6 Multivariate techniques

In Chapters 10-15 in Zuur, Ieno and Smith (2007), principal component analysis, redundancy analysis, correspondence analysis, canonical correspondence analysis, metric and non-metric multidimensional scaling, discriminant analysis and various other multivariate methods are discussed. In this chapter we show how to apply these methods and briefly discuss the output. We will not discuss when and why to apply certain methods in this manual as this is described and illustrated in detail in Zuur et al. (2007). Figure 6.1 shows the panels available from the 'Multivariate' main menu button. We start with principal component analysis.

6.1 Principal component analysis

One of the demonstration data sets available in Brodgar is a lobster data set. This data set consists of time series of catches per unit effort (CPUE) of lobsters in 11 areas in the Atlantic Ocean between 1960 and 1999. The data were available on an annual basis. To open the data, click on Import data – Demo data – Select the Lobster data – Load data – Continue – Save Changes and Finish Data Import Process. We will illustrate the Brodgar implementation of PCA using this data set. Click on the main menu button 'Multivariate', and select PCA in the left panel in Figure 6.1. Clicking on 'Go' in this panel gives the upper left window in Figure 6.2.

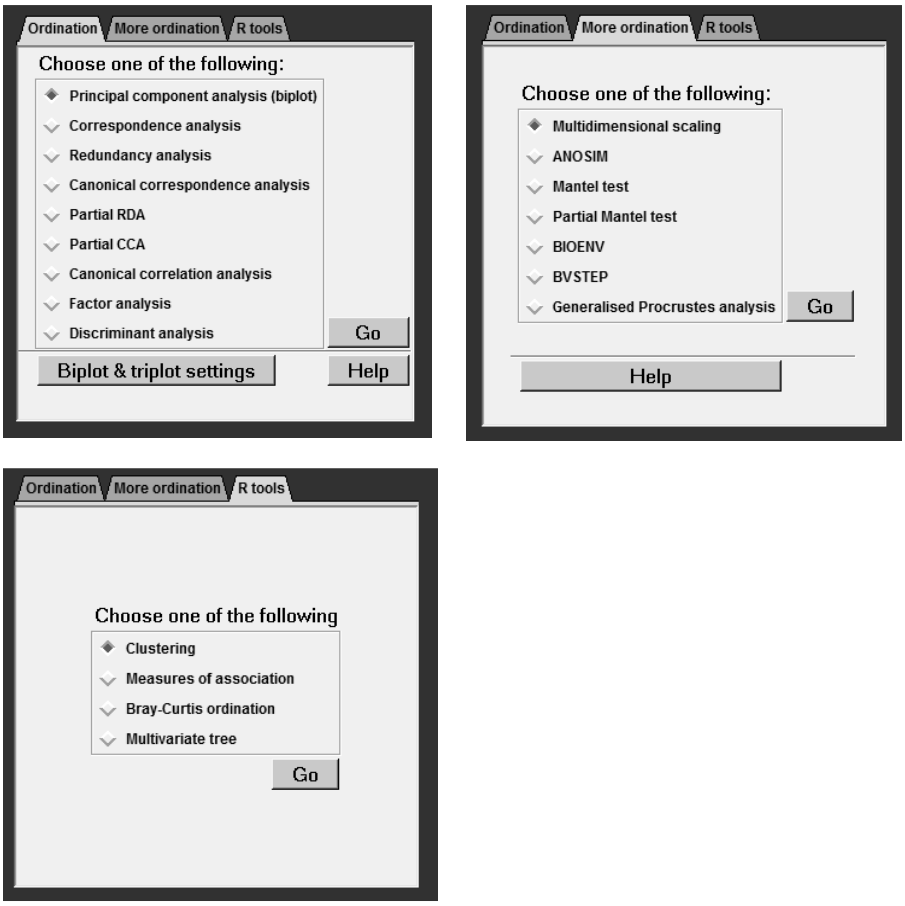

Figure 6.1. Multivariate techniques available in Brodgar.

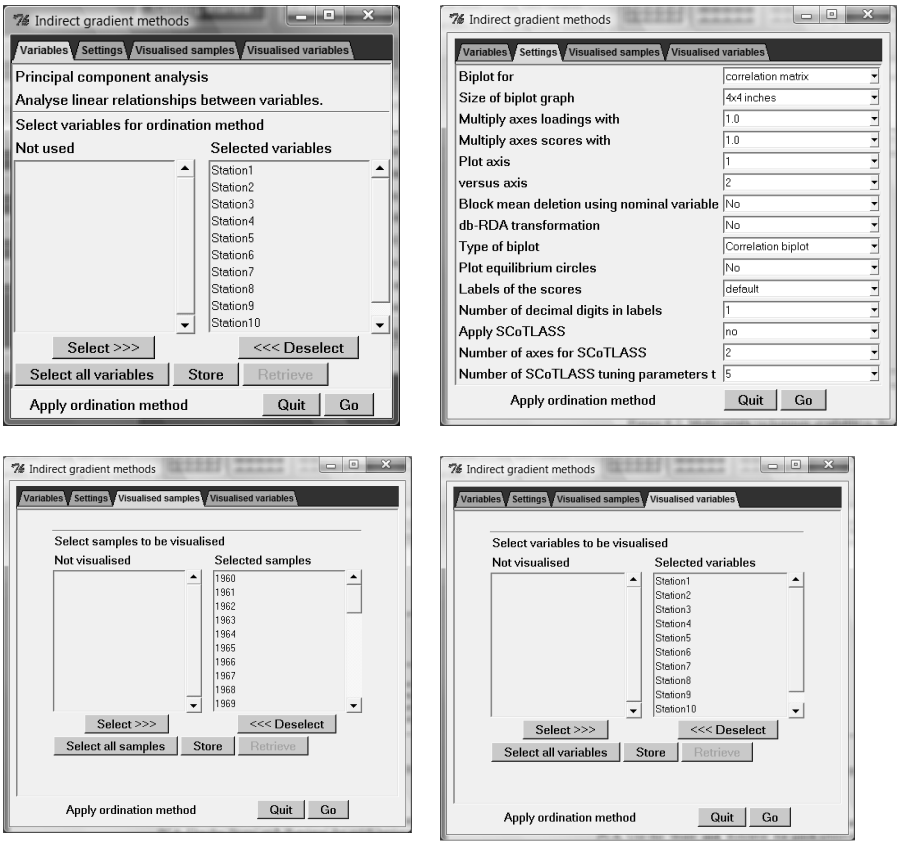

Figure 6.2. Multivariate techniques available in Brodgar.

The user has the following options.

Select variables for ordination. By default, all variables will be used for the PCA. Use the 'Store' and 'Retrieve' buttons for quick retrieval of selected variables.

Under settings:

Biplot for. Choose whether PCA should be applied on the correlation or covariance matrix.

Size of biplot graph. Increases the size (and therefore resolution) of the window. Handy for large screens.

Multiply axes loadings with. Useful if a variable name is only plotted half on the graph. The alternative is to decrease the font size of the graph, or shorten its name.

Multiply axes scores with. Useful if a label is only plotted half on the graph. The alternative is to decrease the font size of the graph, or shorten its name.

Plot axis i versus axis j. Choose which axes you want to plot. Make sure that you choose sensible values. If you make changes to these values, and then apply a PCA on a new data set, make sure that your choice still makes sense!

Block mean deletion using nominal variable. PCA allows the user to apply a mean deletion of the variables based on a nominal variable. For example, if 5 species are measured at 3 transects, and 10 observations are taken per transect (e.g. for the benthic data from Argentina, used elsewhere in this manual), we end up with a 30 by 5 data matrix. An extra column can be made in which the nominal variable 'Transect' identifies at which transect a particular observation was taken. Correlations between the 5 species are high if abundances are all above average at the same transect. Brodgar can remove the transect effect by subtracting the mean value per transect for each variable, prior to the PCA calculations. This can be done by using the 'Block mean deletion using nominal variable' option in the PCA menu. The selected variable needs to be a nominal (categorical) explanatory variable.

Db-RDA transformation. Legendre and Gallagher (2001) showed that various other measures of association can be combined with PCA, namely the Chord distance, Hellinger distance, and two Chi-square related transformations. We advise to use the Chord distance function. See the PCA and RDA chapters in Zuur et al. (2007) for further details.

Type of biplot. Allows the user to select the species or distance biplot. The choice has consequences for the interpretation of the biplot. The species conditional biplot is the ‘scale=0’ option in the biplot function in R, and the distance biplot is ‘scale=1’ in the biplot function in R.

Plot equilibrium circles. Lines beyond the inner circle are important, see also Legendre and Legendre (1998) or Zuur et al. (2007).

Labels of the scores. This option allows the user to select a variable (response or explanatory) and its values are plotted instead of the label. This is a useful tool to show how the scores are related to an explanatory variable. The next option reduces the number of digits.

Number of decimal digits in labels. See above.

Apply ScoTLASS. This is a technique to simplify the interpretation of the loadings and is discussed in Zuur et al. (2007), and references in there.

Number of axes for SCoTLASS. See above, or Zuur et al. (2007), and references in there.

Number of SCoTLASS tuning parameters t. See above.

In the two remaining panels in Figure 6.2, the user can choose which variables and observations not to visualise. Note that this does not mean that the de-selected ones are not used in the analysis, on the contrary. They are just omitted from the graphs.

Clicking on the 'Go' button in Figure 6.2 results in the biplot in Figure 6.3. There are various different scaling options possible in a biplot, see Chapter 12 in Zuur et al. (2007) for more details. The default scaling in Brodgar is the correlation biplot. In this scaling angles between lines (variables) represent correlations. This is the $\alpha = 0$ scaling in Jolliffe (1986). The exact mathematical form of the scores and loadings can be found on page 78 of Jolliffe (1986). Distances between points (years/observations) are so-called Mahanalobis distances. Years can be projected on any line, indicating whether the response variable in question had high or low values in those years. The biplot in Figure 6.3 indicates that:

- The series corresponding to the stations 1, 2, 3, 5 and 6 are highly correlated with each other.
- The series corresponding to the stations 8, 9, 10 and 11 are correlated with each other.
- The line for station 4 is rather short. This either means that there is not much variation at this site (the length of a line is proportional to the variance at a site), or that this site is not represented well by the first two axes.
- Based on the position of the years and the lines, it seems that values at the stations 1, 2 ,3, 5 and 6 were high during the 60s. Further insight can be obtained by looking at the time series plot also, and making use of the option to change colour of the lines (by clicking on the corresponding legend).

The button labelled 'Numerical info' in Figure 6.3, gives the following numerical output for PCA: (i) eigenvalues, (ii) eigenvalues as a percentage of the sum of all eigenvalues, and (iii) the cumulative sum of eigenvalues as a percentage. For the CPUE series, the first two axes represent 73% of the variation. If labels are rather long, it might happen that they lie outside the range of the figure (as is the case here). It is possible to increase the axes ranges by multiplying them with a small number (see above).

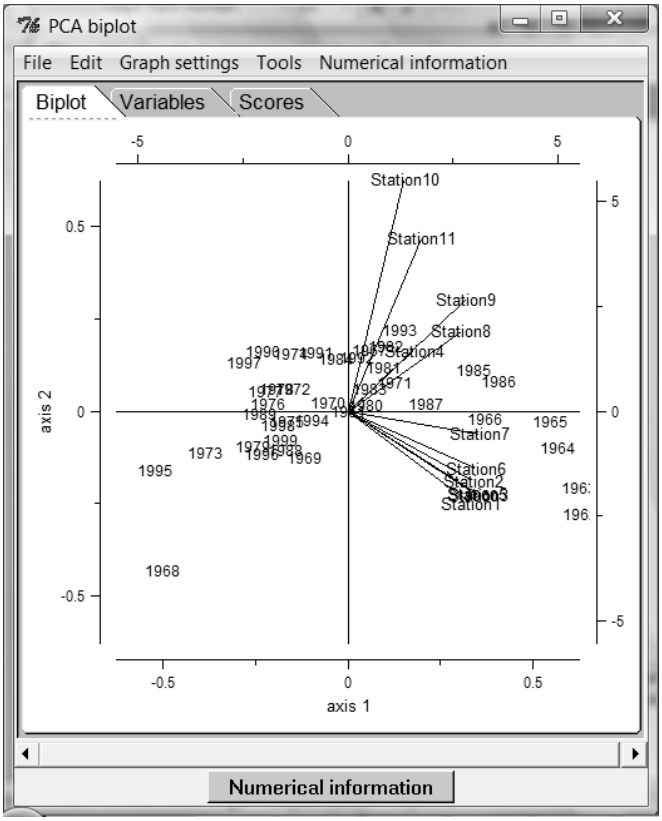

Figure 6.3. Biplot for CPUE series. The menu options in this window allow the user to chance colours, labels and titles, access the scores and loadings and make coenoclines. Clicking on Control-C in this graph, and Control-V in Word, will paste it as a high quality EMF graph.

6.2 Redundancy analysis

We imported the Argentine zoobenthic data (this is the "Zoobenthic data Argentina" in the demo data). To apply RDA, click on the main menu button 'Multivariate' and select RDA. Clicking on the 'Go' button gives the window in Figure 6.4. In this panel, the user can select response variables. By default, all variables are selected.

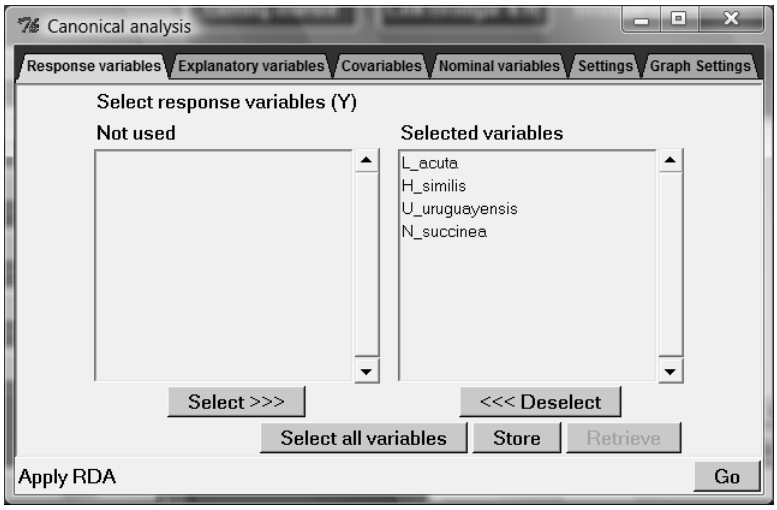

Figure 6.4. First panel for RDA in Brodgar: response variables.

Figure 6.5 shows the panel for the second tab; explanatory variables. Again, by default all explanatory variables are selected. The algorithms for RDA and partial RDA apply a standardisation (normalisation) to the explanatory variables. Therefore, one should not apply a standardisation during the Data Import process. From the data exploration in Chapter 4 in Zuur et al. (2007), we know that some of the explanatory variables are highly correlated with each other. Before the RDA analysis is started, Brodgar will calculate so-called VIF values. If these values indicate that the correlation between the explanatory variables is too high (collinearity), the analysis is terminated.

A nominal variable with more than 2 classes requires special attention. The variable Transect is nominal, and has 3 classes. Therefore, we created three new columns in Excel, called TransectA, TransectB and TransectC. If an observation is from transect A, the corresponding row in the variable TransectA is set to 1, and that of TransectB and TransectC to 0. The same was done for observations from transects B and C. Hence, the nominal variable transect with three classes is transformed in three new nominal variables that have only 2 classes (0 or 1). However, the three new variables TransectA, TransectB and TransectC cannot be used simultaneously in the RDA analysis, because they are collinear. One variable needs to be omitted and it does not matter which one. Instead of doing this in Excel, Brodgar can also do it for you; click Import data – Change data to be imported – highlight the column transect – Column – Generate dummy variables – Continue – Save Changes and Finish Data Import Process. The variables are called Transect_1, Transect_2 and Transect_3.

Our selection of explanatory variables is given in Figure 6.5. It is convenient to store the selection of explanatory variables.

Because nominal explanatory variables should be represented slightly different in a triplot (namely by a square instead of a line), Brodgar needs to know which explanatory variables are nominal (if any). This can be done in the fourth panel, see Figure 6.6. In this case, Season, Transect_1 and Transect_2 are nominal. Make sure that the selected nominal explanatory variables were also selected as explanatory variables in Figure 6.5 (although Brodgar will double check this).

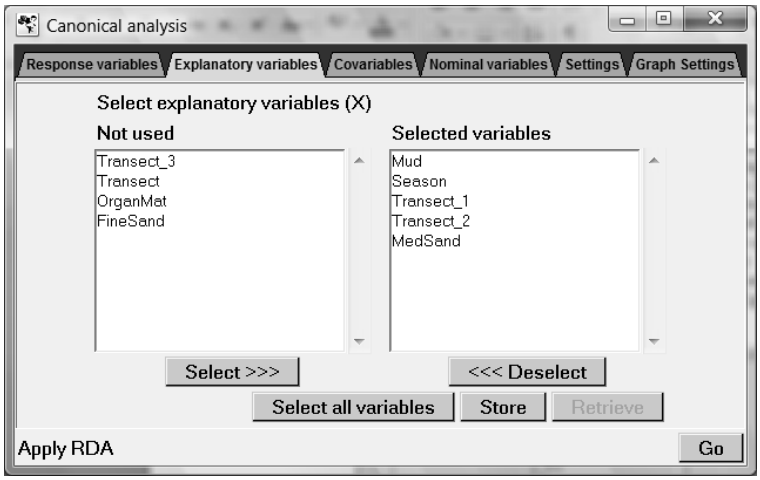

Figure 6.5. Second panel for RDA in Brodgar: explanatory variables.

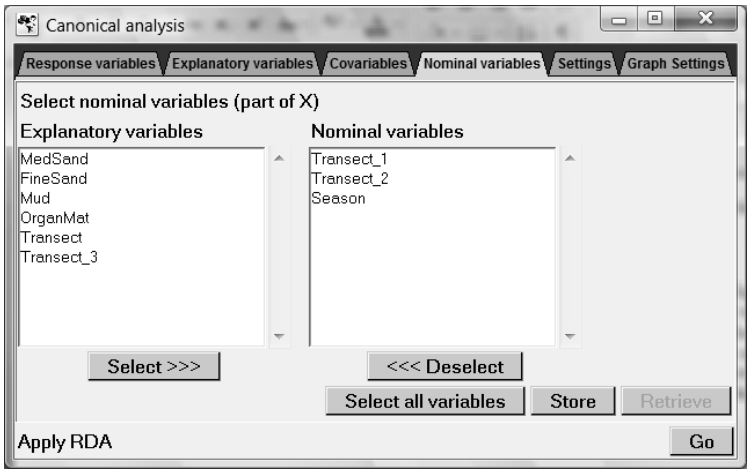

Figure 6.6. Fourth panel for RDA in Brodgar: nominal explanatory variables. Every nominal variable should be coded with 0 and 1.

The fifth panel is presented in Figure 6.7. It allows for various specific settings. Some of the general settings can be accessed from the 'Options' button in the upper left panel in Figure 6.4. The following options can be modified.

Scaling and centering. In RDA, the user can chose from two different scalings; the inter-Y correlation scaling or the inter-sample distances scaling. In ecology, response variables (Y) are species. If the prime aim of the analysis is to find relationships between species and explanatory (environmental) variables, we advise to use the inter-Y correlation (or: inter-species) scaling. This scaling is also called the species conditional scaling. If interest is on the observations, one should select the inter-sample distance scaling. This is also called the sample conditional scaling. Full details can be found in Ter Braak & Verdonschot (1995). The centering option is similar to applying the analysis on the covariance matrix.

Forward selection. To determine which explanatory variables are really important, an automatic forward selection procedure can be applied. This process is identical as in linear regression (Chapter 5), except that the eigenvalues are used instead of an AIC. The user can either select 'automatic selection', in which case all the explanatory variables are ranked from the most important to the least important, or 'best Q variables'. In the last case, Q has to be specified. It is also possible to apply a Monte Carlo significance test. This will give p-values for each explanatory variable.

Monte Carlo significance test. If selected, Brodgar will apply a permutation test. Currently, the only option is: 'all canonical axes'. The method gives the significance of all canonical axes. We advise to use a large number of permutations (e.g. 999 or 1999).

db-RDA transformation. Legendre and Gallagher (2001) showed that various other measures of association can be combined with PCA, namely the Chord distance, Hellinger distance, and two Chi-square related transformations. We advise to use the Chord distance function. See Zuur et al. (2007) for more details.

Interpretation of the triplots, biplots and numerical output is given in Zuur et al. (2007).

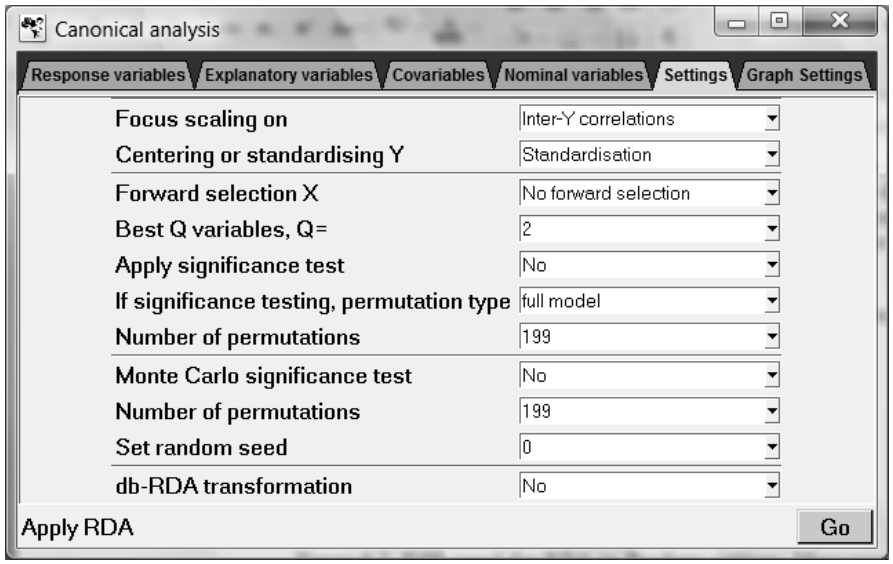

Figure 6.7. Fifth panel for RDA in Brodgar: settings. More recent versions of Brodgar have a sixth panel allowing for changes in basic graphical settings.

6.3 Correspondence analysis and canonical correspondence analysis

Applying correspondence analysis or canonical correspondence analysis in Brodgar is similar to PCA and RDA (see above), and is therefore not discussed here. Due to the nature of the technique, there are considerably less options for CA than for PCA.

6.4 Discriminant analysis

The Fisher Iris data are used in many statistical textbooks to illustrate techniques like discriminant analysis and principal component analysis. The four variables sepal length, sepal width, petal length and petal width were measured on 50 plant specimens of three types of iris, namely *Iris setosa*, *Iris versicolor* and *Iris virginica*. Hence, the data contains 150 observations on 4 variables. We added an extra column with values 1, 2 and 3 that identifies the species. The data can be loaded from the demo data in Brodgar.

Discriminant analysis can be applied because the observations can be divided in three groups (each of size 50 in this case). To tell Brodgar which observations (rows) belong to the same group, the column with the values 1, 2 and 3 is needed. It is called Species, but any name can be used. It is important (for Brodgar) to start labelling the groups at 1. But the rows do not have to be sorted according to the groups. Select the four variables as response variable and Species as an explanatory variable. For optimal presentation of the graphical output, it can be useful to standardise the response variables.

By choosing 'discriminant analysis' in the multivariate analysis menu and clicking the 'Go' button, the discriminant analysis window in Figure 6.8 appears.

Here, the user can (i) select and de-select variables (by default all variables are used), (ii) de-select samples, (iii) identify the groups by selecting a (nominal) explanatory variable, and start the DA calculations. Step 2, de-selecting samples, is optional. Please note that if a row (observation) contains a missing value, the entire row is de-selected. Do no select rows with missing values as this will result in an error message. In step 3, groups can be identified by selecting an explanatory variable (as explained above). Clicking the 'Go' button will carry out the DA calculations.

The numerical output is saved in the file bipl1.out in the project directory. The file bipl2.out is used by Brodgar to generate the graphs. Various examples of DA are presented and discussed in detail, including the numerical output, in Zuur et al. (2007).

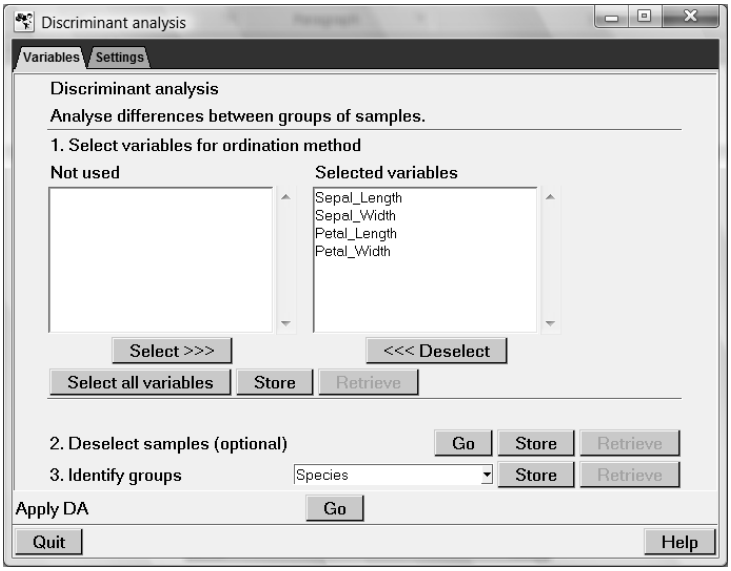

Figure 6.8. Steps in discriminant analysis.

6.5 Measures of association

Before you can run this tool, you first need to install the vegan package in R. To do this, start R and click on: Packages – Install Package(s) – select a CRAN mirror and click OK – select vegan and click ok – close R.

To calculate measures of association, click on the main menu button 'Multivariate' in Brodgar. It will show the left panel in Figure 6.1. Other panels available from the multivariate menu are given in the same figure. By selecting 'Measures of association' under 'R tools' and clicking on the 'Go' button in the right panel in Figure 6.1, the window in Figure 6.9 appears.

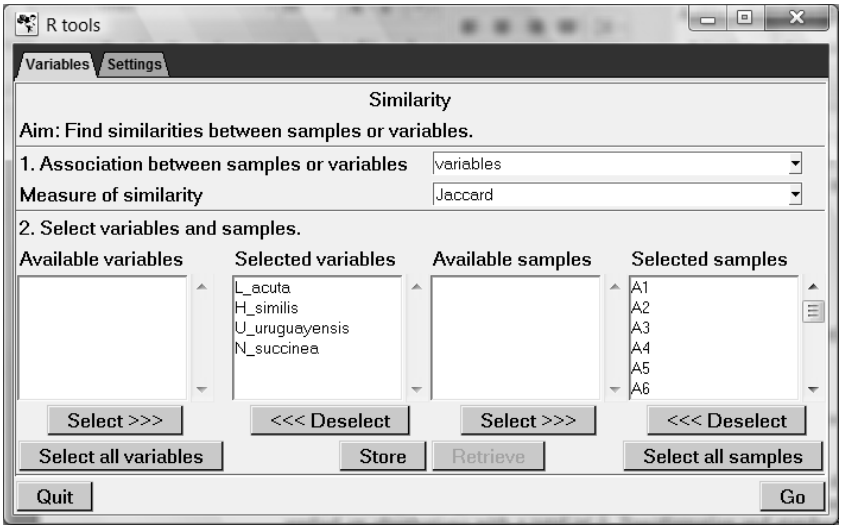

Figure 6.9. Selection of variables and settings for measures of association.

The user has the following options:

Association between samples or variables. Calculate the measure of association between the variables or sites.

Measures of similarity. Make sure that the selection is valid for the data. For example, Brodgar will give an error message if the Chi-square distance function is applied on observations with a total of 0. Transformation and standardisation, which might have been selected during the data import process, can be another problem. For example, a log transformation might result in negative numbers, in which case it does not make sense to select the Jaccard index function. There are only a few measures of association that can cope with normalised data (e.g. the

correlation function). Summarising, if an error message appears, the user is advised to check whether the selected measure of association is valid.

Select variables and samples. Click on 'Select all variables' and 'Select all samples' to use all the data. The 'Store' and 'Retrieve' functions can be used to speed up the clicking process.

Under settings, items like the title, labels, graph type and graph name can be specified. Some of the available measures of association are similarity matrices, whereas other ones are dissimilarity matrices. The MDS function in R requires a dissimilarity matrix as input. Therefore, some indices are transformed into a dissimilarity matrix using the formula (in matrix notation): D2 = max(D1) - D1, where D1 is the similarity matrix and D2 the dissimilarity matrix. This is done for the Jaccard index, Community coefficient, Similarity ratio, Percentage similarity, Ochiai coefficient, correlation and covariance functions, Sorensen and Relative Sorensen functions.

Once, the appropriate selections have been made, click on the 'Go' button in Figure 6.9. This results in Figure 6.10. It shows the MDS ordination graph, which was obtained by applying the R function `cmdscale` to the selected dissimilarity matrix. The actual values of the measure of association can be obtained from the menu in Figure 6.10; click on 'Numerical output'. The measures of association can either be opened in a text file, or copied to the clipboard as tab separated data, and from there it can be pasted into Excel. The user can also access these values in the project directory. Look for the file \YourProjectName\meassimout.txt. One can also open the file similarity.R (which is in the Brodgar installation directory) and source all the code directly in R.

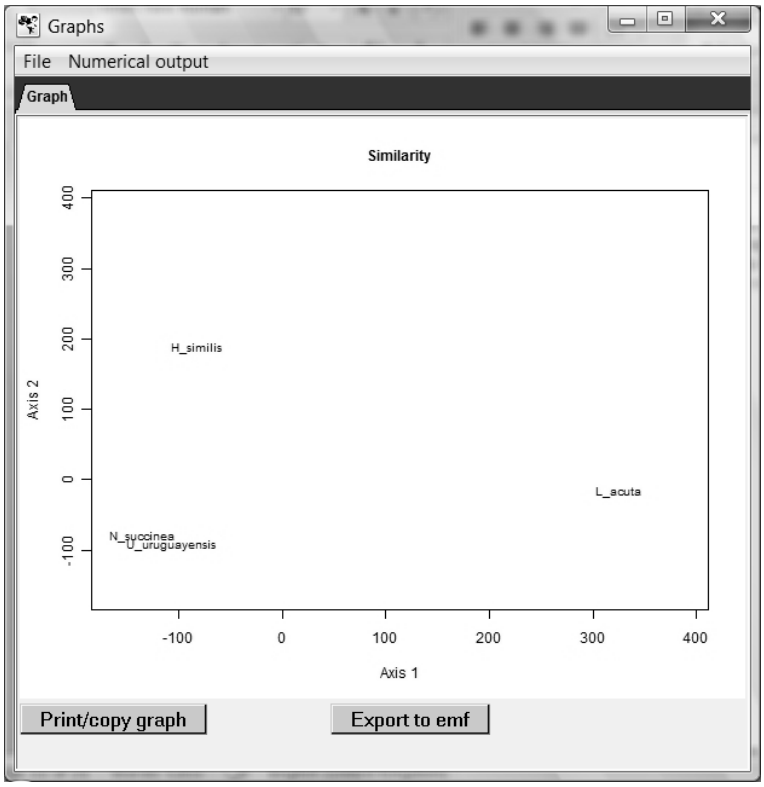

Figure 6.10. Results of MDS for Euclidean distance measure.

6.6 Bray-Curtis ordination in Brodgar

To carry out Bray-Curtis ordination in Brodgar, select it in Figure 6.1 (Multivariate – R Tools – Bray Curtis ordination), and click on the 'Go' button. The options for Bray-Curtis ordination are given in the two panels in Figure 6.11. The user can make the following selections under the 'Variables' panel.

Visualise association between samples or variables. Apply Bray-Curtis ordination on the variables or observations.

Measure of similarity. These were discussed in Chapter 10 in Zuur et al. (2007). Make sure that the selection is valid for the data. For example, Brodgar will give an error message if the relative Sørenson index is applied on a data set in which there are observations with zero values everywhere.

Select variables and samples. Click on 'Select all variables' and 'Select all samples' to use all the data.

Methods to select endpoints. This is discussed in Zuur et al. (2007).

Setting axes limits. Brodgar will automatically scale the axes. If one has long species names, part of the name might fall outside the graph. In this case, change the minimum and maximum values of the axes. The input needs to be of the form: 0,100,0,100 for two axes, and 0,100 for 1 axis. Note the comma separation!

Number of axes. Either 2 axes (default) or 1 axis is calculated. Details of the algorithm implemented in Brodgar can be found in McCrune and Grace (2002). To obtain a second axis, residuals distances are obtained before calculation on the second axis starts. The size of the graph labels and angle of the labels (for 1 axis only) can be changed. Other options include the title, labels and graph name.

Numerical output is available from the menu in the Bray-Curtis window (not shown here). Other information available is the distance matrix and the scores. If 2 axes are calculated, Brodgar will also present the correlations between the original variables and the axes in a biplot. Results of Bray-Curtis ordination for the Argentine zoobenthic data are presented in Zuur et al. (2007).

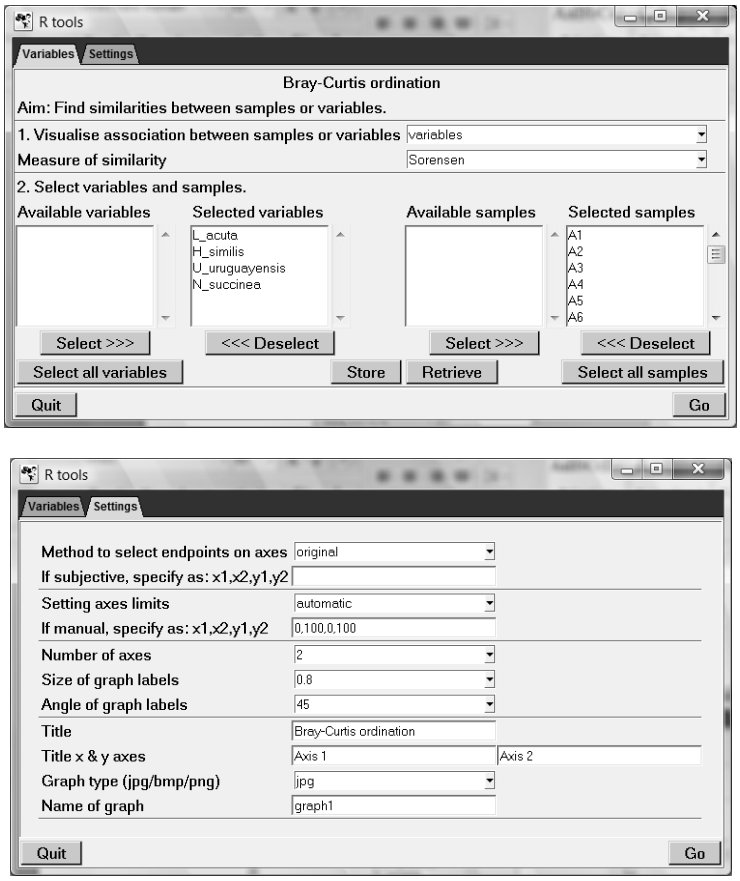

Figure 6.11. Bray Curtis options in Brodgar.

6.7 Generalised Procrustes analysis

GPA can be used for the analysis of 3-way data. Examples of such data are:

- *N* species measured at *M* sites in *T* years.
- *N* species sampled at *M* different areas in *T* years.
- *N* fish species sampled in *M* hauls by *T* different boats.
- *N* panel members assessing the quality of *M* products, during *T* assessments.

One option is to stack all data in one matrix and apply a dimension reduction technique like PCA or MDS. Different labels can be used to identify the original groups. Alternatively, nominal variables can be used to identify one of the three

factors and using redundancy analysis or canonical correspondence analysis. For some data sets this approach might work well. However, one might also end up with non-interpretable results, especially if T is larger than 3 or 4. A different approach is Generalised Procrustes Analysis (Krzanowski 1988). In generalised Procrustes analysis (GPA), 2-way data are analysed for each value of the third factor.

In the data contains N species measured at M sites in T years, then there are T tables containing species-by-sites data. A dimension reduction technique (e.g. MDS) is applied to on each of these N-by-M tables. If interest is on relations between species, the GPA can be set up such that the resulting ordination diagrams contain points representing species. Species close to each other are similar (or correlated if PCA is used), species not close to each other are dissimilar. This interpretation is based on distances between points. These distances are relative. The ordination diagram can be turned upside down, rotated, enlarged or reduced in size without changing the interpretation. Making use of this characteristic, GPA calculates the 'best' possible rotation, translation and scaling for each of the T ordination diagrams, such that an average ordination diagram fits each of the T diagrams as good as possible. Formulated differently, GPA calculates an average ordination diagram, based on the T ordination diagrams. An analysis of variance indicates how well (i) individual species, (ii) each of the ordinations and (iii) each of the axes are fitted by this average ordination diagram.

6.7.1 Fisher's Irish data

In Section 6.4 we used Fisher's Iris data. The four variables sepal length, sepal width, petal length and petal width were measured on 50 plant specimens of three types of iris, namely *Iris setosa*, *Iris versicolor* and *Iris virginica*. Hence, the data contains 150 observations on 4 variables. The spreadsheet has been organised in such a way that the first 50 rows are from species 1, the second block of 50 observations from species 2, and the last 50 rows are from species 3. Summarising, the structure of the data is

$$E = \begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix}$$

Each D_i is of dimension 50 – by – 4. The general data format for GPA is of the form:

$$E = \begin{bmatrix} D_1 \\ .. \\ D_T \end{bmatrix}$$

Missing values are allowed in this technique. We show how GPA can be used to analyse whether relationships between the 4 variables are different between the three species.

The algorithm for GPA will calculate the similarity between the columns of each D_i. So, if D_1 contains the data from the first 50 specimen of species 1, then we end up with a 4 – by – 4 (dis-) similarity matrix for species 1; the same holds for D_2 and D_3.

If a variable was not measured in year s, fill in NA in the entire column in Ds. The matrix E can be imported in Brodgar in the usual way, but you need to sort the data in Excel by species (group).

Click on the 'Multivariate' button from the main menu, and then on the tab labelled 'Misc', and proceed with GPA. The window in Figure 6.12 appears (assuming you have loaded the demo lobster data).

Under the 'Settings' panel, Euclidean distances are chosen as measure of dissimilarity. Click the 'Go' button to deselect observations. The panel in Figure 6.13 appears. Select the rows 50, 100 and 150 as end points of the groups and click 'Finish'.

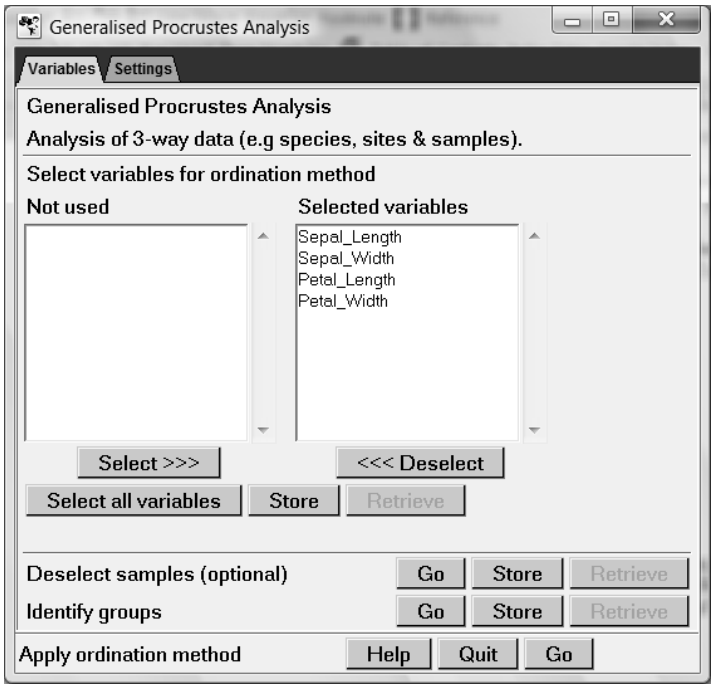

Figure 6.12. GPA window for Fisher's Iris data.

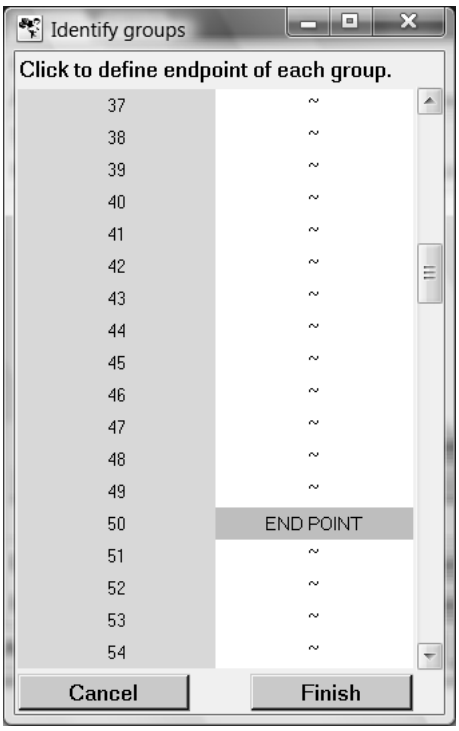

Figure 6.13. Select row 50, 100 and 150 as endpoints.

Brodgar will indicate whether GPA can be applied, and enable the GPA button in Figure 6.12. Clicking it starts the calculations and results are presented in Figure 6.14. The upper left panel shows the average GPA ordination diagram. Results indicate that on average, sepal and petal length are similar to each other, and the same holds for sepal and petal width, but length and width values are different.

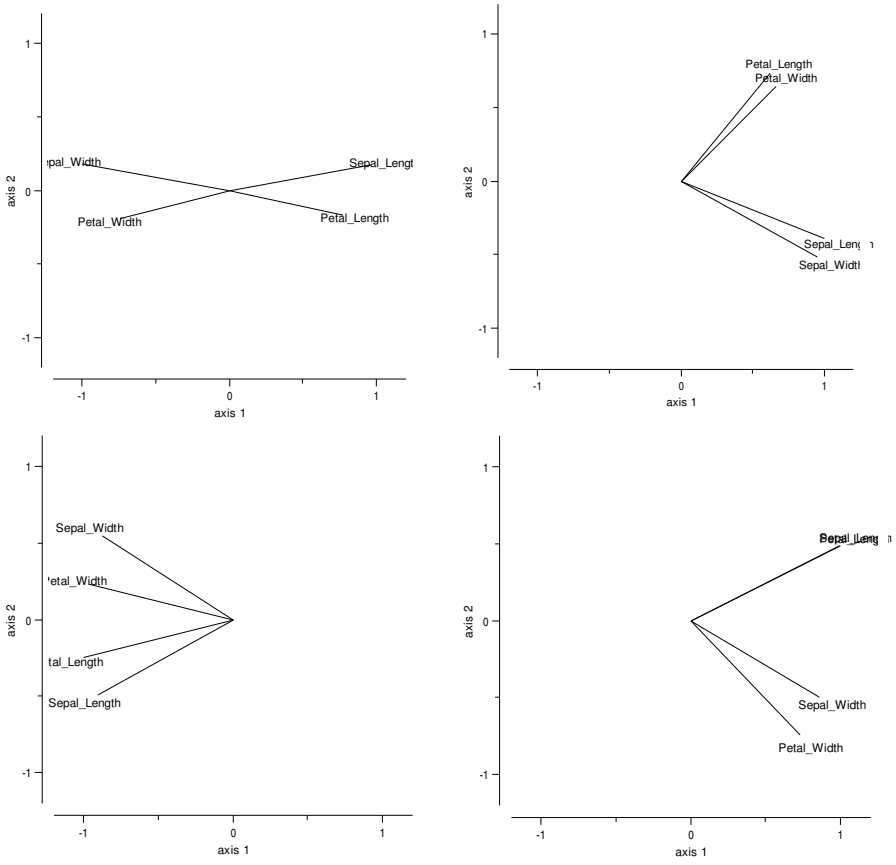

Figure 6.14. Upper left: Results of GPA. This is the average of the other three ordination diagrams. Upper right: MDS ordination diagram for the data of species 1. Upper left: Ordination diagram for species 2. Lower right: Ordination diagram for species 3.

6.8 Clustering

Clustering should only be applied if the researcher has a priori information that the gradient is discontinuous. There are various choices that need to be made, namely:

1. The measure of similarity (Section 6.1).

2. Sequential versus simultaneous algorithms. Simultaneous algorithms make one step at a time.
3. Agglomeration versus division.
4. Monothetic versus polythetic methods. Monothetic: use one descriptor. Polythetic: multiple descriptors.
5. Hierarchical versus non-hierarchical methods. Hierarchical methods use a similarity matrix as starting point. These methods end in a dendrogram. In non-hierarchical methods objects are allowed to move in and out of groups at different stages.
6. Probabilistic versus non-probabilistic methods.
7. Agglomeration method.

Brodgar makes use of the R function `hcust`, which allows for hierarchical cluster analysis using a dissimilarity matrix. The user can choose:

1. Whether clustering should be applied to the response variables are the samples.
2. The measure of similarity. The options are the community coefficient, similarity ratio, percentage similarity, Ochiai coefficient, chord distance, correlation coefficient, covariance function, Euclidean distance, maximum (or absolute) distance, Manhattan distance, Canberrra distance, binary distance, and the squared Euclidean distance. See also Chapter 6 in Zuur et al. (2007).
3. The agglomeration method. This decides how the clustering algorithm should link different groups with each other. Suppose that during the calculations, the algorithm has found four groups, denoted by A, B, C and D. At the next stage, the algorithm needs to calculate distances between all these groups and decide which two groups to fuse. One option (default in Brodgar) is the average linkage; all the distance between each point in A to each point in B are averaged, and the same is done for the other combinations of groups. The two groups with the smallest average distance are then fused. Alternative options are the single, complete, median, Ward, centroid and McQuitty linkage. Whichever linkage function is made, it might change the results drastically, or it might not. Some linkage functions are sensitive to outliers, absolute abundance, distributions of the species all influence, see also Jongman et al. (1996).
4. Which variables and observations to use. It is convenient to store these settings for later retrieval.
5. Titles and labels.

6.9 Multivariate tree models

Multivariate tree models are discussed in De'Ath (2002), and the reader is advised to read this paper before applying multivariate tree models in Brodgar. Nearly all tools discussed in De'Ath (2002) are available in Brodgar. In order to

apply this method in Brodgar, the user needs R version 1.8 or higher. To install the package `mvpart` in R: Start R and click on: Install package(s) – Select a CRAN mirror – OK – select mvpart – OK, and close R. You only need to do this once.

Once this has been done, applying multivariate tree models in Brodgar is straightforward. Selecting 'Multivariate tree' and clicking the 'Go' button in the right panel in Figure 6.1 pops up the multivariate tree menu. It allows the user to:

1. Select and de-select multiple response variables.
2. Select and de-select continuous and nominal explanatory variables.
3. Change various settings.
4. Select and de-select observations.

Selecting response variables, explanatory variables and observations is trivial and is not discussed further. Figure 6.15 shows the possible settings for the multivariate tree models.

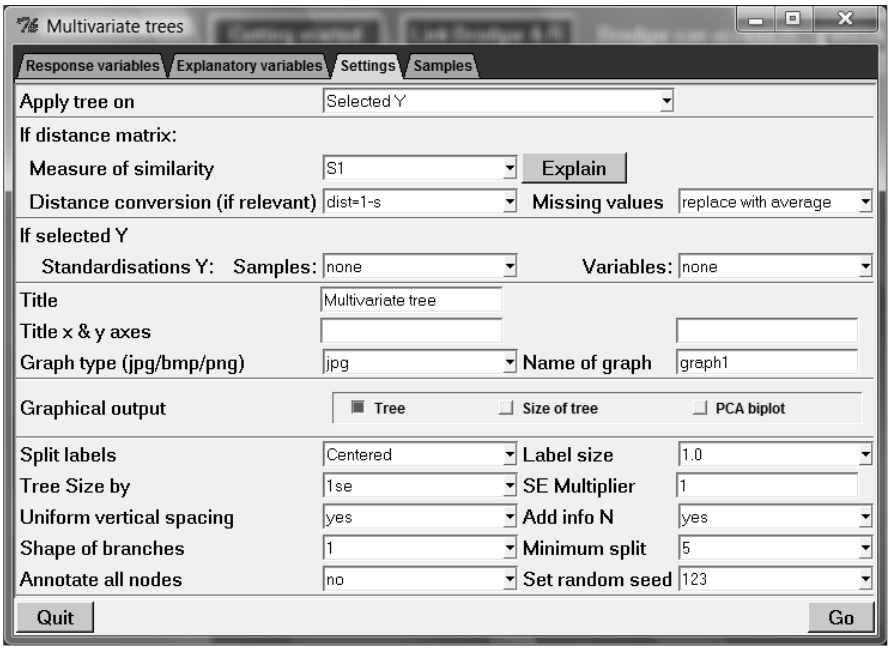

Figure 6.15. Settings for multivariate trees.

First, the user has to decide whether the tree model should be applied on the response variables or on a distance matrix, see De'Ath (2002) for details. If it is applied on a distance matrix, a measure of similarity has to be chosen and a conversion method of similarities to distances (if relevant). The similarity measures are

discussed below (see MDS in Subsection 6.10.3). If during the data import process no standardisation was applied, the user can select row (observations) and column (variables) standardisations. The default values are none.

Titles, labels for the axes and graph name and type can be specified. The graphical output consist of the tree, a cp-plot which can be used to determine the size of the tree and a PCA biplot applied on the mean values per group, see also De'Ath (2002). The tree size can be determined by the 1 standard error rule (default) or it can be determined interactively. All other options are self-explanatory and further information can be obtained from the following help files (see also the `tree` function in Chapter 5):

- The help file for `mvpart`. Start R, type: `library(mvpart)` and on the next line type: `?mvpart`
- Telp file for rpart. Start R, type `library(rpart)` and on the next line type: `?rpart`

Note that if a distance matrix is used, the R code does not apply cross-validation.

The `mvpart` package is distributed under the GPL2 license and Brodgar complies with the R GPL license. We like to thank the authors Terry M Therneau, Beth Atkinson, Brian Ripley, Jari Oksanen and Glenn De'Ath of the `rpart`, `vegan` and `mvpart` packages for their effort.

6.10 Other techniques

6.10.1 Canonical correlation analysis

Canonical correlation analysis (CCOR) can only be used if there are more observations than response variables. Technically, CCOR calculates a linear combination of the response variables:

$$Z_{i1} = c_{11}Y_{i1} + c_{12}\,Y_{i2} + \ldots + c_{1N}\,Y_{iN}$$

and a linear combination of the explanatory variables:

$$W_{i1} = \mathrm{d}_{11}X_{i1} + d_{12}\,X_{i2} + \ldots + d_{1Q}\,X_{iQ}$$

such that the correlation between Z and W is maximal. Further axes can be calculated. Within ecology, CCOR is less popular than CCA.

6.10.2 Factor analysis

Factor analysis (FA) is not popular in ecology, and therefore it is only briefly discussed here. In FA, the user can change the estimation method (maximum likelihood or principal factor), the factor rotation (none, varimax, quartimax, equamax and oblimin), and whether a row normalisation (Kaiser) should be used. The varimax, quartimax and equamax are orthogonal rotations, whereas the oblimin is an oblique rotation. Most software packages use the maximum likelihood estimation method and a varimax rotation. Results of the dimension reduction techniques are stored in the files biplot.out and bipl2.out in your project directory. The correlation matrix used in FA can be found in facorrel.txt and the residuals obtained by FA are stored in faresid.txt.

6.10.3 Multidimensional scaling

Brodgar can also perform metric and non-metric multidimensional scaling, see the middle panel in Figure 6.1. The difference between this option and the one via the R-tools menu is that the R function is metric MDS, whereas the one in the middle panel in Figure 6.1 also allows for non-metric MDS. Legendre and Legendre (1998) described a large number of measures of similarity and Brodgar contains most of them. The notation used in Brodgar is identical to Chapter 7 in Legendre and Legendre (1998), and we strongly advise to read this chapter. The following measures of associations are available in Brodgar.

Symmetrical binary coefficients. These transfer the data to presence-absence data.

- S1: The simple matching coefficient.
- S2: Coefficient of Rogers & Tanimoto (variation of S1).
- S3, S4, S5 and S6: Four other measures of similarity (Equations 7.3 - 7.6 in Legendre and Legendre 1998).

Asymmetrical binary coefficients. These transfer the data to presence-absence data, but the measures of association exclude the double zeros.

- S7: Jaccard's coefficient.
- S8: Sørensen's coefficient (gives double weight to joint presence).
- S9: Variant of S7 that gives triple weight to joint presence.
- S10: Counterpart of S2.
- S11: See Equation (7.14) in Legendre and Legendre (1998).
- S12: See Equation (7.15) in Legendre and Legendre (1998).
- S13: Binary version of Kulczynski's coefficient S18 for quantitative data.
- S14: Ochiai index.

Symmetrical quantitative coefficients. These do not transfer the data to presence-absence data.

- S15: Gower's coefficient: General index making use of partial indices.
- S16: Similar as S15.

Asymmetrical quantitative coefficients. These do not transfer the data to presence-absence data.

- S17: See Equation (7.24) in Legendre and Legendre (1998).
- S18: Kulczynski's coefficient.
- S19: Kulczynski's coefficient.
- S20: See Equation (7.27) in Legendre and Legendre (1998).
- S21: Chi-square similarity

Probabilistic coefficients.

- S22: Probabilistic Chi-square with P degrees of freedom
- S23: Goodall's similarity index; see Equation (7.31) in Legendre and Legendre (1998).
- S27: See Equation (7.33) in Legendre and Legendre (1998).

Distance coefficient: metric distances

- D1: Euclidean distance.
- D2: Average distance.
- D3: Chord distance.
- D4: Geodesic metric.
- D5: Mahalanobis generalised distance.
- D6: Minkowski's metric.
- D7: Manhattan metric (taxicab/city-block metric).
- D8: Mean character difference.
- D9: Whittaker's index of association.
- D10: Canberra metric.
- D11: Coefficient of divergence.
- D12: Coefficient of racial likeness.
- D15: Chi-square metric
- D16 Chi-square distance
- D17: Hellinger distance

Distance coefficient: semimetric distances

- D13: Nonmetric coefficient.
- D14: Percentage difference (D14 = 1 - S17).

Note that some of these coefficients cannot be applied on data containing negative values. The options under the MDS settings window are self-explanatory and are not discussed here.

6.11 Missing values

In PCA, CA, RDA and CCA, missing values in the response variables are replaced by the mean values of the corresponding response variables. In RDA and CCA, missing values in the explanatory variables are replaced by the mean values of the corresponding explanatory variables. In MDS, the user can choose whether (i) the similarity coefficient between two response variables should only be based on observations measured for both, or (ii) missing values should be replaced by averages. In FA, a slightly different approach is followed. Means and variances are computed from all valid data on the individual variables. Covariances and correlations are computed only from the valid pairs of data. Finally, in discriminant analysis, only those observations are used which do not contain missing values.

6.12 (Partial) Mantel test, ANOSIM, BIOENV & BVSTEP

The statistical background of all these techniques are discussed in Chapter 10 in Zuur et al. (2007). The Brodgar clicking process in each of these methods is trivial and follows that of MDS and PCA.

6.12.1 ANOSIM

The ANOSIM (Analysis of similarities) permutation method allows for testing for group structure in the observations. If the original data contains abundance of M species measured at N sites, an N – by – N matrix of (dis-)similarities D is calculated. Suppose that the observations are from four different transects. As a result, the matrix D contains a block structure:

$$\mathbf{D} = \begin{pmatrix} D_{11} & D_{12} & D_{13} & D_{14} \\ D_{21} & D_{22} & D_{23} & D_{24} \\ D_{31} & D_{32} & D_{33} & D_{34} \\ D_{41} & D_{42} & D_{43} & D_{44} \end{pmatrix}$$

The sub-matrices D_{ii} represent the (dis-)similarities between observations of the same transect, and D_{ij} between observations of different transects. A statistic based on both the between and within sub-matrices is used to test the differences among the 4 groups. Further details can be found in Legendre and Legendre (1998), or in Chapter 10 of Zuur et al. (2007). A p-value for the statistic is obtained by permu-

tation. To tell Brodgar which samples belong to the same group, a nominal variable needs to be selected as the blocking variable.

ANOSIM can only be applied if "Association between samples" is selected. ANOSIM can be applied on 1-way data, 2-way nested data, 2-way crossed data with replication and 2-way crossed data with no replication.

6.12.2 Mantel test and Partial Mantel test

In the Mantel test, two (dis-)similarity matrices are calculated; one for the Y variables (e.g. *M* species measured at *N* sites) and one for the X variables (e.g. *Q* environmental variables measured at the same *N* sites). To assess the relationship between the two sets of variables, the two *N* – by - *N* (dis-)similarity matrices are compared. This is done by calculating the correlation (either Pearson, Spearman or Kendall) between the elements of these two matrices. A permutation test is then used to obtain a *p*-value for the correlation coefficient. If the observations are from different transects, it might be sensible to permute the observations only within the same transect. This can be done with the strata variable. If a strata variable is selected, the permutations will be applied within the levels of the strata variable. The strata variable must be a nominal variable.

In the partial Mantel test, a third set of variables (taken from X) is used. This gives three matrices of (dis-)similarities. The correlation (either Pearson, Spearman or Kendall) between the elements of the first two matrices is calculated, while partialling out the effect of the third matrix. Again, a permutation test is used to obtain a *p*-value for the correlation coefficient. It might be interesting to use spatial or temporal effects in the third matrix.

If the first data matrix contains species abundance and the second matrix explanatory variables, it might be sensible to use two different measures of association, e.g. Bray-Curtis for the species data and Euclidean distances for the explanatory variables. The strata variable should not be selected in any of the matrices. Missing values can either be replaced by variable averages or omitted in the pairwise calculation of the measure of association. For the final analysis, we advise to use 9999 permutations instead of the default value of 999. The conversion from similarity to dissimilarity is applied automatically, if required. The default conversion is: distance = 1 - similarity.

The sample labels should be unique. Besides testing for relationships between species and explanatory variables, the (partial) Mantel test can be applied on various other types of data.

6.12.3 BIOENV & BVSTEP

BIOENV is yet another way to link BIOtic and ENVironmental variables. Just as in the Mantel test, the user needs to select variables for two data matrices Z_1 and Z_2. In a typical ecological application Z_1 contains the species data and Z_2 the explanatory (environmental) variables. Hence Z_1 and Z_2 are of dimension *N* – by –

M and *N* – by – *Q* respectively, where *N*, *M* and *Q* are the number of samples, species and explanatory variables respectively. The method calculates two distance matrices D_1 and D_2, both of dimension *N* – by - *N*. To link D_1 and D_2, the Mantel test described above calculates the correlation between the elements of D_1 and D_2.

Instead of using all variables in Z_2 for the calculation of D_2, BIOENV uses a subset of variables. To be more precisely, it will first take each individual variable in Z_2, calculate D_2, and then the correlation between the elements of D_1 and D_2. This process is then repeated using two variables in Z_2, three variables, four variables, etc. It will try every possible combination of variables. For each combination of explanatory variables, only the 10 highest correlations between D_1 and D_2 are shown. If there are more than five explanatory variables in Z_2, BIOENV might become rather slow (depending on the sample size) as every possible combination is used. Therefore, if there are more than twelve variables in Z_2, a random selection of the large number of possible combinations is presented. In this case, it might be better to use the function BVSTEP.

The difference between BIOENV and BVSTEP is that instead of trying every possible combination of variables in Z_2 at each step, BVSTEP applies a forward selection.

The correlation coefficient can be Pearson, Spearman, Kendall or weighted Spearman. The option "Number of variables in matrix 2" allows the user to specify an upper limit on the number of variables in Z2.

BIOENV and BVSTEP are mainly used in ecological applications in which Z_1 contains the species data and Z_2 the environmental data. However, nothing stops the user to apply these methods on other types of data.