in CA/CCA. In the short dialogue, CANOCO uses the default values for the number of segments (26), the number of rescaling to be done (4) and the rescaling threshold (0.0). These defaults can be modified by using a CANOCO.INI file

3.5 Scaling of ordination scores with covariables in the analysis

Most of what has been said so far continues to hold, if there are covariables in the analysis, except that the correlations are, strictly speaking, not (partial) correlations, but partial covariances. In my opinion, partial correlations are more difficult to interpret, because the scale of species and environmental variables then depends on the covariables in the analysis. (To obtain partial correlations, one needs to divide the partial covariances by the square root of the residual variances for species and environmental variables after fitting covariables). Especially if the residual variances are small, partial correlations are less stable than partial covariances.

4 Check on influence

Constrained ordination is an extension of multiple regression. As in regression, samples that have extreme values in the explanatory variables, have more influence on the results than central samples. This influence can be measured by the leverage (Montgomery & Peck, 1982). The leverage is equal to the squared Mahalanobis distance of the sample plus 1/n and thus measures how extreme the position of the sample is in the space of the environmental variables. If the leverage of a sample is more than three times the average leverage, then CANOCO reports the sample number and how many times the average its leverage is.

CANOCO checks for each sample the leverage
(1) in the space of the covariables
(2) in the joint space of the covariables and environmental variables
(3) for each separate environmental variable.

Check (3) detects univariate outlier and uses a higher cut off point, namely five times the average leverage. This corresponds to samples that have a value that is more than 3 standard deviations out of the mean. This check is skipped for indicator variables (0/1-variables). There is an easy formula to transform univariate leverages to standard deviation units: if the leverage is k times the average, the value is sqrt(2*k-1) standard deviations out of the mean.

What to do if samples with high influence are detected? The first thing is to check that the cause is not a recording or typing error. If not, try to understand why the sample is an outlier and whether it really belongs to the population you want to describe. If it does, it may be instructive to check whether removal of the sample would modify your essential conclusion. But, always be hesitant to remove the sample in the analysis you report. More discussion on this topic can be found in every modern book on regression and on outliers.

5 Ordination diagnostics

5.1 Introduction

Usually, an ordination diagram is not an exact representation of the data. Overall measures of quality of the approximation are given in the 'Summary of the ordination' in terms of percentages of variance accounted for. But, neither all species nor all samples are equally well represented in the data. CANOCO has ordination diagnostics to find out which species and which samples are ill-represented and which are well represented. There are three types of statistics: measures of fit for species, residual distances for samples, tolerances for species ('niche widths') and heteregeneity for samples. Tolerance and sample heterogeniety are not defined in PCA/RDA. The fit measure and residual distance are not available in DCA (segments).

Ordination diagnostics are also of interest to see whether passive samples (e.g., historic samples) fit into the structure found for the active samples (e.g., modern samples).

5.2 Fit for species and residual distances

In the long dialogue, CANOCO poses the question, in CA/CCA,

```
*** Species and sample diagnostics ***
0 = no diagnostics
1 = Chi-square- fit and residual distances
2 = tolerances
3 = both 1 and 2
Range of valid answers:        0       [3]
```

Here we describe option 1. The corresponding question in PCA/RDA is:

```
*** Species and sample diagnostics ***
0 = no diagnostics
1 = fit and residual distances
Range of valid answers:        0       [3]
```

(The default value of 3 yields the same result as the answer 1).

Example output for CCA for species is:

```
 DUNE MEADOW SPECIES DATA (M. BATTERINK AND G. WIJFFELS, 1983)
 CCA  Canonical axes:  4  Covariables:   0  Scaling:  2

No transformation
 CFit: Cumulative fit per species as fraction of variance of species

   N        NAME      AX1        AX2        AX3        AX4        VAR(y) % EXPL

     FR FITTED         .2180      .1409      .0757      .0632

   1  ACH  MIL        .3252      .3923      .3926      .4441       2.17      49.35
   2  AGR  STO        .5065      .7199      .7311      .7365       1.17      78.20
   3  AIR  PRA        .0383      .2624      .3437      .3635      14.26      37.32
   .
   .
   .
  27  TRI  REP        .0009      .0222      .1816      .2091        .57      45.46
  28  VIC  LAT        .0416      .1904      .2050      .2718       6.93      66.67
  29  BRA  RUT        .0293      .1315      .1622      .2430        .63      34.95
  30  CAL  CUS        .3698      .3972      .4173      .4259       7.42      48.64
```

The column headed VAR(y) contains the variances of each of the species. In CA/CCA, this is the chi-square statistic (Greenacre, 1984, eq 2.4.2), divided by y_{+k}, calculated for each species:

$$\begin{aligned}\mathrm{var}(y_k) &= \Sigma_i\, (y_{++}/y_{i+})(\, y_{ik}/y_{+k} - y_{i+}/y_{++}\,)^2 \\ &= \Sigma_i\, (y_{i+}/y_{++})(\, (y_{ik} - e_{ik})/e_{ik}\,)^2 \\ &= y_{+k}^{-1}\, \Sigma_i\, (y_{ik} - e_{ik})^2/e_{ik}\end{aligned}$$

where $e_{ik}=y_{i+}y_{+k}/y_{++}$. Note that

$$\text{total inertia} = \text{chi-square} / y_{++} = \Sigma_k\, (y_{+k}/y_{++})\, \mathrm{var}(y_k)$$

From the species scores (b_{ks}) and eigenvector sample scores (x_{is}) on axis s one can calculate the fitted value for y_{ik} as follows, if the scaling < 0 (for CA: Greenacre, 1984, eq. (4.1.25); ter Braak, 1983; analogous for CCA)

$$f_{ik} = y_{i+}y_{+k}/y_{++}\ (\ 1 + b_{k1}\, x_{i1} + b_{k2}\, x_{i2} + \ \dots\ )$$

The fraction of the variance of a species fitted, in this way, by axis 1 is given under the heading AX1. (This fraction is sometimes termed the contribution of dimensions to the inertia of the species, or the relative contribution; Greenacre, 1984, p.70). The fit by axes 1 and 2 is given under the heading AX2, etc. The percentage fit by all q environmental variables together (q axes) is given in the last column, headed % EXPL.

Due to internal rescaling of the data (CAN21, p.30, pp. 70-72), the variance given in PCA/RDA is not the usual variance but proportional to it. The proportionality constant depends on the total sum of squares and the number of samples (CAN21, p.71). The cumulative fit for s axes is the coefficient of determination (R^2) in a regression model with the s axes as explanatory variables: it is the cumulative fraction of the variance explained by the first s axes. The percentage variance accounted for by all environmental variables is given in the last column (% EXPL = 100*R^2).

With covariables in the analysis, VAR(y) is unchanged. All fractions are therefore with respect to the original variance, rather than with respect to the residual variance. The fit due to the current environmental variables is shown. This fit is additional to the fit by the covariables.

Species influence the ordination more the larger their variance and the larger their weight. In PCA/RDA, the weights are usually equal and it is sufficient to look at the species variance. Species with extreme variance may have an unduly large influence. A remedy is to transform the species data by, for example a log or square-root transformation. If that does not help enough, consider given a species less weight in the option for weights for species in the long dialogue. In CA/CCA, species with a large value for weight*variance may have a large influence.

Example output for samples is:

```
 SqRL: Squared residual length per sample with s axes (s=1...4)

   N        NAME      AX1        AX2        AX3        AX4         SQLENG  % FIT

     FR FITTED         .2180      .1409      .0757      .0632

   1    ......1      2.4239     2.3515     2.2302     2.1596       3.06     29.38
   2    ......2       .8389      .8579      .8270      .4398       1.17     62.37
   3    ......3       .9140      .6169      .4876      .4337        .96     54.69
```

In CA/CCA, SQLEN is the squared chi-square distance between the sample point and the centroid in m-dimensional species space (Greenacre, 1984, p. 35; the formula is analogous to that of the variance of species in CA/CCA). After fitting s axes, the squared distance between the sample point and the s-dimensional ordination space is given under the heading AXs (s=1,...,4). The percentage fit (% FIT) is (within rounding error) equal to 100 * (1 - entry AX4/SQLEN).

In PCA/RDA, squared Euclidean distances are shown. These are proportional to the ones calculated from the raw data, because the total sum of squares in the species data is set to 1 in CANOCO (CAN21, p.30, pp. 70-72).

The percentage fit for samples (% FIT) can take negative values in constrained analyses. Then, the residual length is larger than the length, i.e. the sample point is farther from the ordination plane than from the centroid of the data. This can happen when there is a strong species-environment relation, but an odd sample couples an almost 'average' species composition to marked environmental values.

5.3 Tolerance and sample heterogeneity

CA/CCA/DCA are based on the assumption that the species 'distribution' (the response function) is unimodal. (That is at least one way of looking at CA/CCA/DCA). The species score, which (proportional to) the weighted mean of the sample scores, indicates the centre of this distribution. The width of the distribution can, similarly, be quantified by the standard deviation (Chessel et al, 1982), or as I prefer to term it, the tolerance (ter Braak and Barendregt, 1986; ter Braak and Looman, 1986; ter Braak and van Dam, 1989). The tolerance is a measure of niche width. Green (1971) proposed this niche measure in his variant of discriminant analysis. His analysis is equivalent with CCA applied to presence-absence data (Chessel et al, 1987; Lebreton et al, 1988). Green (1971) is thus a precursor to CCA. After Green's paper appeared, a series of papers in Ecology discussed niche measures in canonical space (Dueser and Shugart, 1978; Dueser and Shugart, 1979; Carnes and Slade, 1982; Van Horne and Ford, 1982; Dueser and Shugart, 1982). CANOCO follows the round up by Carnes and Slade (1982) in providing standard deviations of scores per axes (see Green, 1971, Fig.2) and the root mean square standard deviation across the 4 axes (RMSTOL) as a summary niche breadth. The population standard deviation is used (divisor n instead of n-1). An example for the CCA is:

```
Tol : Species tolerance (root mean squared deviation for species)

   N        NAME      AX1       AX2       AX3       AX4        RMSTOL     N2

     FR FITTED        .2180     .1409     .0757     .0632

   1  ACH  MIL       .3702     .7193     .8210    1.0546      78.11      6.10
   2  AGR  STO       .8635     .9474    1.1364     .8557      95.75      9.14
   3  AIR  PRA       .5708     .2919     .4588     .0326      39.45      1.92
```

When the scaling is not 2 or -2, species scores are not weighted averages of sample scores. Then, CANOCO uses the species score instead of the weighted mean in calculating the standard deviation. CANOCO thus calculates the spread around the species point of its occurrences. When the scaling is 2 or -2, this is fine also for interpreting the ordination diagram in terms of where the species may occur. But, for a fair statistical comparison of the standard deviations, one should take into account the effective number of occurrences: if N2 is close to 1, then the standard deviation is always zero! One may account for this, by dividing the standard deviation by sqrt(1- 1/N_2). For presence-absence data, one so obtains precisely the sample (instead of: population) standard deviation.

The last column contains N2, which is the effective number of occurrences of the species. It is analogous to the N2-diversity measure of Hill (1973b). N2 can be understood as follows. For presence-absence data, N2 is simply the number of occurrences. With abundance data, a species may occur with abundances 1000, 1, 1, say. CA/CCA/DCA are based on weighted averages. The weighted average for this species is effectively determined by the sample in which it occurs with abundance 1000 and the value of N2 is close to 1.

For samples, one can calculate the same measures of spread as for species. One obtains, for example, in our CCA,

```
Het : Sample heterogeneity (root mean squared deviation for samples)
   N      NAME      AX1       AX2       AX3       AX4      RMSTOL    N2

     FR FITTED     .2180     .1409     .0757     .0632

   1    ......1    .3524     .3640    1.1338     .1370    62.47     3.77
   2    ......2    .7291     .4110     .2346    1.4485    84.46     9.09
   3    ......3    .5317     .6950     .8218     .6445    68.13     8.25
```

The measure N2 is now the inverse of the Simpson diversity measure.

The numbers after FR FITTED are the fraction explained variance of the species data, which we given earlier cumulatively in the Summary of ordination.

In the evaluation of ordination diagnostics of passive samples, one should be aware of the following. If a active and passive sample have the same species composition (as sample 17 and PAS 17 in DUNE.SPE), then their diagnostics will be identical only in an unconstrained analysis. This is because CANOCO will always use, for the active sample, the eigenvector sample scores, but, for the passive sample, the sample scores which are derived from the species. In a constrained analysis, the eigenvector sample scores are those derived from the environmental variables, whence the difference. With covariables in the analysis, an addition difference can be whether or not values for covariables were entered for the passive sample.

Note: The 'eigenvector sample scores' are, in an unconstrained analysis, the sample scores (which are derived from the species) and, in a constrained analysis, the sample scores which are linear combinations of environmental variables.

6 Forward selection of environmental variables

The purpose of selection of environmental variables is to find out a minimal set of variables that explain the species data about as well as the full set. Selection of variables is a standard topic in books on multiple regression, e.g. Montgomery and Peck (1982). CANOCO generalizes forward selection of variables from univariate regression to the multivariate case. See also Escoufier & Roberts (1979). At each step, the variable is selected that adds most to the explained variance of the species data. The explained variance is a straight sum of squares of regression in RDA and is inertia in CCA (see Summary of the ordination). With CANOCO, one can test at each step whether the variable to be added is statistically significant by means of a Monte Carlo permutation test. This test replaces the F- or t-test in forward selection in univariate multiple regression and shares the shortcomings of these tests: when applied stepwise, the tests do not control in any way the overall size of the test. In practical terms, this means that too many variables will be judged significant.

Forward selection can be chosen in the long dialogue (or by setting this option in the CANOCO.INI file).

Let us give an example using a CCA of the Dunemeadow data. At the point where CANOCO normally starts with the calculation of the eigenvalues, CANOCO now gives:

```
**** Start of forward selection of variables ****

   N     Name Extra fit
   6  PASTURE       .10
   5 HAYPASTU       .13
   8       BF       .14
   9       HF       .15
   4 HAYFIELD       .15
   7       SF       .20
   1       A1       .22
   3   MANURE       .24
  10       NM       .32
   2 MOISTURE       .41

Type number of variable to be selected
 "   -number to test the variable
 "   -999     to test the best variable
 "    0       to stop forward selection
Range of valid answers:     -999       [2]       10
```

The environmental variables are shown in order of the 'Extra fit'. With no variable yet selected, the extra fit is equal to the eigenvalue of a CCA if the corresponding variable were the only environmental variable. The same list could thus be obtained manually in ten runs of CANOCO, each run with another environmental variable. If we press RETURN, CANOCO selects the variable with the highest extra fit, in our case, MOISTURE and reports:

```
Variance explained by the variables selected:      .41
   "          "       "        all variables     :    1.22
```

The value 1.22 is the sum of all constrained eigenvalues (See summary of the ordination), which is the total variance (inertia) of the species data that is explained by all ten environmental variables. Moisture alone explains an inertia of 0.41. CANOCO continues with:

```
  N     Name Extra fit
  8       BF       .08
  6  PASTURE       .08
  9       HF       .11
  1       A1       .12
  5 HAYPASTU       .13
  4 HAYFIELD       .15
  7       SF       .18
  3   MANURE       .23
 10       NM       .26

Type number of variable to be selected
 "   -number to test the variable
 "   -999     to test the best variable
 "    0       to stop forward selection
Range of valid answers:     -999      [10]
```

With moisture already selected, the extra fit is the increase in explained inertia when the analysis with moisture alone is compared with the analysis with both moisture and the corresponding variable. The value of 0.23 for MANURE could thus be obtained manually by running a CCA with MOISTURE and MANURE. The explained inertia (the sum of all constrained eigenvalues) of this CCA is 0.64, which is 0.23 more than with MOISTURE alone.

If we now for some reason wish to add MANURE to the model, we type 3 and press RETURN. CANOCO selects MANURE, instead of the variable with the highest extra fit, and reports:

```
Variance explained by the variables selected:     .64
   "          "      "      all variables    :   1.22

     The inertia explained by MOISTURE and MANURE is, indeed, 0.64

  N     Name Extra fit
  5 HAYPASTU       .05
  7       SF       .06
  8       BF       .09
  6  PASTURE       .09
  4 HAYFIELD       .10
 10       NM       .11
  9       HF       .11
  1       A1       .13

Type number of variable to be selected
 "   -number to test the variable
 "   -999     to test the best variable
 "    0       to stop forward selection
Range of valid answers:     -999      [1]      10
```

We see that NM is not the best variable to add at this point. This can be explained by noting that the Nature Management meadows do not receive manure, so that the variable NM is largely exchangeable with MANURE. The best variable to add is now A1. Let us test whether the additional effect of this variable on the species is statistically significant. We do this by typing either -1 or -999. The questions CANOCO asks now are described in the section on Monte Carlo testing. After the requested 99 random permutations of the values of the variable A1, CANOCO summarizes the test as follows:

```
P-value   .090 (variable   1; F-ratio=  1.55; number of permutations=   99)
```

and asks again whether to select or test a variable or to stop. Because the additional effect of the best variable (A1) is not significant at the conventional 5%-level, we might wish to stop adding more variables. We do this by answering 0. If there is enough data space, CANOCO continues by performing a CCA on the selected variables (variables 2 and 3). Before doing this, CANOCO reports that the other environmental variables are omitted. Variables that are multicollinear with the selected variables will not be omitted, because they do not harm the subsequent analysis. This feature of CANOCO guarantees that if 2 dummy variables of a nominal variable with 3 classes are selected, the third one is automatically included in the subsequent analysis.

It is no problem for CANOCO, if there are covariables in the analysis at the start of the forward selection. The extra fit is calculated as usually. The results can be mimicked manually (without forward selection) by running CANOCO with the same covariables and the variable under consideration as environmental variable. The additional effect of the variable can then be tested also with a Monte Carlo permutation test (after a request for more analyses with the current data). The permutation results are not exactly the same, however, because the sequence of random permutation differs between the two ways of obtaining a Monte Carlo test with CANOCO. With forward selection, the Monte Carlo test is much quicker.

Technical note

Each time an environmental variable is selected, it is transferred to the covariables. The number of covariables may therefore be reported to be too small during forward selection, even if you did not enter yourself any covariable.

7 Monte Carlo permutation test

The methods used for the Monte Carlo test have been changed. In CANOCO version 2.x, the method is based on the null hypothesis of exchangeability of sample values of the environmental variables (as used for example by Collins, 1987). In CANOCO version 3.10, the method uses exchangeability (whence permutation) of the residuals of the species after fitting covariables and environmental variables. The latter method is closely related to the bootstrap Monte Carlo tests proposed by the Hall & Titterington (1989). Instead of bootstrap samples from the residuals, CANOCO uses permutations of the residuals. It can be shown that this gives difference in results of the order 1/n (ter Braak, 1990b). For the rest, the methods are the same. The CANOCO method is based on the randomization model that Kempthorne (1952) poses in his "Design and analysis of experiments". The advantages of the version 3.x method are:

- interactions effects (e.g. product variables) can be validly tested.

- the correlation structure of the explanatory variables (covariables and environmental variables) is not changed during permutation.

- more power by permutation of residuals under the full model or, equivalently, the alternative hypothesis (Hall & Titterington, 1989).

Furhter, the test statistics is changed from a sum of squares (eigenvalue or trace) to an F-type criterion by dividing by the residual error. The advantage of using an asymptotic pivotal statistic is shown by Hall & Titterington (1989).

Permutation of residual under the null model (the null hypothesis) is optional. With this option, the old and new method yield equivalent results in simple situations such as overall tests without covariables and tests with a conditioning on all covariables (randomized block designs). With permutation under the null model, the test is less dependent on the model being analyzed. For example, in a randomized block design with conditional permutation, no additive block effect needs to be assumed for validity of the test. On the negative side, the test is less powerful this way.

Permutation under the null model is the default in tests of the first canonical axis only, i.e. when the overall test is not required. This is because, with permutation under the full model, the alternative hypothesis is not one-dimensional, but q-dimensional.

Technical note

The F-ratio for the overall test is defined as:

$$F = (\mathrm{trace}/q)/(\mathrm{rss}/(n-p-q-1))$$

with trace the sum of the canonical eigenvalues, rss the residual sum of squares, q the number of environmental variables and p the number of covariables and n the number of samples. The residual sum of squares is, for the data without permutation,

rss = sum of all unconstrained eigenvalues - trace

For the data after permutation, the rss is the residual sum of squares in the multivariate regression model:

$$Y^{+} = Z_1 M_1 + Z_2 M_2 + E$$

with Y^{+} the permuted residuals. The trace is the sum of squares due to Z_2, after fitting Z_1.

The F-ratio for axis 1 is

$$F = \lambda_1 / (rss/(n-p-q-1))$$

with rss = the sum of all unconstrained eigenvalues - λ_1. During permutation, the rss is that of the above multivariate regression model with a rank 1 restriction on the matrix of regression coefficients (M_2).

In CCA, the regression model is slightly different. The data Y are the residuals under the independence model in a contingency table

$$r_{ik} = y_{ik} - y_{i+}y_{+k}/y_{++}$$

and the regression uses sample (row) weights y_{i+} and, for sums of squares over species, species (column) weights y_{+k}.

8 Permutation types

8.1 Introduction

The validity of permutation test hinges on the validity of the type of permutation for the particular research design at hand. For completely randomized designed experiments (Cox, 1958), completely random perputation is appropriate, whereas for randomized block design the permutation must be conditioned on the blocks. Data from line transects, time series, rectangular grids and repeated measurement studies (e.g, BACI-designs) require specialized permutation types. CANOCO 3.10 can automatically generate valid permutation types for such data, when recorded at equal intervals. If your data require another type, you can feed permutations from an external file into CANOCO. For example, Legendre et al (1990) propose, for one-way (M)ANOVA tests, a permutation type for data from an unregular grid. Permutations generated with their program COCOPAN can be fed into CANOCO. Permutation tests for time series data and spatial data as performed by CANOCO, form a nonparametric way to overcome the difficulty of statistical tests in the presence of autocorrelation or spatial correlation (Besag & Clifford, 1990: section 5; Ter Braak, 1980: part II, chapter 3). They thus form a viable alternative for traditional parametric tests based on precise modelling of the autocorrelation structure.

8.2 Without covariables

Without covariables in the analysis, CANOCO poses the question:

```
*** Type of permutation ***
0 = permutations read from file
1 = unrestricted
2 = restricted permutation for time series, line transects or grids
Range of valid answers:       0      [1]      2
```

Unrestricted permutation yields completely random permutations. If the data are from a time series, line transect or a rectangular spatial grid, CANOCO can generate more appropriate permutations. The idea is as follows. For stationary time series (with the sample points at equal time intervals), two series are unrelated if the starting point of the one serie is randomly linked to a time point of the other serie. So, the null hypothesis is rejected if the observed correlation is extreme in the distribution of correlations generated by such random links. We still need to face the problem that, after random linking, the start of the second serie has no first serie's points linked to it. Similarly, the end of the first serie has no linked points. Rather than using the linked points only, we use the tric of bending the time series into a circle, so that start and end meet. This mathematical tric works fine, provided there is no trend; it only corrupts the autocorrelation structure of each series at beginning and end of each series. For line transects, the dependence stucture is not unidirectional as in time series. Usually, a point is related to its neighbours at both sides. Therefore, each observation serie along the transect can also be mirrored (the series of points 1, 2, 3, 4 and 4, 3, 2, 1 are statistically equivalent). However, the distinction between line transects and time series is not essential here. The test statistic used in CANOCO is correlation-based and the autocorrelation at lag h is equal to that at lag -h. Under the null-hypothesis, a trend-free time series can therefore be mirrored also. The general idea is that, with a correlation-based test statistic as is used CANOCO, the test of association must use permutations which preserve marginal correlations, but change cross-correlations (Ripley, pers. comm.).

The general idea can be applied to data on a rectangular grid (with equal horizontal and vertical spacing). Data on a rectangular grid are wrapped around a torus (so that opposite sides meet) and the points on the torus for the species data set are randomly shifted with respect to the points on the torus for the environment data. If there is no trend, the grid can be rotated 180 degrees without changing the autocovariance function (i.e the autocovariance function c(h) equals c(-h), where h is the shift $h=(h_1,h_2)$). Therefore, both sides of the grid can also be mirrored before the shift (grid D below). If the autocovariance function is symmetric ($c(h_1,h_2)=c(-h_1,h_2)$), we may mirror either one of the sides. Then, the following four grids have the same correlation structure and random shifts can be made, starting from each of the four equivalent grids:

```
        A                  B                  C                  D

 1  2  3  4        17 18 19 20       4  3  2  1        20 19 18 17
 5  6  7  8        13 14 15 16       8  7  6  5        16 15 14 13
 9 10 11 12         9 10 11 12      12 11 10  9        12 11 10  9
13 14 15 16         5  6  7  8      16 15 14 13         8  7  6  5
17 18 19 20         1  2  3  4      20 19 18 17         4  3  2  1
```

Isotropic spatial processes have a symmetric covariance function.

The question that CANOCO poses to choose among these possibilities is:

```
Type number of rows of the rectangular grid
type 1 for time series and line transects
    -number to disable random shifts of the mirror image
Range of valid answers:     -10       [1]      10
```

For time series and line transects, the appropriate answer is 1. If the data are from a grid, e.g. with sample numbers and layout as in the left most grid above, the answer must be the number of rows (5). The general rule for determining what CANOCO considers rows, is that consecutive samples in the same row must have consecutive sample numbers. By default, CANOCO assumes that the autocovariance function is asymmetric and generates random shifts starting from grid A and D only. This default can be changed in the CANOCO.INI file (option 23). If this option is set to one, shifts starting from all four grids are used.

For the above tests to work, the series need to be trend-free. It is therefore wise to linearly detrend the series before the test is applied. This is done in CANOCO by using covariables. For time series, use the time as covariable, for line transects the position and for grids both spatial coordinates of the sample (i.e. two covariables: one for the horizontal and the other for the vertical position).

8.3 With covariables

With covariables in the analysis, CANOCO poses the question:

```
*** Type of permutation ***
0 = permutations read from file
1 = unrestricted
2 = restricted permutation for time series, line transects or grids
3 = permutation within blocks
4 = permutation for repeated measurements, e.g. BACI designs
Range of valid answers:       0       [1]      4
```

When there are covariables, there are two more possibilities. Type 3 was available in CANOCO 2.x as well (CAN21, section 4.11). After choosing the types 2 or 3, the next question is:

```
*** Specification of blocks ***
Type the number of covariables
onto which the permutation must be conditioned
Range of valid answers:      [0]      3
```

With the type of farms as covariables (SF, BF, HF, NM), the answer 3 garantees that farm type as treated as a block. CANOCO reports which samples belong to each permutation group (block), so you can check whether CANOCO does what you wished. If the permutation type is 2 (time series, line transects or spatial grids) and the only covariables are time/spatial coordinate(s), answer 0 here. With a nonzero answer, CANOCO continues to ask, for each group, for the number of rows in the layout of the samples, as above. To avoid confusion, number samples of the same time series, line transect or grid consecutively.

In repeated measurement designs (type 4), each unit must have been recorded the same number of times. Consecutive samples in time must be given consecutive numbers in the input files. After choosing type 4 (repeated measurement designs), CANOCO asks:

```
*** Specification of block + time variables ***
Type the number of covariables
onto which the permutation must be conditioned
Range of valid answers:      1      [3]
```

CANOCO reports (with answer 3), for example:

```
Permutation class   1 (block   1, time   1) contains the samples numbered:
    1    6   11   16
Type number of times these units are sampled
Range of valid answers:      [1]      5
```

If we answer 5, then the design is fully specified: there were 4 units, each sampled 5 times. The units belonging to time 2, 3, 4, and 5 will be reported by CANOCO. With the answer 2, CANOCO will notice that there are remaining samples. There must therefore be a block 2 in the design and CANOCO will ask how many times the units of that block were sampled.

With type 4, CANOCO randomly permutes the labels of the units: the samples at different time points of a unit are hold together. Permuting the statistical units (instead of the individual time samples) garantees valid statistical inference.

8.4 Example of BACI-design

An application of this type of permutation is given in the files BACI.* on the distribution disks (Data from Manger and Schouten, 1989). A liming experiment is carried out in three forests. In each forest, there are six plot, each recorded four times (one time before and three times after the treatments are applied). Recorded are percentage abundance of nematodes in four food-groups (data in BACI.SPE). The treatment is the application of three doses of lime: 0, 3 and 9 ton/ha lime. The experiment is thus of the BACI-design: Before-After-Control-Impact design (Green, 1979). Not all BACI-experiments are repeated measurement studies, but probably many are, because often the same locations are resampled instead of that a new random sample is drawn from each area. The model for statistical analysis is:

abundance = plot effect + time effect + lime effect + error

We are interested in the lime effect. Plot and time must therefore be eliminated from the analysis by making them covariables (file BACI.COV), and lime must be in the environmental data (file BACI.ENV). For the Monte Carlo test, it is important to note that, under the null hypothesis of no lime effect, only the plots within each forest are exchangeable among each other. Forests thus comprise blocks. That is why the forest indicator variables are added in the file BACI.COV. Block and time variables are the first 7 covariables. CANOCO will note some dependencies among the covariables; one forest and one time variable will be removed, leaving 5 independent block and time variables. (Also three plot indicators will found dependent). After choosing permutation type 4, we specify to condition the permutation on the first five covariables. For the number of times each sample of the first block is sampled we answer 4. For the second and third block we answer the same (see file BACI.CON). Neither an analysis with CCA nor with RDA on double centred log-data shows a significant lime effect. By contrast, there seems to be a time effect, which is easily mistaken to be a liming effect!

If your design requires a more specialized permutation type than provided by CANOCO, you can enter the permutations from file. After choosing permatution type 0, CANOCO asks for the file with the permutations. An example file is DUNE.PER. The permutations can be in free format, each one starting at a new line. If there are n active samples, the numbers 1 to n must be permuted. The numbers correspond to the first n samples listed in the samples scores on, for example, the solution file (even if these samples have other identifying numbers). To avoid confusion it is therefore safe to make samples numbers consecutive. It is NOT permitted to specify a bootstrap sample from the numbers 1 to n; the CANOCO algorithms do not allow this. CANOCO will detect an error if the values read do not form a permutation.

9 Redundancy analysis with a statistical weighting of species

In the long dialogue for RDA, CANOCO poses the question:

```
*** Centring/standardization by species ***
0 = none          (non-centred PCA)
1 = centring      (for PCA/RDA on a covariance matrix)
2 = standardization by species norm
3 = both 1 and 2 (for PCA/RDA on a correlation matrix)
4 = standardization using error variance
Range of valid answers:       0       [1]      4
```

The fourth option is new in CANOCO version 3. A disadvantage of RDA compared to canonical correlation analysis is that the result depends on the particular units of scale of measurement for each response variable (species). On the other hand, canonical correlation analysis is unattractive when the number of species is of the same order of magnitude as the number of samples. An intermediate solution, proposed in the discussion of Ter Braak (1990a), is to weight each species inversely to its error variance. CANOCO now incorporates this and reports the relative weights given to species on the output file, in the solution file (as weight for species alongside the species scores) and, if requested, CANOCO.PUN. If the R^2 of a species exceeds 0.9, than its weight is truncated as were the R^2 equal to 0.9. This is done to avoid extreme weights (larger than 10) for species that happen to fit extremely well.

Technical details

If standardization using error variance is requested, CANOCO first centres and standardizes the species as if option 3 was chosen. For the species data so standardized, CANOCO regresses each species on to the environmental variables to obtain the error variance. The reported variances of species are therefore all equal with this option.

The weights given to species are not reestimated in permutations for a Monte Carlo test.

CANOCO uses the error variance in the full rank model. By contrast, Van der Leeden (1990) uses the error variance from the reduced rank model. The advantage of the CANOCO approach is that it does not depend on the reduced rank assumption and that the CANOCO solution is much simpler, in casu non-iterative.

10 Biplot of t-values of multivariate regression coefficients

RDA is a form of multivariate multiple regression (CAN21, section 7.3). To the principal results of a regression analysis belong regression coefficients and associated t-values (Jongman et al, 1987, chapter 3). RDA attempts to represent the regressions of all species jointly in a low-dimensional space. RDA can thus yield not only low-dimensional approximations to fitted values (see Summary of ordination and Ordination diagnostics), but also low-dimensional approximations to the regression coefficients (CAN21, section 7.3) and their t-value (Ter Braak, 1990a, Fig. 2).

An example result for an RDA on the Dunemeadow data is as follows:

```
 DUNE MEADOW SPECIES DATA (M. BATTERINK AND G. WIJFFELS, 1983)
 RDA  Canonical axes:  4  Covariables:   0  Scaling:  2
Cent./stand. by samples:  0  0 by species:  1  0
No transformation
 StBi: Species coordinates for t-value biplot

   N       NAME      AX1        AX2        AX3        AX4        VAR(y)  % EXPL

              EIG    .2644      .1701      .0671      .0413

   1  ACH  MIL      -.5379      .0969      .0000      .0000        .55     59.88
   2  AGR  STO       .3124     -.2560      .0000      .0000       2.57     69.59
   3  AIR  PRA       .4829      .8740      .0000      .0000        .22     43.11
   .
   .
```

```
 DUNE MEADOW SPECIES DATA (M. BATTERINK AND G. WIJFFELS, 1983)
 RDA  Canonical axes:  4  Covariables:   0  Scaling:  2
Cent./stand. by samples:  0  0 by species:  1  0
No transformation
EtBi: Environmental coordinates for t-value biplot

   N       NAME      AX1        AX2        AX3        AX4

              EIG    .2644      .1701      .0671      .0413

   1          A1     .0234     -.0789      .0000      .0000
   2    MOISTURE     .4490     -.3396      .0000      .0000
   3      MANURE    -.0738     -.1661      .0000      .0000
   .
   .
```

The output is tailored to a biplot in two dimensions. The dimension is set in the CANOCO.INI file (the default dimension is 2). These biplot scores are given by CANOCO when the user asks for output of the t-values. (The t-values of the regression/canonical coefficients (CAN21, Table 4.7) of the species axes on to the environmental variables are given together with the regression coefficients). The scores given for species and environmental variables must be plotted on the same scale. The interpretation of the plot is as Fig 2 in Ter Braak (1990a) and can be summarized as follows.

The points for the environmental variables can be projected on the arrow for a species. If the projection point falls precisely on the head of the species' arrow the approximated t-value is 2. If it falls on other side of the origin of the coordinate system at the same distance, the t-value is -2. If the projection points fall closer to the origin, the t-value is less than 2 in absolute value. Points farther away indicate t-values that are larger than 2. The regression coefficients of the corresponding environmental variables are inferred to be significantly different from 0.

The heads of the species arrow thus determine a natural unit of measure measurement for t-values, namely 2 (the critical t-value at the 5% significance level, provided the number of degrees of freedom exceeds 20).

The species scores and the 'regression/canonical coefficients for standardized variables' can be added to the plot (see Fig. 2 in Ter Braak 1990a) in the same unit of scale (at least with scaling 2). Jointly they approximate the regression coefficients for standardized variables (CAN21, section 7.3).

With PCA, the same plots can be made. The fit to the regression coefficients and t-values is, however, worse in PCA than in RDA.

With CA/CCA, the regression coefficients and t-values are those from the weighted regression described at the end of the section on the Monte Carlo permutation test.

With covariables in the analysis, the regression coefficients are partial regression coefficients (given in the matrix M_2 in the Monte Carlo Section).

11 Detrending

The promise of detrending-by-polynomials was shown to be false by Knox (1989) and Ter Braak (unpublished conference contribution). For the artificial data sets generated and analyzed by Minchin (1987), detrending-by-segments performed consistently better than detrending-by-polynomials. (For a reasonable performance of DCA for these data sets, a log-transformation was essential). As a result, detrending-by-segments is again the default in DCA. In DCCA, detrending-by-segments may cause numerical problems. If so, remember that detrending is not needed in a constrained analysis if the set of environmental variables is reduced to the essential ones (Ter Braak & Prentice, 1988).

12 Choice of method

In addition to what is said about this on page 17-18 of the manual, an important aspect is also that in CA/DCA/CCA the focus is on relative abundance (given the sample total), whereas in PCA/RDA is focus is on absolute abundance. If an environmental variable influences the total biomass, but leaves the species composition (relative abundance) unchanged, the variable will be important in PCA/RDA, but not important at all in CA/DCA/CCA. An idea for analysis is to analyze total biomass separately by regression. A completely different aspect of the data, namely the species composition, is subsequently analyzed by CCA. The analyses are fully complementary. By contrast, a default PCA/RDA would probably yield much the same results as the regression analysis on total biomass. CA/CCA are not unique in analyzing composition, as shown in the manual in section 7.1: double centred PCA/RDA on log transformed values also analyzes composition, but only if there are no zero values.

13 CANOPLOT

CANOPLOT makes print plots from the solution file produced by CANOCO. It is much less sophisticated than CANODRAW. It does, for example, not produce graphical output on the screen. But it may still be of interest to you for routine plots with many names.

Most of the questions speak for themselves, so after running CANOCO (and producing a solution file, e.g. CANOCO.SOL) you can just try to run this program.

A question that may give some difficulties is the question about the scaling of the plots: In most cases the defaults will do, so pressing RETURN in answer to the questions about the scaling is the easiest way to obtain reasonable plots. The scale of the plots is defined by the maximum range of scores to be plotted. Usually, this range is set different for each set of scores (each indicated by a scale number, e.g. 1 = range of species scores). There are a number of cases in which plots can be best superimposed if they have the same scaling:

1. Species and samples in DCA, CA and CCA. This is termed the joint plot scaling in CANOPLOT (scale number 8).
2. Centroids of nominal environmental variables superimposed on sample scores. (That is because the centroids are average sample scores).
3. Species scores and biplot scores for environmental variables in scaling 2 of PCA/RDA. In this scaling the scores are correlations (without covariables) and always lie between -1 and +1. By default, the plot is scaled to these limits.
4. A plot for advanced users with regression coefficients, species and environmental scores for the t-value biplot. The projections of species on to environmental vectors that end precisely on the environmental point indicate that the t-ratio of the regression coefficient of that variable for that species is approx. equal to 2. For this to be true, the scores must of course be in the same scaling. In scaling 2, the scaling is from -1 to +1 to enable the plot to be combined with the plot of the previous point. Regression coefficients might fall outside these limits. If they do, there is probably a high multicollinearity among the environmental variables (check the variance inflation factors) so that this plot is useless.

By default CANOPLOT uses all 8 letters of a name in the solution file, but the user may choose to use only a few letters, e.g. 3. CANOPLOT then uses the first 3 letters of the name. This is useful if the plot is crowded or, if the first few letters of a name represent a higher grouping, to highlight this grouping in the plot.

You can manually delete lines with axes scores from the solution file without affecting the working of CANOPLOT, for example to delete rare species or species with a low fit. If all scores are positive (say), which may happen sometimes, CANOPLOT does not draw axes. For producing a plot with axes lines (i.e. with an origin), it suffices to add a item with a blank name field with negative scores.

CANOPLOT attempts to read a CANOCO.CON in the current directory to fetch the name of a solution file. This name then becomes the default solution file instead of CANOCO.SOL. Next, CANOPLOT searches for a CANOCO INI-file in the same way as CANOCO does. In this file, CANOPLOT searches for a CANOPLOT paragraph. This paragraph looks like the following:

```
*CANOPLOT (values start in position 2)
 1  = range [0,1]   = (01) pagemode of screen
 25 = range [10,100]= (02) number of lines on a screen
 8  = range [1,8]   = (03) number of characters of names to plot
 12 = range [0,999] = (04) characternumber so that CHAR(no) yields a newpage
 118    = number of characters per line (maximum: print line length)
 0.145 = width of print character in cm
 0.423 = height of print character in cm
 CANOPLOT.PLT                                  = output file for  print plot
*ENDCANOPLOT
```

The width and height of a printed character require special attention. Ordination diagrams should have the same scale on the horizontal and the vertical axis. To ensure that this results in the same scale unit on paper on your printer you may have to adapt the width and height values. The default values of 0.145 and 0.423 are appropriate for an EPSON condensed mode print with noncondensed line spacing. CANOPLOT automatically initializes your printer in the way (in the DOS-version). In the VAX-version of CANOPLOT, for example, the default width and height are set to 0.25 and 0.420, respectively, and the number of characters on a line is set 108 (to avoid problems with too many lines per page). The newpage number is 49 (a one) on the VAX.

Technical notes

Determination of the range in the uniform scaling.

CANOCO Scaling:	Extremes of axes are extremes of:
1	Species scores
2	Sample scores (SAMS+SAME)
3	Species scores + Sample scores

Installation note

The width and height of a character are set in line 81 and 83 of the main program (SIZE1 and SIZE2). The value of NCHAR on line 86 is 18 less than the number of characters on a line noted in the INI file. For MS-DOS, CANOPLOT.FOR has been compiled using Microsoft FORTRAN 5.0.

13 Glossary

See Table 1.1 in the CANOCO manual

Additions:

Eigenvector sample scores

The 'eigenvector sample scores' are, in an unconstrained analysis, the sample scores (which are derived from the species) and, in a constrained analysis, the sample scores which are linear combinations of environmental variables.

Acknowledgement

I wish to thank Dr. B. D. Ripley for settling the finer points in the permutation test for time series and spatial data.

15 References

Besag, J. and Clifford, P. (1989). Generalized Monte Carlo significance tests. *Biometrika.* **76**, 633-642.

Carnes, B. A. and Slade, N. A. (1982). Some comments on niche analysis in canonical space. *Ecology.* **63**, 888-893.

Chessel, D., Lebreton, J. -D. and Prodon, R. (1982). Mesures symétriques d'amplitude d'hbitat et de diversité intra-échantillon dans un tableau espèces-relevés: cas d'un gradient simple. *C.R. Acad. Sc. Paris, Series III.* **295**, 83-88.

Chessel, D., Lebreton, J. B. and Yoccoz, N. (1987). Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie. *Revue Statistique Appliquée.* **35**, 55-72.

Collins, M. F. (1987). A permutation test for planar regression. *Australian Journal of Statistics.* **29**, 303-308.

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

Dueser, R. D. and Shugart, H. H. (1978). Microhabitats in a forest-floor small-mammal fauna. *Ecology.* **59**, 89-98.

Dueser, R. D. and Shugart, H. H. (1979). Niche pattern in a forest-floor small-mammal fauna. *Ecology.* **60**, 108-118.

Dueser, R. D. and Shugart, H. H. (1982). Reply to comments by Van Horne and Ford and by Carnes and Slade. *Ecology.* **63**, 1174-1175.

Escoufier, Y. and Robert, P. (1979). Choosing variables and metrics by optimizing the RV-coefficient. In *Optimizing methods in Statistics,* J. S. Rustagi (eds), 205-219. New York: Academic Press.

Escoufier, Y. (1985). L'analyse des correspondances: ses propriétés et ses extensions. *Bulletin of the ISI, proceedings of the 45 th session (Amsterdam).*

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

Green, R. H. (1971). A multivariate statistical approach to the Hutchinsonian niche: bivalve mollucs of central Canada. *Ecology.* **52**, 543-556.

Green, R. H. (1979). *Sampling design and statistical methods for environmental biologists.* New York: Wiley.

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

Hall, P. and Titterington, D. M. (1989). The effect of simulation order on level accuracy and power of Monte Carlo tests. *J. R. Statist. Soc. B.* **51**, 459-467.

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

Hill, M. O. (1973b). Diversity and evenness: a unifying notation and its consequences. *Ecology.* **54**, 427-432.

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

Jongman, R. H. G., ter Braak, C. J. F. and van Tongeren, O. F. R. (1987). *Data analysis in community and landscape ecology.* Wageningen: Pudoc (in U.S.A: UNIPUB, 4611-F Assembly Drive, Lanham, Maryland 20706-4391)

Kempthorne, O. (1952). *The Design and Analysis of Experiments.* New York: Wiley.

Knox, R. G. (1989). Effects of detrending and rescaling on correspondence analyis: solution stability and accuracy. **83,** 129-136.

Lebreton, J. D., Chessel, D., Prodon, R. and Yoccoz, N. (1988). L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. *Acta Oecologia Generalis.* **9,** 53-67.

Legendre, P., Oden, N. L., Sokal, R. R., Vaudor, A. and Kim, J. (1990). Approximate analysis of variance of spatially autocorrelated regional data. *Journal of Classification* **7** .

Manger, R. and Schouten, A. J. (1989). Onderzoek naar de effecten van bekalking op de nematodenfauna van drie bosopstanden in Boswachterij St. Antonis (Peel-regio), rapport 718823001. Bilthoven: RIVM.

Minchin, P. R. (1987). Simulation of multidimensional community patterns: towards a comprehensive model. *Vegetatio.* **71,** 145-156.

Montgomery, D. C. and Peck, E. A. (1982). *Introduction to linear regression analysis.* New York: Wiley.

ter Braak, C. J. F. (1980). *Binary mosaics and point quadrat sampling in ecology.* Newcastle upon Tyne: MSc thesis.

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 reponse model. *Biometrics.* **41,** 859-873.

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 Barendregt, L. G. (1986). Weighted averaging of species indicator values: its efficiency in environmental calibration. *Math. Bio.* **78,** 57-72.

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

ter Braak, C. J. F. and van Dam, H. (1989). Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia.* **178,** 209-223.

ter Braak, C. J. F. (1990a). Interpreting canonical correlation analysis through biplots of structural correlations and weights. *Psychometrika.* **55,** .

ter Braak, C. J. F. (1990b). Permutation versus bootstrap significance tests in multiple regression and ANOVA. In *Bootstrapping and related resampling techniques (in prep),* K. -H. Jockel (eds), . Berlin: Springer Verlag.

van der Leeden, R. (1990). *Reduced rank regression with structured residuals.* Leiden: DSWO Press.

Van Horne, B. and Ford, R. G. (1982). Niche breadth calculation based on discriminant analysis. *Ecology.* **63,** 1172-1174.

Whittaker, J. (1984). Model interpretation from the additive elements of the likelihood function. *Appl. Statist.* **33,** 52-64.