

Canonical Analysis

Principles of Canonical Analysis
Redundancy Analysis (RDA)
 Worked Example of RDA
Canonical Correspondence Analysis (CCA)
Canonical Correlation Analysis (CCorA)
Discriminant Analysis (DA)
 Worked Example of DA

Canonical Analysis

Canonical analysis is the simultaneous analysis of two, or eventually several data tables. It permits biologists to do a *direct comparison* of two data matrices. Hence, canonical analysis and its derivatives are known as *direct ordination methods*.

Often, in ecology, one is interested in the relationship between a first table describing species composition and a second table of environmental descriptors, observed *at the same locations* (i.e., objects or samples).

Canonical Analysis

Previous to this, we have considered *indirect ordination* methods (PCA, PCO, NMDS, CA, DCA) in that we would ordinate a species \times stand matrix and then conduct some form of correlation or regression analysis on the ordination vectors to relate objects or descriptors to externally obtained environmental information. This procedure is performed *a posteriori*.

In canonical analysis, with two matrices (\mathbf{X} and \mathbf{Y}), one is constrained by the other, and both are examined simultaneously. This permits one to directly test *a priori* hypotheses by bringing all of the variance of \mathbf{Y} that is directly related to \mathbf{X} and allowing formal tests of the hypotheses.

1. Direct Gradient Analysis.....2

2. Few species.....4

4. Monotonic responses to gradients (low beta).....Linear regression

4. Nonmonotonic responses to gradients (high beta).....Generalized linear models

2. Many species.....5

5. Monotonic responsesRDA

5. Nonmonotonic responses.....6

6. concerned about arch effect.....DCCA

6. not concerned about arch effect.....CCA

1. Indirect Gradient Analysis.....3

3. Only distance values are available.....7

7. Monotonic responsesPCoA

7. Nonmonotonic responses.....NMDS

3. Raw data available.....8

8. Monotonic responses9

9. Variables noncommensurate.....PCA - corr. matrix

9. Variables commensurate.....PCA - cov. matrix

8. Nonmonotonic responses.....10

10. Feel OK about prespecifying number of dimensions, not worried about local optima, not interested in species scores.....NMDS

10. Not as above, but willing to accept either arch effect or detrending/rescaling.....11

11. Don't like arch, detrending OKDCA

11. Arch OK, or only interested in axis 1.....CA

Dichotomous Key for Ordination Methods

Not 100% accurate, but a good place to start.

(Palmer 1998)
<http://www.okstate.edu/artsci/botany/ordinate/index.html>

Canonical Form

In mathematics, a *canonical form* is the simplest and most comprehensive form to which certain functions, relations, or expressions can be reduced without loss of generality.

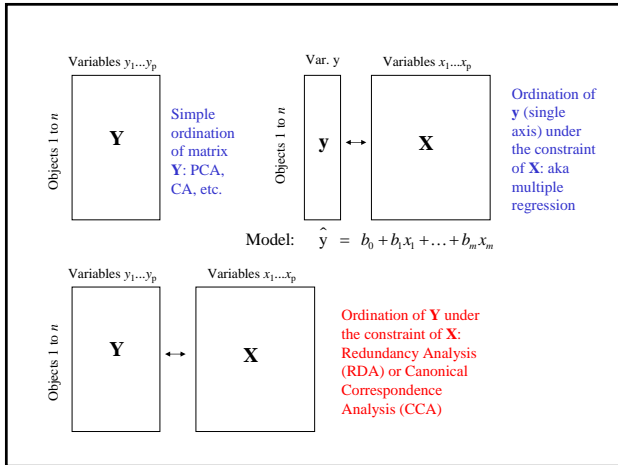
For example, the canonical form of a covariance matrix is its matrix of eigenvalues.

In general, most methods of canonical analysis employ eigenanalysis (some extensions have been described using NMDS).

Canonical Analysis

Canonical analysis combines the concepts of ordination and regression. It involves a response matrix **Y** and an explanatory matrix **X**. (See next slide.)

Like previous ordination methods, canonical analysis produces orthogonal axes from which scatter diagrams may be plotted.



Problems of canonical analysis can be represented via a partitioned covariance matrix resulting from the fusion of **Y** and **X** data sets and producing a joint dispersion matrix S_{Y+X} ...

$$S_{Y+X} = \begin{bmatrix} s_{y_1, y_1} & \dots & s_{y_1, y_p} & s_{y_1, x_1} & \dots & s_{y_1, x_m} \\ \vdots & & \vdots & \vdots & & \vdots \\ s_{y_p, y_1} & \dots & s_{y_p, y_p} & s_{y_p, x_1} & \dots & s_{y_p, x_m} \\ \hline s_{x_1, y_1} & \dots & s_{x_1, y_p} & s_{x_1, x_1} & \dots & s_{x_1, x_m} \\ \vdots & & \vdots & \vdots & & \vdots \\ s_{x_m, y_1} & \dots & s_{x_m, y_p} & s_{x_m, x_1} & \dots & s_{x_m, x_m} \end{bmatrix} = \begin{bmatrix} S_{YY} & S_{YX} \\ S_{XY} & S_{XX} \end{bmatrix} = \begin{bmatrix} S_{YY} & S_{YX} \\ S'_{YX} & S_{XX} \end{bmatrix}$$

Submatrices S_{YY} (order $p \times p$) and S_{XX} ($m \times m$) concern each of two sets of descriptors, respectively, where S_{YX} ($p \times m$) and its transpose $S'_{YX} = S_{XY}$ ($m \times p$) account for the covariances among the descriptors of the two groups.

Redundancy Analysis

In redundancy analysis (RDA), each canonical ordination axis corresponds to a direction, in the multivariate scatter of objects (**Y**), which is maximally related to a linear combination of the explanatory variables **X**. A canonical axis is thus similar to a principal component.

Two ordinations of the objects are obtained: (1) linear combinations of the **Y** variables (matrix **F** in PCA), (2) linear combinations of the fitted **Y-hat** variables (matrix **Z**), which are thus also linear combinations of the **X** variables.

RDA preserves the Euclidean distance among objects in matrix **Y-hat** containing values of **Y** fitted by regression to the explanatory variables **X**.

Canonical Correspondence Analysis

Canonical correspondence analysis (CCA) is similar to RDA. The difference is that it preserves the χ^2 distance (as in CA), instead of the Euclidean distance among objects.

Calculations are a bit more complex since \hat{Y} contains fitted values obtained by weighted linear regression of matrix \bar{Q} of correspondence analysis on the explanatory variables \mathbf{X} . As in RDA, two ordinations of the objects are obtained.

Canonical Correlation Analysis

In canonical correlation analysis (CCorA), the canonical axes maximize the correlation between linear combinations of the two sets of variables \mathbf{Y} and \mathbf{X} .

This is obtained by maximizing the among-variable-group covariance (or correlation) with respect to the within-variable-group covariance.

Two ordinations of the objects are again obtained.

Canonical Discriminant Analysis

In canonical discriminant analysis, the objects are divided in to k groups, described by a qualitative descriptor.

The method maximizes the dispersion of the centroids of the k groups. This is obtained by maximizing the ratio of the among-object-group dispersion over the pooled within-object-group dispersion.

Canonical Analysis

Unfortunately, we do not have the time to develop the details of the algebra of each of the 4 methods of canonical analysis previously described. But, you have now gained all of the necessary skills necessary to interpret the details on your own should you need to pursue one of these analyses.

Two excellent sources of information on these methods can be found in Legendre and Legendre (1998), ter Braak and Šmilauer (1998), and Lepš and Šmilauer (2003).

Canonical Analysis

As an alternative to a detailed treatment of mathematics behind each method, I would like to develop some worked examples using one or more software applications.

Let's develop a data set using the number of fish observed at 10 sites along a transect running from the beach of a Caribbean island, with water depths going from 1 to 10 m. The first three sites are on sand and the others alternate between coral and "other substrate" (coded as 0/1).

Tropical Fish Data Set

Site No.	Sp-1	Sp-2	Sp-3	Sp-4	Sp-5	Sp-6	Sp-7	Sp-8	Sp-9	Depth (m)	Coral	Sand	Other
1	1	0	0	0	0	0	2	4	4	1	0	1	0
2	0	0	0	0	0	0	5	6	1	2	0	1	0
3	0	1	0	0	0	0	0	2	3	3	0	1	0
4	1	4	0	0	8	1	6	2	0	4	0	0	1
5	1	5	17	7	0	0	6	6	2	5	1	0	0
6	9	6	0	0	6	2	10	1	4	6	0	0	1
7	9	7	13	10	0	0	4	5	4	7	1	0	0
8	7	8	0	0	4	3	6	6	4	8	0	0	10
9	7	9	10	13	0	0	6	2	0	9	1	0	1
10	5	10	0	0	2	4	0	1	3	10	0	0	0
Σ	60	50	40	30	20	10	45	35	25				

Tropical Fish Data Set

Because we wish to conduct a direct gradient analysis (i.e., we have both species data and environmental data from the same samples), and we have numerous species (9), with roughly monotonic responses (although one may be unimodal; e.g., 7) we select RDA as the method of choice.

RDA is particularly appropriate when the gradients are short and species distributions are linear (or generally monotonic).

The software of choice for this type of analysis is CANOCO. Mathematically, this software is excellent, however its ease of use is not the best and graphics are poor.

Tropical Fish Data Set

	A	B	C	D	E
1	Depth	Coral	Sand	Other	
2	Site1	1	0	1	0
3	Site2	2	0	1	0
4	Site3	3	0	1	0
5	Site4	4	0	0	1
6	Site5	5	1	0	0
7	Site6	6	0	0	1
8	Site7	7	1	0	0
9	Site8	8	0	0	1
10	Site9	9	1	0	0
11	Site10	10	0	0	1

First, create two separate files in Excel. One file should be for the environmental data, the other for the species abundance data. NB: both must have the same number of rows!

	B	C	D	E	F	G	H	I	J	K
1	Sp-1	Sp-2	Sp-3	Sp-4	Sp-5	Sp-6	Sp-7	Sp-8	Sp-9	
2	Site1	1	0	0	0	0	0	2	4	4
3	Site2	0	0	0	0	0	0	5	6	1
4	Site3	0	1	0	0	0	0	0	2	3
5	Site4	1	4	0	0	6	1	6	2	0
6	Site5	1	5	17	7	0	0	6	6	2
7	Site6	9	6	0	0	6	2	10	1	4
8	Site7	9	7	13	10	0	0	4	5	4
9	Site8	7	8	0	0	4	3	6	6	4
10	Site9	7	9	10	13	0	0	6	2	0
11	Site10	5	10	0	0	2	4	0	1	3

Tropical Fish Data Set

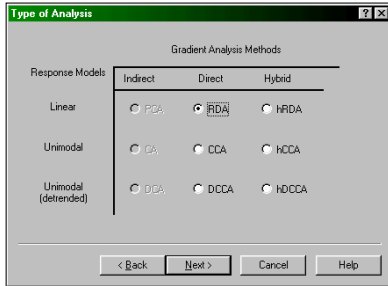
CANOCO requires a very specific data format. Fortunately, it has a utility to put an Excel datasheet automatically in to that format (WCanolmp). Follow the directions provided.



Convert each XLS file to a CANOCO DTA data file and save.

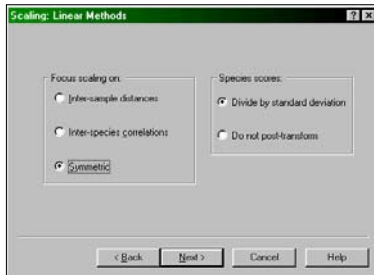
When finished, you will have two DTA files that look like this...

Tropical Fish Data Set



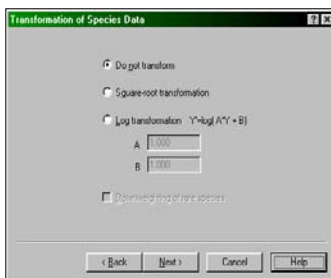
Select the procedure you wish to run. In this case RDA.

Tropical Fish Data Set



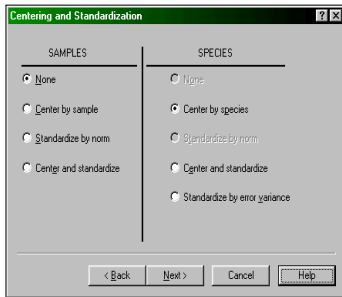
Specify here whether you predominantly want to interpret relationships among samples or among species from the ordination diagram (or whether you prefer a symmetric scaling). Your choice is unimportant if the eigenvalues of the axes of interest are similar.

Tropical Fish Data Set



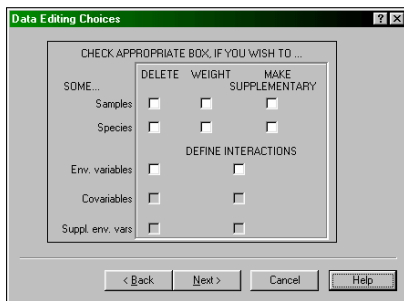
Species abundance values often display a highly skewed distribution. You can prevent a few high values from unduly influencing the ordination by transforming the data. This is probably not necessary for our example data set.

Tropical Fish Data Set



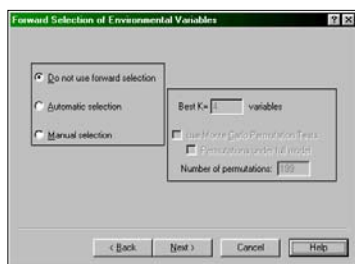
Specify here whether you want to center and/or standardize the species data table by samples and/or by species (rows and columns of the species data file, respectively). Ordinary RDA (based on a covariance matrix) is obtained by centering by species only. Each species is then weighted by its variance. Standardized PCA/RDA (based on a correlation matrix) is obtained by centering and standardization by species.

Tropical Fish Data Set



Checking a box here allows you to specify later which samples or species you wish to delete, weight, or make supplementary (= passive).

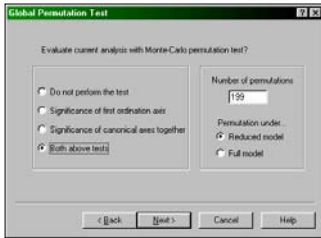
Tropical Fish Data Set



Forward selection is useful for ranking environmental variables in importance for determining the species data or for reducing a large set of environmental variables. Variables can be selected automatically or manually. In automatic selection, the K best variables are selected sequentially on the basis of maximum extra fit. You can limit the number of selected variables (K).

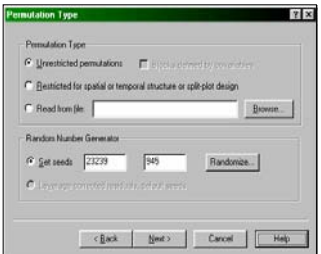
Optionally, the statistical significance of each selected variable can be judged by a Monte-Carlo permutation test.

Tropical Fish Data Set



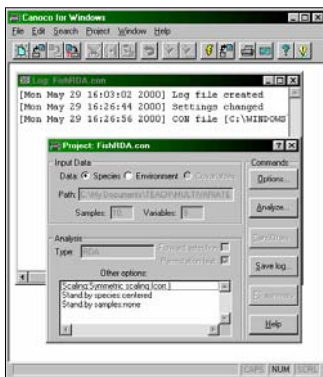
Specify here whether you want to determine the statistical significance of the relation between the species and the whole set of environmental variables, given the covariables. Two test statistics are available: one based on the first canonical eigenvalue and one based on the sum of all canonical eigenvalues. The resulting tests determine the significance of the first ordination axis and that of all canonical axes together, respectively.

Tropical Fish Data Set



Experimental design and sampling design determine the appropriate permutation type. Unrestricted permutation is appropriate for completely randomized and randomized block designs and for simple random sampling and stratified random sampling. It is also the default for studies without any additional structure. In designs with blocks or strata, exchanges of samples between the blocks or the strata must be excluded. This is achieved by checking Blocks here and defining them by covariables later. If samples are taken in a number of different locations, defining location as blocks provides a test for common within-location variation.

Tropical Fish Data Set

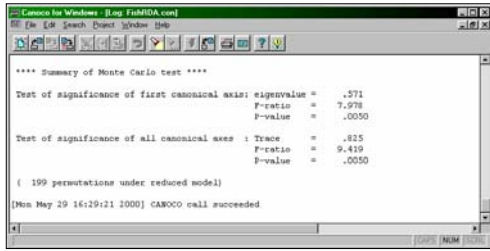


Click Finish at the end of the wizard and specify where to save the canoco output file (*.CON).

When finished, click the analyze button to run the analysis.

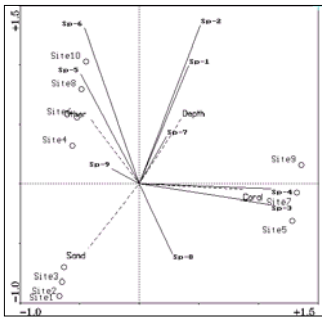
Data will appear in the log file (middle window).

Tropical Fish Data Set



The last panel summarizes the results of the global permutation tests to judge the significance of the relation between species and environment in the data sets provided. Both the first and second canonical eigenvalues have significant F-ratios.

Tropical Fish Data Set



After much ado, one can generate a triplot using the [very unfriendly DOS application] named Canodraw.

The result is shown here. Species, sites, and environmental variables are all shown on the same plot!

Discriminant Analysis

A common situation arises in EEB applications where one starts with an already known grouping of objects, and one wishes to assess how well a group of quantitative descriptors can explain the object groups. Thus, the problem is no longer how to define or delineate groups, but rather how to interpret them. This is the realm of *discriminant analysis*.

Discriminant analysis is a method of linear modeling, like analysis of variance, multiple regression, and canonical correlation analysis. DA is frequently used in systematics.

Discriminant Analysis

DA proceeds in two steps:

(1) It first tests for the differences in the explanatory variables (\mathbf{X}), among the predefined groups. This part of the analysis is identical to the overall test performed in the MANOVA.

(2) If the test supports the alternative hypothesis of significant differences among groups in the \mathbf{X} variables, the analysis proceeds to find the linear combinations (called *discriminant functions*) of the \mathbf{X} variables that best discriminate the groups.

Discriminant Analysis

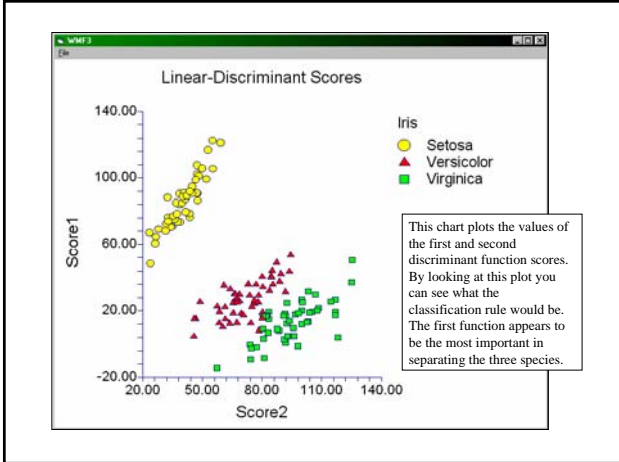
Like one-way ANOVA, discriminant analysis considers a single classification criterion (i.e., division of the objects into groups) and allows one to test whether the explanatory variables can discriminate among the groups. Testing for differences among group means in DA is identical to ANOVA for a single explanatory variable and to MANOVA for multiple explanatory variables.

When it comes to modeling, i.e., finding the linear combinations of the variables (\mathbf{X}) that best discriminate among the groups, DA is a form of "inverse analysis" where the classification criterion is considered to be the response variable (\mathbf{y}) whereas the quantitative variables are explanatory (matrix \mathbf{X}).

Discriminant Analysis

Note that discriminant analysis (DA) is also called *canonical variates analysis* (CVA). This method was first proposed by Fisher (1936) where he published the now famous data set where he described the morphology of 150 specimens of irises (Iridaceae) using 4 measured flower characters (lengths and widths of sepals and petals) belonging to three species.

Again, in the interest of time, we will bypass the mathematical treatment of DA and work through the iris data set using a software application (NCSS).



The End!
 Stay Tuned for Miles.
 Have a Nice Summer.
