

17/11/02

CANOCO Reference Manual and CanoDraw for Windows User's Guide

**Software for Canonical Community Ordination
(version 4.5)**

Cajo J. F. ter Braak and Petr Šmilauer

Copyright © 1997
by Cajo J. F. ter Braak and Petr Šmilauer
All rights reserved.
No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or by any information storage or retrieval system, without the prior written permission of the copyright owner.

Biometris, Wageningen and České Budějovice, 2002

17/11/02

To Margriet and Marie

Suggested Citation:

ter Braak, C. J. F. & Smilauer, P. 2002: CANOCO Reference manual and CanoDraw for Windows User's guide: Software for Canonical Community Ordination (version 4.5). Microcomputer Power (Ithaca, NY, USA), 500 pp.

Printed in the United States of America

© 1998-2002 by Biometris, Wageningen University and Research Centre, Wageningen, The Netherlands (Chapters 1-8) and by Petr Smilauer (Chapters 9-14)

Cover design: Margriet Stapel and Anja ten Hove with thanks to the Dutch Butterfly Conservation

Table of Contents

1.	INTRODUCTION	11
1.1	Research questions and the ordination methods in CANOCO	11
1.2	Canoco for Windows	12
1.3	Organization of the chapters	12
1.4	Practical information	13
1.5	Further reading	14
1.6	Acknowledgments	15
2.	GETTING STARTED	17
2.1	Installation	17
2.2	The first run: DCA	20
2.3	Importing data from a spreadsheet	26
2.4	The second run: CCA	27
2.5	The third run: RDA	28
3.	BACKGROUND THEORY	33
3.1	General objective	33
3.2	Terminology	33
3.3	Models, methods, and algorithms	35
3.4	The two faces of (canonical) correspondence analysis	39
3.5	Interpreting ordination diagrams	39
3.6	Supplementary species, samples, and environmental variables	42
3.7	Permutation tests	42
3.7.1	Introduction	42
3.7.2	The basic idea	43
3.7.3	Permutation type: how are samples shuffled?	44
3.7.4	What is shuffled?	47
3.7.5	Model-based permutations	48
3.7.6	Multifactorial analysis of variance	51
3.8	RDA and CCA as regression procedures: reduced-rank regression	54
3.9	Compositional data	56
3.9.1	Introduction	56
3.9.2	Compositional data not containing zeroes: log-ratio analysis	57
3.9.3	Compositional data containing zeroes: CA, DCA and CCA	58
3.9.4	The two faces of compositional data revisited	60
3.10	Nominal response data	61
3.11	Canonical Variates Analysis (CVA) and discriminant analysis	62
3.12	Principal coordinates analysis	64
3.13	The stability of ordination axes	64
4.	DATA INPUT	67
4.1	Importing from spreadsheets: CanoImp, WCanoImp	67
4.1.1	Processing capacity	68
4.1.2	Properties of the output files	69

4.2	Merging data tables with CanoMerge program	69
4.3	Linking samples in different data files	71
4.4	Native CANOCO formats	71
4.4.1	Full format	71
4.4.2	Condensed format	75
4.4.3	Free format	78
5.	PROJECT SETUP AND ANALYSIS IN CANOCO FOR WINDOWS	81
5.1	Introduction	81
5.1.1	How to use Canoco for Windows	82
5.1.2	The Canoco for Windows workspace on startup	83
5.1.3	The Canoco for Windows workspace with project and log views	84
5.2	Project View	85
5.3	Selecting data sets and analysis type	85
5.3.1	Available Data	86
5.3.2	Data Files	87
5.3.3	Type of Analysis	88
5.4	Number of canonical axes and detrending method	89
5.4.1	Canonical Axes	89
5.4.2	Detrending Method	90
5.5	Scaling of ordination scores	90
5.5.1	Scaling: Linear Methods	91
5.5.2	Scaling: Unimodal Methods	92
5.6	Transformation of the species data	93
5.6.1	Centering and Standardization	93
5.6.2	Transformation of Species Data	94
5.7	Data editing options	95
5.7.1	Data Editing Choices	95
5.7.2	Delete Items	96
5.7.3	Set Weights for Samples / Species	97
5.7.4	Supplementary Samples / Species	98
5.7.5	Interactions of Variables	99
5.8	Forward selection	100
5.8.1	Forward Selection of Environmental variables	100
5.8.2	Forward Selection report	101
5.8.3	Manual Forward Selection	102
5.9	Global significance tests	103
5.9.1	Global Permutation Test	103
5.10	Specifying the randomization model	104
5.10.1	Permutation Type	104
5.10.2	Definition of Blocks	105
5.10.3	Permutation Restrictions	106
5.10.4	Grid Dimensions	107
5.10.5	Split-Plot Design I	108
5.10.6	Split-Plot Design II	109
5.11	Saving the project	111
5.12	Running the analysis and saving the log	111
5.13	Creating ordination diagrams	112
5.14	Specifying program options	113
6.	RESULTS OF THE ANALYSIS	115
6.1	Introduction	115

6.2	Log-window	115
6.2.1	Log of reading of the project CON-file and the data files	115
6.2.2	Collinear environmental variables and collinear covariables	118
6.2.3	Outliers in the explanatory variables: check on influence	119
6.2.4	Correlation matrix, means, standard deviations, inflation factors	120
6.2.5	Summary of the ordination	122
6.2.6	Global permutation test	126
6.2.7	Forward selection of environmental variables	129
6.3	Solution file	132
6.3.1	Introduction	132
6.3.2	Relationships between ordination scores	134
6.3.3	Weights and the eigenvector sample scores	156
6.3.4	Species scores	157
6.3.5	Sample scores $\{ x_i^* \}$ that are derived from the species	160
6.3.6	Regression coefficients and associated t-values	163
6.3.7	Sample scores $\{ x'_i \}$ that are derived from the environment	166
6.3.8	Inter-set correlations of environmental variables with axes	167
6.3.9	Biplot scores of environmental variables	168
6.3.10	Centroids of environmental variables in the ordination diagram	171
6.3.11	Ordination diagnostics	174
6.3.12	t-Value biplot	179
6.4	Species-by-environment table	182
7.	CONSOLE VERSION OF CANOCO	185
<hr/>		
7.1	Introduction	185
7.1.1	Differences with Canoco for Windows	185
7.1.2	Differences with CANOCO 3.x	186
7.2	Ways to answer the questions	186
7.3	Initialization file	187
7.4	Introductory questions	188
7.5	Selecting data sets and analysis type	189
7.6	Number of canonical axes and detrending options	192
7.7	Forward selection of environmental variables	194
7.8	Scaling of ordination scores	195
7.9	Ordination diagnostics	197
7.10	Omitting samples and selecting explanatory variables	197
7.11	Transformations of species data	200
7.12	Centering and standardization in linear methods	204
7.13	Output options	206
7.14	Additional analyses	208
7.15	Supplementary environmental variables	209
7.16	More ordination axes	209
7.17	Monte Carlo permutation tests	209
7.17.1	Introduction	209
7.17.2	Unrestricted permutation; common questions	210
7.17.3	Specifying blocks	211
7.17.4	Restricted permutation types	212
7.17.5	Restricted permutation types within blocks	215
7.18	Forward selection dialogue	216
7.19	Canoco project files	217
7.20	Nonstandard analyses	219
8.	CANOCO EXAMPLES	221

8.1	Introduction	221
8.1.1	How to analyze the examples	222
8.2	Examples from “Unimodal models to relate species to environment” re-analyzed	223
8.2.1	Example SPIDER1 - A niche study by CA and DCA	226
8.2.2	Example SPIDER2 - A niche study by CCA	230
8.2.3	Example DYKE - CCA of presence / absence data	234
8.2.4	Example ALGAE - A study of a pollution gradient	237
8.2.5	Example DUNEBOOK - CCA and RDA on observational data	241
8.2.6	Example WEEDS - A multi-species trend surface	245
8.2.7	Example SEASHORE - A vegetation succession study	247
8.2.8	Example EPIALGAE - Conditional and marginal effects	249
8.2.9	Example STREAMS - Partial CCA on macro-invertebrates	251
8.2.10	Example VEGCHANG - A study of change in time and space	252
8.2.11	Example DISEASES - Reduced-rank regression	255
8.3	Examples of significance tests	258
8.3.1	Example DUNETEST - Simple and partial tests	258
8.3.2	Example PLOUGH - A randomized block experiment	262
8.3.3	Example E40 - A multifactorial experiment with fixed factors	265
8.3.4	Example LINES - Line transect(s) across the seashore	270
8.3.5	Example GRID - Samples in a rectangular spatial layout	272
8.3.6	Example SPLITPLT - A split-plot analysis	275
8.3.7	Example BACI1SPE - A univariate, unreplicated BACI analysis	279
8.3.8	Example BACIMSPE - A multivariate, unreplicated BACI analysis	282
8.3.9	Example BACI3SIT - A multivariate BACI analysis with three sites	284
8.3.10	Example BACI_REP - A multivariate, replicated BACI analysis	286
8.3.11	Example PRC_SIM - Displaying time-dependent effects by PRC	287
8.3.12	Example PRC - Testing time-dependent effects by PRC	291
8.3.13	Example WELCH - A design-based test of interaction	293
8.4	Other ordination methods that are also available in CANOCO	295
8.4.1	Introduction	295
8.4.2	Example LOGRATIO - Log-ratio analysis of compositional data	295
8.4.3	Example CVA - Canonical Variates Analysis	298
8.4.4	Example MULTREGR - Multiple regression with CANOCO	301
9.	PROGRAM PRCOORD	305
9.1	Working with PrCoord	305
9.2	Implementation details	307
9.3	How to calculate db-RDA with Canoco software	308
10.	CANODRAW INTRODUCTION	312
11.	CANODRAW CONCEPTS	314
11.1	Types of items in Canoco and CanoDraw projects	314
11.2	Indices of items	315
11.3	Window types	315
11.4	Graph types	317
11.5	Graph object types	318
11.6	Graph scaling and coordinate units	319
11.7	Limiting contents of graphs	320
11.8	Application-wide and project-specific options	320

12.1	File	322
12.1.1	New Project	322
12.1.2	Open Project	323
12.1.3	Open Graph	324
12.1.4	Close	324
12.1.5	Save	324
12.1.6	Save As	325
12.1.7	Export G	325
12.1.8	Visual Attributes G	327
12.1.9	Print G	328
12.1.10	Print Preview G	328
12.1.11	Print Setup	328
12.1.12	Recently used files	329
12.1.13	Exit	329
12.2	Edit	329
12.2.1	Undo	329
12.2.2	Redo G	330
12.2.3	Cut P	330
12.2.4	Copy	330
12.2.5	Paste P	330
12.2.6	Delete	330
12.2.7	Change text G	330
12.2.8	Make label vertical / horizontal G	331
12.2.9	Copy labels to Clipboard	331
12.3	View	331
12.3.1	Diagram Settings	331
12.3.2	Visual Attributes	348
12.3.3	Workspace Settings	349
12.3.4	Properties Sheet G	351
12.3.5	Tree view G	354
12.3.6	Zoom G	354
12.3.7	Project Details P	354
12.3.8	Bars	356
12.4	Project	357
12.4.1	Settings	357
12.4.2	Nominal variables	365
12.4.3	Classify	366
12.4.4	Define Groups of	373
12.4.5	Define Series of	376
12.4.6	Suppress	378
12.4.7	Enforce	379
12.4.8	Import variables	379
12.4.9	Export Statistics	382
12.4.10	Manage graphs	384
12.5	Create	385
12.5.1	Simple Ordination Plot	385
12.5.2	Scatter Plots	386
12.5.3	Biplots and Joint Plots	390
12.5.4	Triplots	394
12.5.5	Attribute Plots	395

12.5.6	Recreate graph G	403
12.5.7	Range of axes G	403
12.5.8	Lock legend G	404
12.5.9	Unlock legend G	404
12.6	Object G	404
12.6.1	Select Suchlike	404
12.6.2	Select Similar	405
12.6.3	Lock selected	405
12.6.4	Unlock all	405
12.6.5	Graph tool	406
12.7	Window	407
12.7.1	Cascade	407
12.7.2	Tile	407
12.7.3	Arrange Icons	407
12.7.4	Close all	407
12.7.5	Close graphs of active project	407
12.7.6	Open graph project G	407
12.7.7	List of windows	408
12.8	Help	408
12.8.1	Help Topics	408
12.8.2	About CanoDraw	408
12.8.3	Tip of the Day	408
13. WORKING WITH CANODRAW		409
13.1	Selecting graph objects	409
13.1.1	Manual selection with mouse	409
13.1.2	Rubber-band selection	409
13.1.3	Selection by example: similar and "suchlike" items	410
13.1.4	Selecting whole class of items	410
13.1.5	Selecting graph object by its label and vice versa	410
13.1.6	Locking and unlocking objects	410
13.2	Finding particular object	411
13.3	Modifying graph contents	411
13.4	Creating graph legend	413
13.5	Exploring graph contents	414
13.6	Fitting regression models	418
13.6.1	Generalized Linear Models (GLM)	418
13.6.2	Generalized Additive Models (GAM)	421
13.6.3	Loess (locally-weighted regression) models	.. 422
13.6.4	Regression diagnostics in CanoDraw	423
14. CANODRAW EXAMPLES		427
14.1	SPIDER1	427
14.2	SPIDER2	435
14.3	DYKE	441
14.4	DUNEBOOK	443
14.5	WEEDS	451
14.6	SEASHORE	454
14.7	DISEASES	458
14.8	PRC_SIM	461

15.	REFERENCES	467
16.	APPENDIX A: THE EXTENDED DUNE MEADOW DATA	475
17.	APPENDIX B: MATHEMATICAL DERIVATIONS	479
17.1	Environmental biplot scores represent covariances or weighted averages	479
17.2	Environmental centroid scores represent class means or class totals	480
17.3	Sum of all canonical eigenvalues (trace)	482
18.	APPENDIX C: FORMAT OF (W)CANOIMP FILES	483
19.	APPENDIX D: CANODRAW SOFTWARE SETUP	484
20.	INDEX	485
21.	LIST OF FIGURES	490
22.	LIST OF TABLES	496

1. Introduction

1.1 Research questions and the ordination methods in CANOCO

Knowing about biological communities and their relation to the environment is both fascinating and important for human beings. Ordination can help biologists infer relations from large data sets on plant and animal communities and their environment. The data may arise from the field or the laboratory, and can be observational or experimental. Ordination analysis with CANOCO can provide insights into the structure of biological communities and into the impact of natural and human-induced environmental disturbances on biological assemblages. CANOCO has been used in the past to help answer research questions such as:

- How does the vegetation develop on abandoned cultivation sites within a tropical rain forest?
- How does agricultural management practice affect meadow vegetation?
- What are the effects on forest undergrowth if liming is used to mitigate the effects of acid rain?
- Do diatoms respond so strongly to lake pH that they can be used to monitor trends in acidity?
- How long does it take before an invertebrate community recovers from an application of the insecticide chlorpyrifos? How does the time to recover depend on the concentration?

CANOCO contains four main classes of ordination methods

1. Methods to describe the structure in a single data set. For instance, the structure of a biological community or the correlation structure of a set of environmental variables (ordination, indirect gradient analysis).
2. Methods to explain one data set by another data set. For instance, to explain or to predict species abundances from environmental data (canonical ordination, direct gradient analysis).
3. Methods to explain one data set by another data set, after accounting for variation explained by a third data set (covariable data). For instance, to explain species abundances from environmental data, adjusted for observer and seasonal effects (partial canonical ordination).
4. Methods to describe the structure in a single data set after accounting for variation explained by a second data set (covariable data). For instance, the community structure adjusted for observer and seasonal effects (partial ordination).

Within each of these four classes, you can choose between three response models: a linear model, an unimodal model, and an unimodal model with detrending. The basic ordination methods, Principal Component Analysis, Correspondence Analysis, and Detrended Correspondence Analysis are thus extended to canonical, partial, and partial canonical forms. These methods are also effective outside biology.

The principle output of an analysis consists of:

- an ordination diagram with a numerical summary of the variance explained,
- the variance explained by the environmental variables, if present,
- the statistical significance of the environmental variables, and
- the statistical significance of the first ordination axis of an analysis with explanatory variables.

The ordination diagram graphically represents the community structure and, in canonical analyses, the community response to the environmental variables. The significance is determined by permutation tests. Experimental design and sample design determine the appropriate permutation type. Also, the effects of environmental variables after accounting for specified covariables can be tested (a partial test). Significance tests guard against over-interpretation of canonical ordination diagrams.

1.2 Canoco for Windows

The first version of CANOCO was developed in 1985 on a VAX computer as an extension of the computer program DECORANA (Hill 1979) to include canonical correspondence analysis and its detrended form (Ter Braak 1986). The second version, released in 1987, also included principal components analysis and its canonical form which originates from Rao (1964) and is now known as redundancy analysis (Van den Wollenberg 1977) or reduced-rank regression (Davies & Tso 1982). Onno van Tongeren gave, for the time, a user-friendly interface to CANOCO, and helped to migrate the program to MS-DOS. Petr Šmilauer developed CanoDraw for drawing ordination diagrams. In the third version, released in 1990, ordination diagnostics and forward selection were added and the facilities for permutation tests were greatly extended.

The versions of CANOCO starting from 4.0 have a Windows user interface under Microsoft Windows™ 98, ME, Windows NT, 2000, or XP. A console version of CANOCO 4.x continues to be available, also for other operating systems. The facilities for permutation tests are further extended in versions 4.x to include tests for split-plot designs and related multi-level designs, and the scaling of the species scores was also improved. The documentation has been extended and hopefully improved. CANOCO 4.5 is now bundled with CanoDraw for Windows, which replaces the two programs used in version 4.0 (CanoDraw 3.1 and CanoPost). Data can be imported from spreadsheets with the Windows utility, WCanoImp, and several data tables can be combined with the CanoMerge program. Program PrCoord provides an easy access to the distance-based RDA method for CANOCO users. The whole package is called *Canoco for Windows*.

1.3 Organization of the chapters

The Getting Started chapter describes the installation of the software under Windows operating systems. Furthermore, it provides tutorial sessions on how to carry out ordination analyses, how to make ordination diagrams, and how to import data from spreadsheets under Windows. The next chapter gives the background theory, with topics such as interpreting ordination diagrams, permutation tests, the stability of ordination axes, and the analysis of compositional data. The emphasis here is on aspects that are new or that deserve more attention than given in the Data Analysis textbook (Jongman et al. 1987), the Unimodal Models booklet (Ter Braak 1987) or its extended version (Ter Braak 1996), which is included with the Canoco for Windows package.

CANOCO requires input data in a special format. Chapter 4 describes how to convert data in spreadsheets to one of the CANOCO formats by using the program WCanoImp and how to use the CanoMerge program. It also gives examples of valid CANOCO formats, in case you need to prepare the data files yourself.

Chapter 5, on project setup and analysis in Canoco for Windows, describes the Windows user interface with which you can define, run and modify ordination analyses. In order to run an ordination analysis you must create a CANOCO project with information on your data sets and

the chosen method of ordination. Projects are created and modified with the Project Setup Wizard, which guides you through the available options. How to interpret the results of the analysis is described in Chapter 6. This Results chapter is also relevant if you run the console version of CANOCO. The console version is described in Chapter 7. This chapter is also of interest if you wish to understand CANOCO project files, because these have the same format as the answer file in the console version.

Chapter 8 shows how the results of the ordination examples in the Unimodal Models booklet can be obtained with CANOCO and Canoco for Windows. The Examples chapter also illustrates with real data how the permutation test facilities can be used to test ecological hypotheses.

Chapter 9 describes how to use the PrCoord program to analyze your data by Principal Coordinates Analysis (PCO) and how to perform distance-based redundancy analysis (db-RDA).

Program CanoDraw for Windows is described in the following Chapters 10 to 15.

1.4 Practical information

Canoco for Windows requires Microsoft Windows 98 or NT 4.0 or later versions with at least 32 MB of internal memory and 64 MB of free disk space for virtual memory. The maximum data size that can be analyzed is: 25 000 samples (n), 5000 species (m), 2000 covariables (p), 1000 environmental variables (q). The maximum number of nonzero values in the species data is 750 000, the maximum value for $p*n$ is 1 000 000, and the maximum value for $(q-8)*n$ is 500 000. CanoDraw for Windows is able to visualize analysis results for all datasets that can be analyzed with Canoco for Windows program.

Canoco for Windows consists of the following components:

- canoco.exe console version of the CANOCO 4.0 program
- canowin.exe Windows version of the CANOCO program, supported by
 - cwinterf.dll support dynamic link library (DLL)
 - can45_64.dll CANOCO engine dynamic link library (DLL)
 - canodat.dll data parsing dynamic link library (DLL)
- canoimp.exe console version of the CanoImp program
- wcanoimp.exe Windows version of the CanoImp program (WCanoImp), using
 - animdll.dll support dynamic link library (DLL) for WCanoImp program
- canodrw4.exe CanoDraw for Windows 4.0 program
 - canodatc.dll support dynamic link library (DLL) for reading data files
 - loess.dll dynamic link library (DLL) supporting Loess models
- canomerg.exe CanoMerge program (uses the canodatc.dll library)
- prcoord.exe PrCoord program (uses the canodatc.dll library)
- help files (.HLP and .CNT) for Canoco for Windows, CanoDraw for Windows, WCanoImp, CanoMerge, and PrCoord programs.

The console version of CANOCO 4.x is also available as an executable for MS-DOS, the Apple Macintosh and OS/2. On these systems, CanoDraw 4.0 and the Windows versions of Canoco and other programs can be run only if execution of Windows programs is supported by an emulator. With the CANOCO source code package and a FORTRAN 77 or 90 compiler, the console version of CANOCO 4.5 can be installed on other systems as well. All these items

require a license for the Canoco for Windows package. Please ask the CANOCO distributors for further details.

Technical support is provided by your CANOCO distributor.

Canoco for Windows is distributed by:

Microcomputer Power

Attention: Richard E. Furnas

111 Clover Lane Dept. W1

Ithaca NY 14850-4930, USA.

Fax (607)-272-0782, phone (607)-272-2188,

E-mail: FurnasR@microcomputerpower.com

Web-address: <http://www.microcomputerpower.com/>

and

SCIENTIA Software

Attention: Janos Podani

Box 658

H-1365 Budapest, Hungary.

Fax +36-1-3812-188, phone +36-1-3812-293

E-mail: podani@ludens.elte.hu

Web-address: <http://ramet.elte.hu/~scientia>

1.5 Further reading

The basic theory behind CANOCO can be found in the Ordination chapter of Jongman et al. (1987). A summary is provided by Ter Braak & Prentice (1988, reprinted in *Unimodal Models*, pp 93-138). The extended *Unimodal Models* booklet, which is distributed with Canoco for Windows, describes the state of art on interpreting ordination diagrams on pages 139-188 and provides the mathematical theory behind permutation tests (pp 217-223) and reduced-rank regression (pp 225-258). A recent and comprehensive account of the theory and its ecological applications is given by Legendre & Legendre (1998). The annotated bibliography by Birks et al. (1996) is a rich source for finding interesting applications of canonical ordination techniques. It lists over 300 publications. A basic reference for experimental design and the analysis of variance is Underwood (1996). The ordination web pages of Mike Palmer contain large amount of useful information about ordination methods, as well as links to other information sources: <http://www.okstate.edu/artsci/botany/ordinate/>

A discussion forum for ordination topics in community ecology is provided by ORDNEWS, a listserv moderated by Steve Bousquin. To subscribe, send an e-mail to: listserv@colostate.edu, do not include a subject in the message and as the only text, "subscribe ordnews your-name". Replace "your-name" with your actual name.

The place to look for new information about Canoco for Windows and CanoDraw for Windows are the following WWW pages:

<<http://www.canoco.com>>

1.6 Acknowledgments

Many people helped in many different ways to realize Canoco for Windows and its documentation. A special word of thanks goes to John (H.J.B.) Birks for his meticulous reading of the whole manuscript. Margriet Stapel co-authored Chapter 5. Pierre Legendre, Daniel Borcard and Han van Dobben gave useful comments on the Background Theory chapter in particular. Research with Marti Anderson, Paul van den Brink and Han van Dobben was important for the Permutation Theory chapter. Richard Furnas helped to improve the console version and the examples, and made the Macintosh version. Pieter Vereijken checked the formulae on the relations between ordination scores in section 6.3. Mike Palmer beta-tested Canoco for Windows and commented on the online help. Thank you all for your contributions. Last but not least we thank Mark Hill with whom it all started.

Numbers are fascinating, but on the cover it is the butterflies and flowers that really allure the eye. We wish all CANOCO users interesting research with fruitful results.

2. Getting started

2.1 Installation

Canoco for Windows is distributed on one installation compact disc (CD). Before using the software on your computer, you must install it with the supplied installation program. To remove cleanly the software from your system, you must use the un-install procedure. These procedures are described in more detail in this section.

To install Canoco on your computer, use the following steps:

- If you install on the Windows NT, Windows 2000, or Windows XP platforms, you must log into your computer using an administrative account (typically with the *Administrator* user name).
- When you insert the installation CD into your CD drive, the installation program can startup automatically (depending on the settings used on your computer).
- Alternatively, you can start the installation program in the following way: from the **Start** menu select the submenu **Settings** and the command **Control Panel** (see Figure 2-1). The Control Panel window appears and you must run the **Add/Remove Programs** command in it (by double-clicking its icon or by selecting the **Open** command from its popup menu). In the **Add/Remove Programs** window, select the **Add New Programs** tab (also called **Install/Uninstall** in the older versions of operating systems). *Note that if you are upgrading from the version 4.0 of Canoco for Windows package, it is shown in the list at the bottom of this windows. You do **not** need to uninstall it before installing the new version.* Then click on the **CD or Floppy** button (**Install...** button in older versions). Click the **Next** button when asked to do so. After searching the floppy and CD drives, Windows should find the installation program (named *setup.exe*), as shown in Figure 2-2. Click the **Finish** button to start the Canoco for Windows installer.

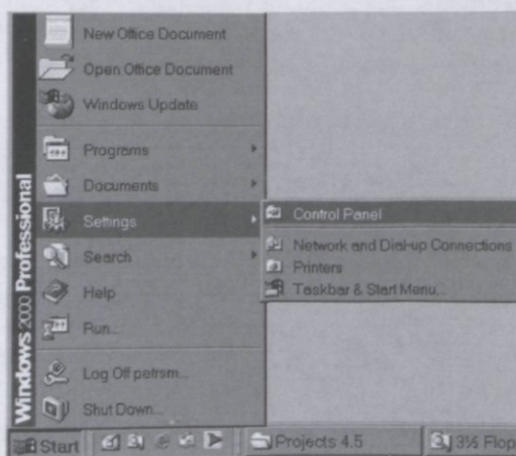


Figure 2-1 Where to start the installation of Canoco for Windows.

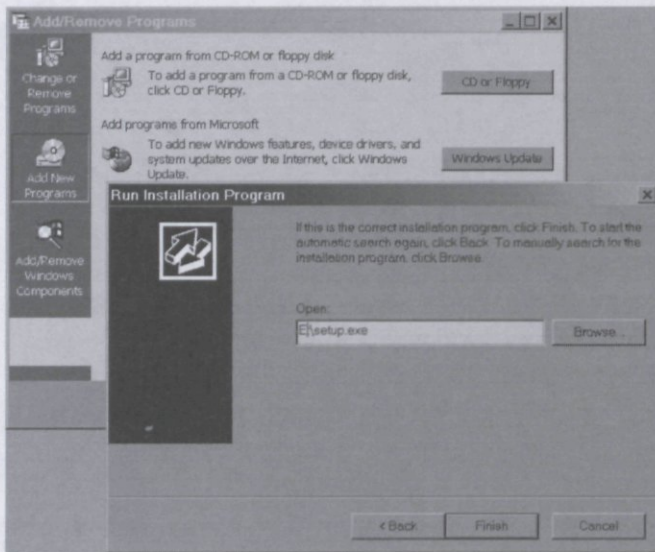


Figure 2-2 Starting the installer program.

- In a short time, you will see the **Welcome** dialog box where you have the possibility to cancel the installation or to proceed by selecting the **Next** button. Note that this installation program can install Canoco for Windows, CanoDraw for Windows, as well as the other programs of the package. The content of the **install.txt** file (which is later stored in the Canoco directory) is displayed on the next page (**Installation Instructions**). Important comments about the installation requirements, as well as the solution for eventual installation problems are given in this file.
- After you leave this window (using the **Next** button), an **Registration Information** page appears (Figure 2-3). Here you must specify a valid user name and name of your company (if you are not affiliated to any institution, specify here **Private**), as well as your official serial number. This number is usually displayed on the original envelope for the installation CD or on your copy of Canoco manual. If you have a downloaded time-limited trial version, specify **CAN0000** there. To continue (after filling-in all the three fields), click the **Next** button.

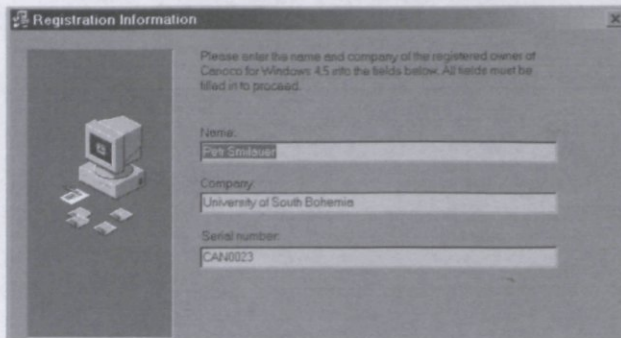


Figure 2-3 User Information dialog box.

- In the next page, you have to select the installation directory where the software (both Canoco for Windows and the accompanying programs) will be installed. If you are upgrading from Canoco for Windows version 4.0, the directory where the original version resides is

displayed. You are advised to keep this choice, in such case. If you do a new installation, the default directory is `c:\canoco` but you can change it either by typing the path or by using the directory-selection dialog box invoked by the **Browse** button. To continue with the installation, click the **Next** button.

- Canoco installation program then asks you whether you want to create backup copies of the replaced files. This page appears only if you are upgrading from older version of Canoco for Windows. You should select the **Yes** option if you think about restoring the older Canoco version in the future.
- In the next step, you must choose the package components to be installed. Only the optional components are displayed: the Canoco program with several additional programs are always installed. The first component (named **Canoco Samples**) contains the sample Canoco projects and data files, most of them are referred to in this manual (Chapter 8); **CanoDraw Program** contains the program CanoDraw for Windows, which is needed to create ordination diagrams from Canoco results. To select or un-select any of the facultative components, make sure that the checkboxes on the left side of their names are checked / un-checked as appropriate (Figure 2-4).

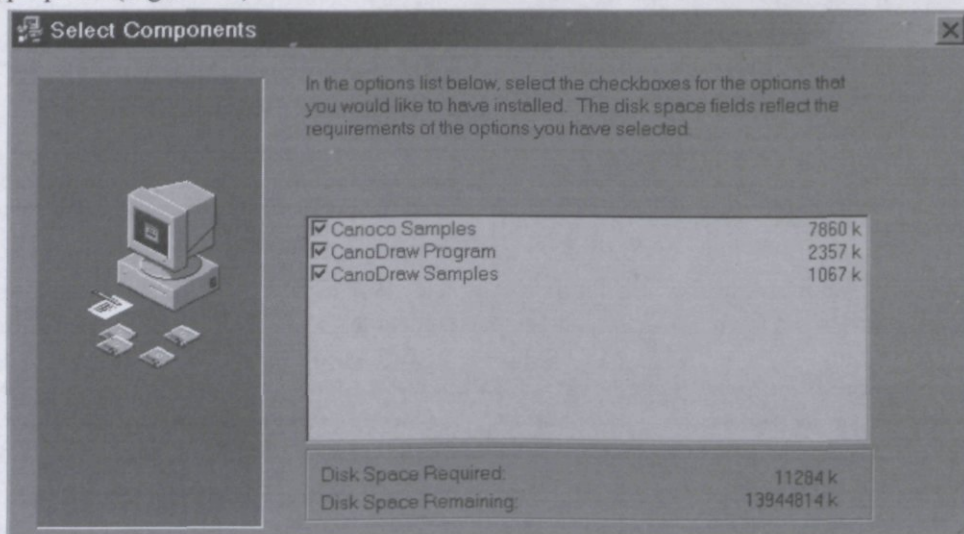


Figure 2-4 Select Components page

- The installation program then asks you to select the Programs menu group, in which the icons used for starting the individual programs are to be placed. The default name offered for the group is **Canoco for Windows**, but you can use the selection box to place the icons into any of the existing program groups or to specify a completely new name. The name of the program group you select here appears as a submenu of the **Programs** menu in the **Start** menu.
- Then, a window with the license agreement for Canoco for Windows software appears. Study its text carefully, as your agreement with it is legally binding. Click on the **Yes** button if you agree, otherwise click the **No** button and the installation is cancelled.
- A similar license agreement is then displayed for the programs CanoDraw for Windows, CanoMerge and PrCoord (all bound with a single agreement). After this page, the **Start Installation** page appears and when you press the **Next** button, files are copied from the installation file into the selected install directory on your hard disk.

- At the end of the installation, the contents of the **README.TXT** file are displayed, where the important, last-minute information is stored together with the contact addresses of the Canoco for Windows retailers.
- The installation program also places two shortcuts in your desktop area, one for starting the Canoco for Windows 4.5 program and the other for starting the CanoDraw for Windows program.

To uninstall the Canoco for Windows package, use the following instructions:

- Select the **Control Panel** item from the **Start / Settings** menu
- In the **Control Panel** window start the **Add/Remove Programs** command. In its window, select the **Change or Remove Programs (Install/Uninstall, in older versions)** tab and look for the **Canoco for Windows 4.5** item in the list of installed applications. Select that item and click the **Change/Remove...** button. After confirming your choice, the Canoco for Windows package is un-installed from your computer, together with related submenus and icons. If the installation directory was modified during the use of the package (typically by creating new files within it), the directory is retained, together with these additional files. The deinstallation program asks you whether you wish a roll-back, which means restoring the original files which were replaced during Canoco for Windows 4.5 installation. If you perform the roll-back, you must always start from the latest upgrade applied to your Canoco installation directory.

2.2 The first run: DCA

This is the first tutorial session with the Canoco for Windows 4.5 program. Throughout this and other tutorials, we will assume that you have installed the whole Canoco for Windows package (including the example data sets) into the **c:\canoco** directory and placed the program icons into the program group named **Canoco for Windows**. Also, to simplify the description, the reader is assumed to use the US version of Windows 2000. In this tutorial, we analyze the structure in the Dune Meadow vegetation data (see **Appendix A**) by Detrended Correspondence Analysis (DCA).

To start the program, select the item **Canoco for Windows 4.5** from the **Start / Programs / Canoco for Windows** submenu. The start-up screen appears while Canoco for Windows loads. Then you see the **Tip of the Day** window overlapping the Canoco for Windows workspace. You can browse through the tips about the use of the program (by means of the **Next Tip** button), but for now close the tips window using the **Close** button.

Study carefully the workspace of the Canoco for Windows program (Figure 2-5): below the title bar is the main menu and below it the toolbar row with buttons representing shortcuts to frequently used menu commands (alternatively, you can use keyboard shortcuts for many of the menu items).

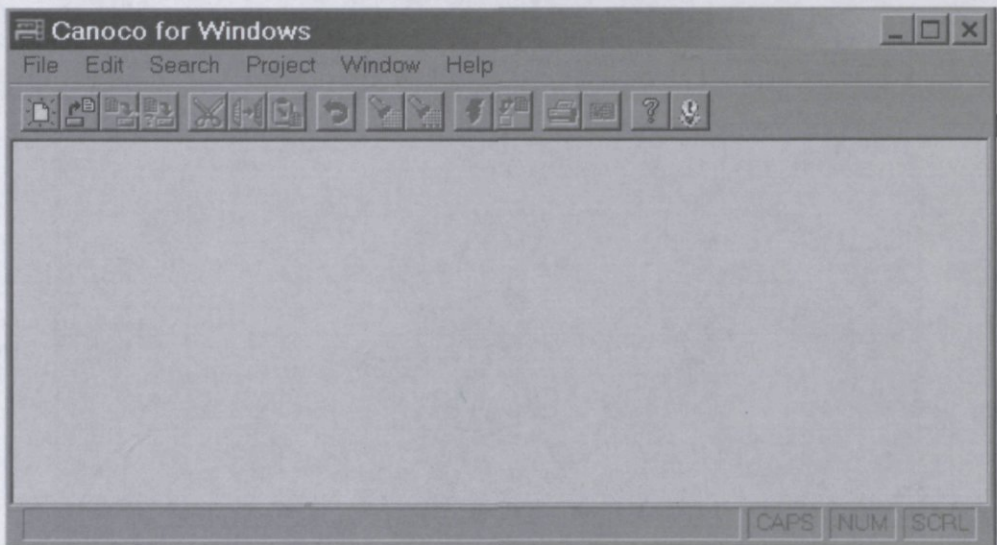


Figure 2-5 The CANOCO workspace.

If you position the mouse pointer over one of the toolbar buttons (for example the one on the far left), you can see the tooltip text “New project” and a more detailed description of the command in the status line at the bottom of the Canoco for Windows workspace (stating *Creates new Canoco project file (.CON)*, for our example). Note that the command corresponding to this button is located in the **File** submenu and if you open that submenu, you can also see that you can invoke this command using the keyboard shortcut **Ctrl-N** (you must hold the key **Ctrl** down and press the key **N** at the same time). As you can see, some commands are not available at the moment (because no Canoco project is open yet) and their menu choices are grayed-out as well as the corresponding toolbar buttons. The white space in the Canoco workspace is not for typing; it will later on contain windows with information on your ordination analyses.

Now you will create your first Canoco for Windows project by selecting the **File/New project** command (or pressing the **Ctrl-N** key combination or by pressing **Alt-F** followed by **N**, or by clicking the leftmost toolbar button). A new project is created and represented in the Canoco for Windows workspace by two related windows (to be described in more detail later). Because the project must be first defined before it can be used, the Project Setup Wizard appears immediately (Figure 2-6).

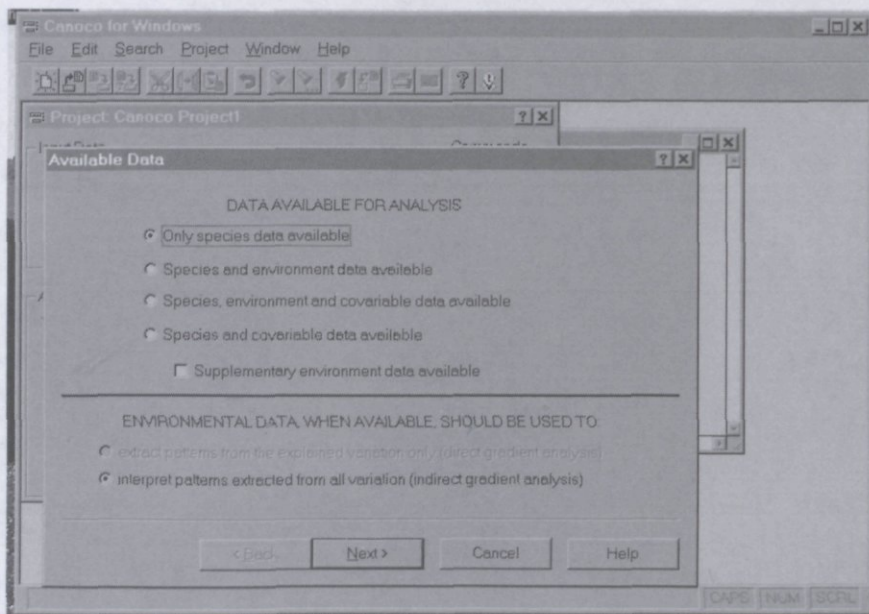


Figure 2-6 The first page of the Project Setup Wizard.

The Project Setup Wizard is used to specify all the options and features of the analyses available under Canoco for Windows. The Project Setup Wizard is quite similar to other “wizards” you might have seen in the Windows operating system when installing new hardware, creating a new graph in your spreadsheet program, etc. The Wizard presents you with a series of pages on various aspects of the ordination method you want to apply to your data. You can progress through the sequence of pages from the beginning to the end using the **Next** button and then finalize the setup (accepting the choices made so far) by clicking the **Finish** button at the last page. Which page appears at a particular moment depends partly on your choices in the preceding wizard pages. This is quite similar to the way the console version of Canoco 4.5 works (and the way the old Canoco versions worked), but two important, time-saving improvements are present:

- you can move backwards to earlier pages to correct mistakes
- you can easily “clone” a particular analysis and quickly change just a few settings in the new copy

Returning to our tutorial example, we will start with a simple ordination method, the classical **Detrended Correspondence Analysis (DCA)**, using the dune meadow data available in the **Samples** subdirectory of the Canoco for Windows directory (i.e. in the **c:\canoco\samples** directory). In the first Project Setup Wizard page, we will keep the choice “**Only species data available**” and click the **Next** button.

As we progress to the next page (**Data Files**), we see two fields that need to be filled (**Species data file name** and **Canoco solution file name**). To fill the first one, we click the **Browse** button on its right side. This displays the standard file selection dialog box, with the **c:\canoco** directory used as the home directory. From there, we navigate to the **Samples** subdirectory (by double-clicking the appropriate folder icon) and there select the **dunespe.dta** file (by selecting it and clicking the **Open** button or by double-clicking it directly). After that, we automatically return to the **Data Files** wizard page, with the path to the species data file already filled-in. The Canoco solution file will contain the analysis results and is also used by the program CanoDraw to display the ordination diagrams. You might like to separate the

project files (to be defined later) and the solution files from the source data and the Canoco for Windows package files by placing them into a separate directory.

To achieve this, click the **Browse** button next to the **Canoco solution file name** field. In the opened file-selection dialog box navigate into the **c:\canoco** directory (if you are not already there), click the right mouse button to invoke the pop-up menu (to perform the correct action, the mouse pointer must not be positioned over any existing file or directory name or icon, when the right mouse button is pressed). In the menu, we select the **New / Folder** command (Figure 2-7) and rename the newly created directory to **Analyses**. Then navigate into this folder and in the **File name** field of the file-selection dialog box type the file name **dune-dca.sol**, and then click the **Open** button. You are automatically returned to the **Data Files** wizard page and you can progress from there using the **Next** button. Canoco for Windows reads the species data file and shows the next wizard page.

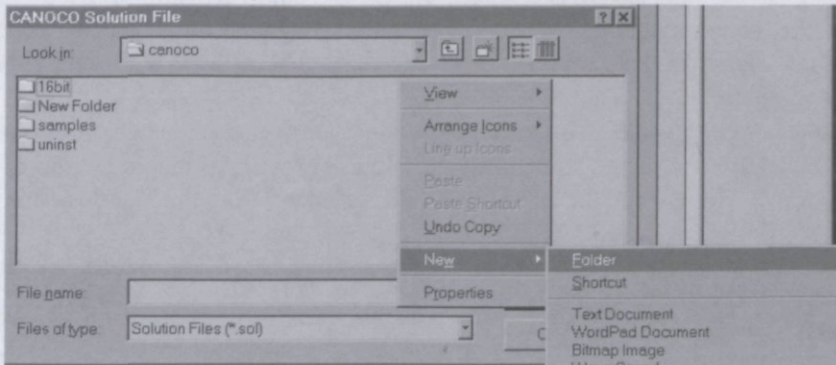


Figure 2-7 Creating new directory in the File Open dialog box.

This wizard page (**Type of Analysis**) is essential for selecting the type of the ordination method to be applied to your data. Given our choices in the previous pages, only indirect gradient analysis methods are available. The **DCA** method is already selected as the default choice and we will keep this selection. But we can stop for a moment on this page to see how to use the Canoco for Windows online documentation. Canoco for Windows features a context-sensitive help available at two levels of detail:

- If you click on the question mark in the upper right corner of the wizard page, the mouse cursor shape changes. Now its image includes not only the arrow, but also the question mark. When in this help-mode, clicking a particular item in the wizard page shows its meaning and its expected way of use. You can move the pointer, for example, to the label **Unimodal** at the left side of the wizard page and after clicking on it, you will see a pop-up window with a short description of methods based on the unimodal response model. The pop-up window disappears as soon as you click somewhere outside of it.
- If you click the **Help** button at the bottom right corner of the wizard page the Windows Help application starts. The displayed help topic refers to the particular wizard page and features a hyper-graphic reproduction of that page. You can get the item-specific information by pointing to that item and clicking it with the left mouse button. You can also browse through the help pages describing the individual Project Setup Wizard pages, using the << and >> buttons in the Help window. The individual help topics can be also printed.

Close the Help window and progress to the next Project Setup Wizard page using the **Next** button. Because we selected a unimodal method with detrending, the next page offers us the choice of the detrending method. We will keep the default choice (**by segments**) and progress to the next page. There we are asked about the transformation to be applied to the species data. As the species values are already on a log-like scale, no further transformation is needed.

The next wizard page allows you to indicate that you want to delete or weight species or samples or to make them supplementary (i.e., passive, in older terminology). All the available choices are deselected, implying that you do not want to make any changes to our data. The choices relating to environmental variables, covariables, and supplementary environmental variables are disabled, as these entities are not present in your project. If you check the delete box for Species and click the **Next** button, a list appears containing the species that occur in the data file. As you do not want to delete any species, click the **Next** button again. The **Finish** wizard page appears now, because you have specified all available options. Here you are reminded that you can move back to change any options. Otherwise you can confirm your choices by clicking the **Finish** button. Do that now.

Because you just defined a new project with no name assigned to it, the file-selection dialog box appears and you must specify a name for the Canoco project file. The file name is automatically given the extension **CON**, if you do not add the extension yourself. The dialog box opens in the **c:\canoco** directory and you must navigate from there to the **Analyses** subdirectory made earlier in this session. Then write **dune-dca** in the **File name** field and click the **Save** button.

You are now back in the Canoco for Windows workspace and you can study in more detail the two views available for each Canoco project. You may wish to maximize the Canoco for Windows workspace window and then select the **Tile** command from the **Window** submenu.

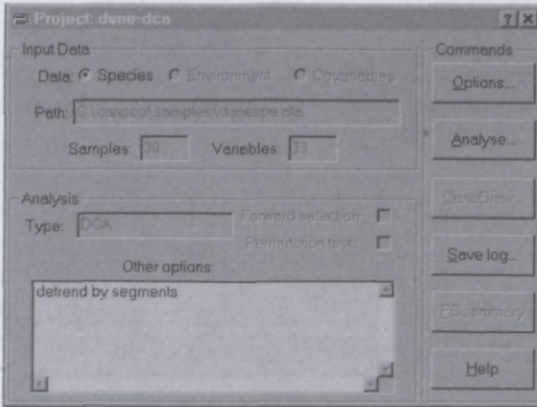


Figure 2-8 The Project View window.

The **Project View** has the title **Project: dune-dca** (Figure 2-8) and summarizes the settings used in the project, including the number of samples and variables (species) in the data files. This view also has a set of buttons on the right side enabling you to execute the most important commands on the corresponding project. The button **Options**, for example, invokes the Project Setup Wizard for modifying the project, the **Analyse** button runs the ordination method on the data sets using the current settings.

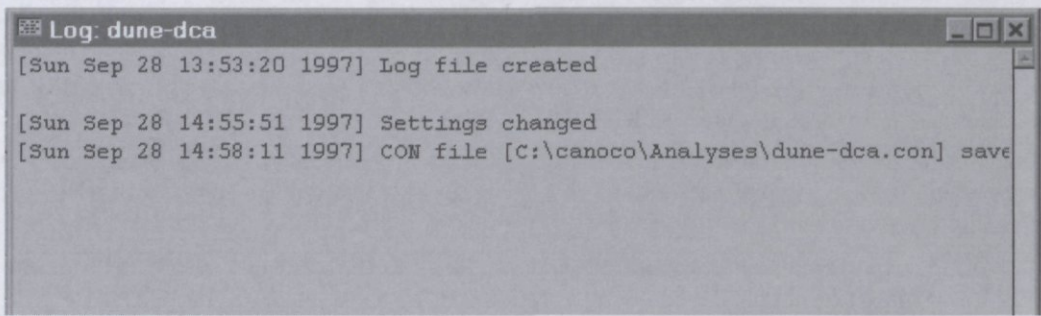




Figure 2-9 The Log View window.

The **Log view** has the title **Log: dune-dca** (Figure 2-9) and records the changes in the project during the particular session (the lines starting with date and time in square brackets), as well as the output from the project analysis by the Canoco program (the other lines). The log is editable, so you can remove text from it or add your own comments. You can also copy the text from the log to the Windows Clipboard and paste it into your preferred word processor, you can save the contents of the log window to a text file, or you can print it.

You can switch between the two related views for a particular project by any of the following methods:

- use the following toolbar button: 
- use the **F3** keyboard shortcut defined for this purpose
- select the appropriate view from the list of open windows at the bottom of the **Window** submenu
- or, if visible, click on the view you wish to activate.

To carry out the actual ordination analysis of the dune meadow data, make sure that the Project View is active and not the Log View, and then click the **Analyze...** button on the right side of the Project View. Alternatively,

- click the  button on the toolbar, or
- select from the **Project** submenu the **Analyze** command, or
- use the **Ctrl-A** keyboard shortcut

The progress box appears, informing you about the progress of the analysis. After the analysis is completed, switch to the Log View by clicking on the log-window or by clicking the switch button on toolbar and maximize the window. The output of the analysis is stored in this window and concludes with the message “[DATE TIME] CANOCO call succeeded”. The summary of the analysis (eigenvalues, lengths of gradient etc.) is shown. You can scroll to the beginning of the log. The log corresponds to the output file of the console version of Canoco 4.5 and is described in section 6.2.

Here the first tutorial session ends and you can close the Canoco for Windows application by the **Ctrl-X** keyboard shortcut. Canoco first asks whether to save the changed log as a file. You can ignore this log and click the **No** button (clicking **Cancel** would cancel the Close command).

2.3 Importing data from a spreadsheet

The Canoco for Windows 4.5 package comes with a simple, but versatile utility for importing data files from your spreadsheet documents. The utility comes in two forms:

- the console program **canoimp.exe** which parses a TAB-separated file and converts it to the requested Canoco-compatible format. This program has a larger capacity, but is less user friendly than the other form
- the Windows - based utility **wcanoimp.exe** which reads the data table from the Windows Clipboard and saves it in a Canoco-compatible format with the requested properties.

In this tutorial, only the second (Windows - based) utility is demonstrated; a more detailed description of both programs appears in Chapter 4: **Data input**. The **samples** subdirectory in the Canoco for Windows installation directory (i.e. the directory **c:\canoco\samples** for our examples) contains a file named **dune_env.xls**, in the format of Microsoft Excel version 3.0.

To transform this spreadsheet into Canoco data file, open it in your favourite spreadsheet program (Excel or other program) and then select the whole data table, including names of the variables (in the first row) and names of the samples (in the first column), as shown in Figure 2-10. Then copy the selected block of data to the Windows Clipboard. To do so, select the **Copy** command, typically placed in the **Edit** submenu of your spreadsheet program. The keyboard shortcut **Ctrl-Ins** works for most programs, as well. After you have done that, start the **WCanoImp** program. You can select the **WCanoImp** item in the **Start / Programs / Canoco for Windows** menu.

	A1	Moisture	Manure	Hayfield	Haypastu	Pasture	SF	BF	HF	NM
1	Sample 1	2.8	1	4	0	1	0	1	0	0
2	Sample 2	3.5	1	2	0	1	0	0	1	0
3	Sample 3	4.3	2	4	0	1	0	1	0	0
4	Sample 4	4.2	2	4	0	1	0	1	0	0
5	Sample 5	6.3	1	2	1	0	0	0	0	1
6	Sample 6	4.3	1	2	0	1	0	0	0	1
7	Sample 7	2.8	1	3	0	0	1	0	0	1
8	Sample 8	4.2	5	3	0	0	1	0	0	1
9	Sample 9	3.7	4	1	1	0	0	0	0	1
10	Sample 10	3.3	2	1	1	0	0	0	1	0
11	Sample 11	3.5	1	1	0	0	1	0	1	0
12	Sample 12	5.8	4	2	0	1	0	1	0	0
13	Sample 13	6	5	3	0	1	0	1	0	0
14	Sample 14	9.3	5	0	0	0	1	0	0	1
15	Sample 15	11.5	5	0	0	1	0	0	0	1
16	Sample 16	5.7	5	3	0	0	1	1	0	0
17	Sample 17	4	2	0	1	0	0	0	0	1
18	Sample 18	4.6	1	0	1	0	0	0	0	1
19	Sample 19	3.7	5	0	1	0	0	0	0	1
20	Sample 20	3.5	5	0	1	0	0	0	0	1
21	Sample 20	3.5	5	0	1	0	0	0	0	1
22										

Figure 2-10 Selecting the data table in the spreadsheet application.

The **WCanoImp** window appears, showing you (in its top half) the instructions for its usage. Note that you have already performed almost all of the required steps. Before you click the **Save** button, you must review the option settings in the bottom half of the **WCanoImp** window. The first option (**Each column is a Sample**) should be checked only if the original data table needs to be transposed to conform to the Canoco data formats, where samples are arranged row-wise. The next two options for generating labels are not appropriate for our situation, as we have both variable and sample labels available in the copied block. The last option (**Save in Condensed format**) is often used with tables containing species data, which are usually very sparse, but this choice brings no advantage for our data (explanatory variables).

After pressing the **Save** button, the file-selection dialog box appears and we must specify a name for the file to be created as well as its directory. Navigate to the `c:\canoco\Analyses` directory and write `dune_env.dta` as the name of the file. The next dialog box allows you to specify the title line stored in the Canoco data file (see the description of file formats in Chapter 4). You can write there anything you like, restricting yourself to a maximum of 80 characters.

After selecting the **OK** button, WCanolmp produces the file and tells us about it with the **Created requested data file** message box. Close the WCanolmp program using the **Exit** button.

2.4 The second run: CCA

In this tutorial, we analyze the relationship between the species and the environmental variables in the Dune Meadow data by **Canonical Correspondence Analysis (CCA)** and determine the statistical significance of this relation by a Monte Carlo permutation test. We do this by modifying the Canoco project we created in section 2.2.

Begin by starting the Canoco for Windows program, using the method outlined in section 2.2. If the **Tip of the Day** window appears, you can close it using the **Close** button. If you open the **File** submenu of the Canoco for Windows menu, you can see a list of the most recently used Canoco projects at its bottom (just above the **Exit** command). If you have worked through the tutorial from section 2.2 recently, the item `C:\canoco\analyses\dune-dca.con` will be present there. By selecting that item, the corresponding project file is opened. We want to keep the previous DCA project and therefore create a “clone” of the original Canoco project. This can be done with the **File / Save As...** menu command or with the toolbar button with the red question mark. Canoco displays the file-selection dialog box, offering the original file name as the default value in the **File name** field. Change the name from `dune-dca` to `dune-cca`. After closing the dialog box, the current project settings are saved into the new project file `dune-cca`. Then Canoco asks us whether to clear the log window. Select the **Yes** button here. We now modify the project settings by invoking the Project Setup Wizard (e.g. by clicking the **Options...** button in the Project View).

Because we want to include environmental data in the analysis, select the **Species and environmental data available** option in the first wizard page. The selection at the bottom is automatically adjusted to the **extract patterns from the explained variation only** option. This is appropriate for obtaining a direct gradient analysis, such as CCA. Click the **Next** button at the bottom of the wizard page.

Because of the changes in the previous page, you need to supply the name of the file with environmental variables in the second edit field in the **Data Files** wizard page. Click the **Browse** button on the right side and select the `duneenv.dta` file in the `c:\canoco\Samples` directory (not the file you created in section 2.3!). Also change the name of the solution file from `dune-dca.sol` to `dune-cca.sol`, at the bottom of this wizard page, but you may also wish to change the name later when you are warned against overwriting by Canoco for Windows.

Clicking the **Next** button brings you to the **Type of Analysis** wizard page, where (due to the change in the first wizard page) the CCA method is already selected. From there you progress to the wizard page titled **Scaling: Unimodal Methods**. Here you can select the method for the scaling of the ordination scores. Given this project was based on the DCA analysis, the value for the **Scaling type** may not be appropriate. Select **biplot scaling** instead of **Hill's scaling** and press the **Next** button.

Change nothing on the **Transformation of Species Data** wizard page. In the **Data Editing Choices** wizard page check the **Delete** option for environmental variables to be able to see which environmental variables are included in the analysis. In the next wizard page, all ten environmental variables of the Dune Meadow data are listed. As we do not want to delete any,

click the **Next** button again. The next wizard page is about **Forward Selection of Environmental Variables**: We will not use this method in our tutorial (the default choice), so click the **Next** button again. In the following wizard page (**Global Permutation Test**), change the choice from **Do not perform the test** to **Both above tests**. This specifies that Canoco will perform two Monte Carlo permutation tests, one based on the first canonical ordination axis and one based on all canonical axes (i.e. on all variability explained by the environmental variables). You are asked to specify the type of the permutation test in the next wizard page. Keep the choice of **Unrestricted permutations**. After clicking the **Next** button, the **Finish Options** page is displayed. Click the **Finish** button there.

Now run the modified analysis project, using (for example) the **Analyze...** button in the **Project View**. The progress dialog indicates how Canoco proceeds with the analysis, including the Monte Carlo permutations. After completion of the analysis, switch to the log-window to see the result of the permutation test. Both reported P-values are less than 0.05 so that we conclude that the relation between the species and the ten environmental variables is statistically significant at the 5% significance level. Higher up in the window you find the summary of the ordination analysis and the means and correlations of the environmental variables.

As the last task in this tutorial, we take a quick tour of the CanoDraw program. The use of this program is described in more detail in the chapters 10 to 14 of this manual. Click the **CanoDraw...** button in the **Project View** to start the program. Canoco for Windows asks CanoDraw to automatically open the active Canoco project file, and CanoDraw asks you about the name, under which the new CanoDraw project should be saved (using the **Save As** dialog box). This is typically the same name as it was used for the Canoco project, except that the file extension is changed from *.con* to *.cdw*. Then select the **Project / Nominal variables / Environmental variables** command from the program submenu. Select all the variables, except the first three ones, in the left-hand list and click the **Select >>** button to move them into the right-hand list. Close this dialog using the **OK** button.

Then in the CanoDraw program select, in the **Create** submenu, the **Biplots and Joint Plots** and, from this submenu, the **Species and env. variables** option. This creates a biplot with species and environmental variables (with nominal environmental variables represented by symbols). The graph can be further adjusted by repositioning the labels, changing symbol types or text of the labels, but we will only store the graph for an eventual later exploration or modification. Select the **File / Save** command and optionally change the graph name in the *File name:* field of the **Save As** dialog box.

You will end our short trip to CanoDraw program here, by selecting the **File / Exit** menu command from the CanoDraw menu (selecting **Yes** when asked about saving the changes in the *dune-cca.cdw* project).

2.5 The third run: RDA

In the next tutorial, you will analyze the dune meadow data using a constrained linear ordination method (**redundancy analysis**, RDA). The beta diversity of the species data is not as high as to make the application of a linear method to this dataset nonsensical. You will also see in this tutorial how the Canoco for Windows program enables us to rank the importance of the individual explanatory variables, using **Automatic Forward selection** of environmental variables.

You will start from the existing project file, namely the **dune-cca.con** project. If you closed the Canoco for Windows application after the last tutorial, you must reopen it now and open the **dune-cca** project as well. You can take a shortcut for both steps: assuming the Canoco for Windows program is not running, go into the **Start** menu, open the **Documents** submenu and

there you will find the **dune-cca** label (preceded by a small icon which Canoco for Windows uses to identify the Project Views). Alternatively, you can use the Windows Explorer application to navigate to the **c:\canoco\analyses** directory and double click the **dune-cca** icon there. Yet another method to open an existing Canoco project file is to drag the file icon from the Windows Explorer area and drop it onto Canoco for Windows workspace.

In the Canoco for Windows workspace, select the Project View and save the project under a new name (using the **File/Save As...** menu command): **dune-rda.con**. Confirm the clearing of the log window, when Canoco asks about it. Now you must change the actual project settings. Click on the **Options...** button in the Project View to open the Project Setup Wizard.

Do not change anything in its first page (**Available Data**) and progress to the second one (using the **Next** button). In the second page, only the name of the solution file has to be changed (to something like **c:\canoco\analyses\dune-rda.sol**). On the following page (**Type of Analysis**), change the selection from **CCA** to **RDA**. Keep the choices on the next two pages (**Scaling: Linear Methods; Transformation of Species Data**). Then the wizard page with title **Centering and Standardization** appears and there you change the option on its right side from **Center by species** to **Center and standardize**, to get an analysis based on the matrix of correlation coefficients (rather than on the matrix of covariances). In the next wizard page (**Data Editing Choices**), you will delete one sample, make one species supplementary (passive in the terminology of the CANOCO 3.x), and define a single interaction term between two explanatory variables, all just for demonstration purposes. You need to tell to Canoco for Windows now that you want to do those operations by checking the corresponding boxes (see Figure 2-11).

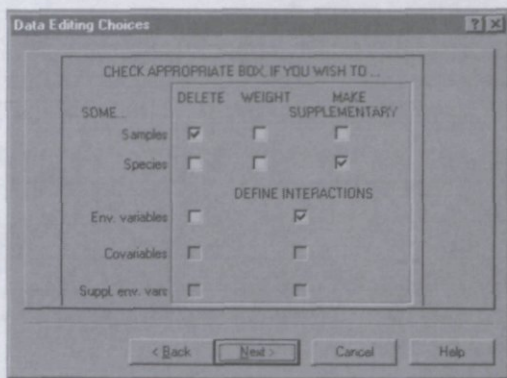


Figure 2-11 Data Editing Choices in the Project Setup Wizard.

If you then press the **Next** button, a wizard page titled **Delete Samples** appears. On its left side is a list of names of samples which occur in the species data file (**dunespe.dta**). This list is labeled **Source pool**. You can move one or many samples from this pool into the pool of samples to be omitted from the analysis by selecting them and then clicking the **>>** button. This command moves the selected sample labels to the list on the right side of the wizard page. Note that we can select a single sample by clicking on it using the left mouse button or we can select a group of contiguous sample labels by selecting the first item in the group and then selecting the last one while holding the **Shift** key down. A non-contiguous set of labels can be selected by combining the clicking on the labels (using the left mouse button) whilst holding the **Ctrl** key down. Also note that because the **dunespe.dta** data set does not have the samples numbered contiguously, empty lines appear in the list (namely between the samples named **Sample17** and **SupplSAM** and the samples **Duplic17** and **Sample18**). As an exercise, you will delete the last sample on the list, with the label **Sample20** (you need to progress to the bottom of the list using

the scroll-bar on its right side). Click on the **Sample20** label to select it and then click the >> button. The label moves to the right-hand listbox. If you change our mind and want to move the sample back (i.e. not to delete it in our analysis), you have to select it in the right listbox and click on the << button.

In the next wizard page (**Supplementary Species**) make the first species (**Ach mil**) supplementary using a similar technique. Species (or samples, if selected in similar way) which are supplementary do not actively influence the solution produced by the ordination method, but they are passively projected into the resulting ordination space, based on their occurrences in the data.

The next wizard page (**Interactions of Env. variables**) allows us to specify an interaction between two environmental (explanatory) variables. The interactions are defined in the Canoco program as simple products of the interacting terms. As you can see, the first environmental variable (**A1** - thickness of the upper soil horizon) is preselected in the top left list and the second environmental variable (**Moisture**) in the top right list. Change the selection, nevertheless, to the variable **Moisture** in the **First variable** list and to the variable **Manure** in the **Second variable** list. Then click the **Add** button. The interaction term appears in the bottom list (titled **Powers and product variables**) - see Figure 2-12. The newly created term appears in both source listboxes, as well, so that more complicated interactions can be defined.

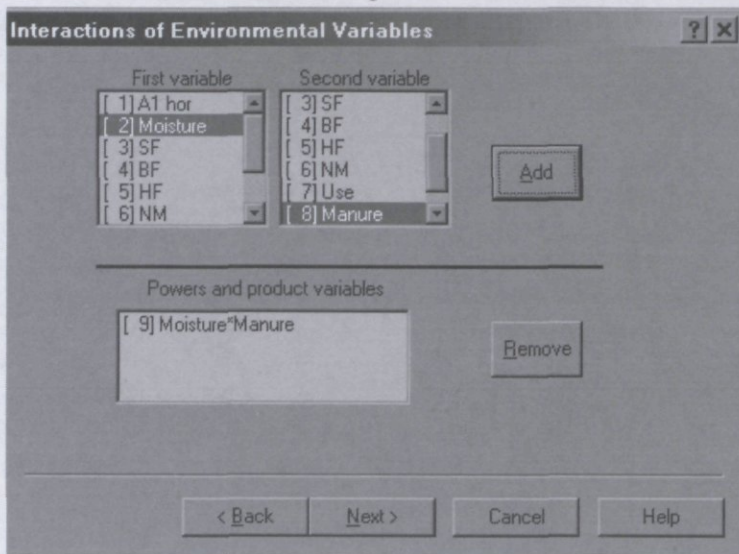


Figure 2-12 Interactions of Environmental variables in the Project Setup Wizard.

Clicking the **Next** button brings us to the page titled **Forward Selection of Env. variables**. Select here the **Automatic selection** option instead of the default one (which says **Do not use forward selection**). Canoco for Windows informs us that we can select at most eleven environmental variables (including the just defined interaction term). Also, the tests of the significance of the effect of the environmental variables in the individual steps of the forward selection is preselected (in the **use Monte Carlo Permutation Tests** option). Each permutation test will be based on 499 random permutations. The next wizard page allows us to specify possible restrictions on the permutation, based on the specific properties of our sampling design. As our dune data set does not have any such specific properties, we keep the **Unrestricted permutations** option, allowing for completely random permutation of the samples. Clicking the **Next** button brings us to the **Finish** option page.

Now save the changed project settings to the project file (using the keyboard short-cut **Ctrl-V**, for example) and run the analysis (by pressing the **Ctrl-A** key combination). You must exercise some patience, as Canoco for Windows does a permutation test on each of the eleven explanatory variables, during the forward selection. You can inspect the log file for a detailed description of the stepwise selection done by Canoco, but Canoco for Windows also provides a short summary of the automatic forward selection. To see it, press the **FS summary** button in the Project View. A dialog box similar to that in Figure 2-13 appears.

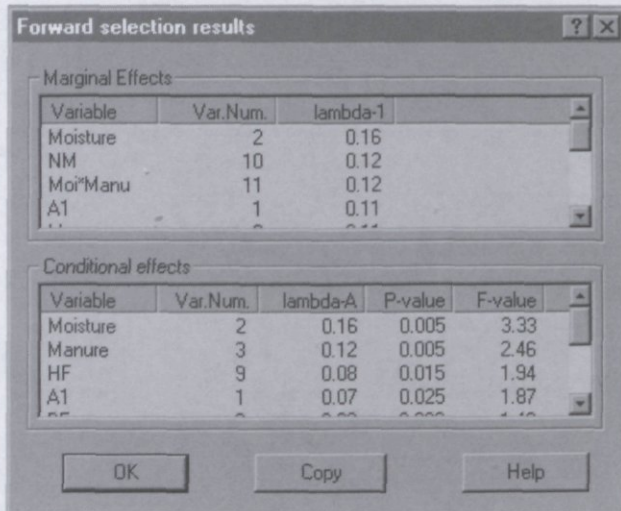


Figure 2-13 Forward Selection summary.

In the table in the upper part of the dialog box, the marginal effects of the explanatory variables are displayed, ordered from the variable with the highest explanatory power (at the top of the list) to the variable with the minimum ability to explain patterns in the species data (at the bottom of the list). In the summary presented in the upper part of this dialog box, each variable is judged separately, without considering the effect of the other explanatory variables.

On the other hand, the sequence of the variables in the bottom part of the dialog box is obtained by the stepwise selection procedure. In this procedure, the explanatory variable best fitting the species data is selected first and then, the next best fitting variable is added. Before each addition, the significance of the explanatory effect of the candidate variable is evaluated using the Monte Carlo permutation test. The table in the bottom part of the dialog box records not only the amount of the variability in species data, explained by the particular environmental variable when included into the set of selected explanatory variables, but also the results of the corresponding Monte Carlo permutation test (the column showing the value of the F statistics and the estimate of the probability of the Type I error). The last two columns are displayed only if Monte Carlo permutation testing during forward selection was selected in the analysis setup.

Both tables can be copied to the Windows Clipboard by clicking the **Copy** button at the bottom. From there, they can be pasted into a document or a spreadsheet. The values are TAB-separated. To paste the tables into a spreadsheet, start your spreadsheet application (or switch to it, if it is already running), click on an empty sheet and select **Paste** from the **Edit** menu.

Note that the Automatic Forward selection procedure is not available in the console version of CANOCO 4.5.

3. Background theory

3.1 General objective

CANOCO, an acronym for **CAN**Onical **C**ommunity **O**rdination, is designed for data analysis in community ecology. Researchers in other disciplines should consult Table 3.1 for the terminology used in this manual. Canonical ordination is a class of techniques for relating the species composition of communities to their environment. Data analysis by canonical ordination can either be exploratory or confirmatory. When used in an exploratory way, it leads to an ordination diagram of samples, species, and environmental variables, which optimally displays how community composition varies with the environment. When used in a confirmatory way, it leads to statistical tests of the effects of particular environmental variables on community composition taking into account the effect of other variables. The theory of this is given in the book by Jongman et al. (1987) and the collection of papers in the book “Unimodal models to relate species to environment”, in particular Ter Braak and Prentice (1987).

3.2 Terminology

The terminology (Table 3.1) used in CANOCO stems from typical applications in community ecology. CANOCO operates on species, environmental variables, and covariables (Table 3.1). Ordination is applied to the species data, which are typically data on abundances or incidences (i.e. presence-absence) of a set of species in a set of samples. The variation in the species data is to be explained via the ordination axes by environmental variables and covariables. Environmental variables are the explanatory variables of prime interest. Covariables are concomitant variables whose effect must be partialled out when estimating the effects of the environmental variables. When one wants a constrained ordination, the number of environmental variables and covariables must be smaller than the number of samples; otherwise constrained ordination and unconstrained ordination coincide. There is no such limit on the number of species; the lower limit is 1 in the case of PCA/RDA and 2 in the case of CA/DCA/CCA.

Of course, there is nothing special about the terms used. They have a formal meaning only (Table 3.1). For example, if one wants an ordination of environmental variables, then this is easily done by entering the name of the data file containing these variables at the point where one usually specifies a file with species data.

Table 3.1 Terminology used in CANOCO, with commonly used synonyms.

Term	Explanation
Abundance/response	value of a response variable, usually positive or 0
Biplot	an ordination diagram of two kinds of entities which can be interpreted by the biplot rule. Interpretation proceeds by projecting points on directions defined by arrows in the biplot (e.g. Fig 3. on page 77 of Unimodal Models). See section 3.5
Canonical axis	an ordination axis that is constrained to be a linear combination of environmental variables
Canonical eigenvalue	eigenvalue of a canonical axis
Canonical ordination	an ordination in which the axes are constrained to be linear combinations of environmental variables
Community	a set of individuals pertaining to several species occurring together in a given area at a given time; assemblage
Covariable	concomitant variable, background variable, explanatory variables corresponding to incidental or nuisance parameters, block factor in experimental design
Direct gradient analysis	external analysis, canonical ordination, ordination constrained by external variables, constrained multivariate regression, reduced-rank regression, direct comparison
Eigenvalue	importance measure of an ordination axis
Environmental variable	explanatory variable (of prime interest), independent variable in a regression equation, external variable, stimulus variable, treatment variable
Gradient	latent environmental variable, see ordination axis
Indirect gradient analysis	internal analysis, "factor analysis", unconstrained ordination, indirect comparison, metric scaling or multidimensional scaling, possibly followed post-hoc by a regression analysis on external variables
Joint plot	an ordination diagram of two kinds of entities which can be interpreted by the centroid principle. See section 3.5
Linear method	method based on a linear model, e.g. linear regression, multiple regression, principal

Term	Explanation
Ordination	components analysis, redundancy analysis see Indirect gradient analysis
Ordination axis	eigenvector, latent variable, theoretical explanatory variable
Ordination diagram	scatter plot of the eigenvector scores; used both for biplots and joint plots
Sample	sampling unit, individual, object, site
Sample score	position of a sample along an ordination axis; eigenvector value of a sample
Species	response variable, dependent variable in a regression equation, internal variable
Species score	value of a species on an ordination axis; eigenvector coefficient; loading in PCA, center of species curve in CA and DCA
Supplementary sample	sample added post-hoc to the ordination by projection. Called passive sample in CANOCO 3.0
Supplementary variable	variable (species or environmental variable) added post-hoc to the ordination by projection. Called passive variable in CANOCO 3.0
Triplot	an ordination diagram with three kinds of entities of which all pairs form biplots. Examples are the RDA and CCA triplots that consist of samples, species and environmental variables (often also called biplots)
Weighted averaging method	method based on a unimodal response model of which the optimum (mode, ideal point) is estimated by weighted averaging, e.g. correspondence analysis

3.3 Models, methods, and algorithms

In this section we outline the methods available in CANOCO, the models on which they are based, and algorithms which are used. For more information consult Chapter 5 in Jongman et al. (1987) and Unimodal Models.

Canonical ordination is a combination of ordination and multiple regression. Ordination techniques such as principal components and correspondence analysis (= reciprocal averaging) are commonly used to reduce the variation in community composition to the scatter of samples and species in an ordination diagram. Subsequently the diagram is interpreted with the help of external data, for example by calculating correlation coefficients between environmental variables and ordination axes, or by multiple regression of the ordination axes on environmental variables.

A difficulty here is that the ordination axes are just particular orthogonal directions in the ordination diagram. Other directions may well be better related to the environmental variables. Canonical ordination is a solution to this difficulty. The regression model is inserted in the ordination model. As a result the ordination axes appear in order of variance explained by linear combinations of environmental variables.

The ordination technique of correspondence analysis was introduced in ecology by way of the reciprocal averaging algorithm (Hill, 1973a), also called the two-way weighted averaging algorithm. It is an iterative ordination algorithm: from initial arbitrary sample scores, species scores are obtained, from which new samples scores are derived, from which new species scores are derived, and so on.

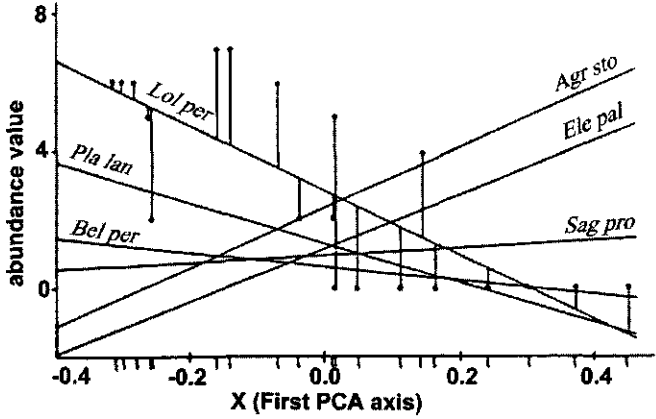


Figure 3-1 Linear response model in PCA and RDA.
 Straight lines for the abundance of six plant species along the first axis of principal components analysis (X), applied to the dune meadow data.

Principal components analysis can be obtained by a similar algorithm by taking weighted sums, instead of weighted averages (Jongman et al. 1987: section 5.3). Canonical ordination techniques can be obtained by carrying out multiple regressions within the iterative algorithm: each time new sample scores are derived, they are regressed on the environmental variables (instead of just once after an ordination). CANOCO uses this kind of iterative ordination algorithm. Details of the algorithm are given in the appendix of Ter Braak & Prentice (1988, pages 93 - 138 in *Unimodal Models*). The resulting species scores are parameters of response curves of species with respect to the ordination axis. In linear methods to which principal components analysis belongs, the response "curves" are straight lines (Figure 3-1) and the species scores are slope parameters. In weighted averaging methods to which correspondence analysis belongs, the response curves are unimodal (Figure 3-2) and the species scores can be considered as the centers of the curves. For symmetric response curves that are sampled over their full range, the center is equal to the optimum of the curve.

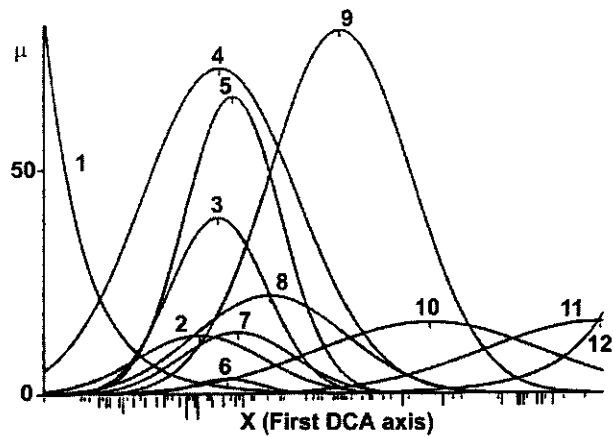


Figure 3-2 Unimodal response model in (D)CA and CCA.

Response curves for the count (μ) of 12 species of wolf spiders in a dune area, along the first axis of detrended correspondence analysis (X) applied to data of Van der Aart & Smeenk-Enserink (1975).

In their theory of gradient analysis, Ter Braak & Prentice (1988) introduce six types of data-analysis problems on the basis of the linear and unimodal response models in Figure 3-1 and Figure 3-2:

1. **Regression:** When there is just a single, known explanatory variable, the slope of each line in Figure 3-1 would have been estimated by simple linear regression and the center of each curve in Figure 3-2 by weighted averaging "regression".
2. **Calibration:** If there are some samples for which the value of the explanatory variable is missing, the values can be estimated from the species composition of those samples by seeking for each such sample the value of the environmental variable that is most likely to give the observed species composition as judged by the response curves in the Figures. This gives linear calibration in Figure 3-1 and weighted averaging calibration in Figure 3-2.
3. **Ordination:** When all values of the explanatory variable are missing, one could still attempt to construct a theoretical variable that best fits the species data according to a linear model or a unimodal model. The theoretical variable is the first ordination axis found by the iterative ordination algorithm. The algorithm is essentially a converging sequence of regressions and calibrations. The sample scores are the values that the theoretical variable takes in the samples. The theoretical variable/ordination axis has no environmental basis. In CANOCO, the linear method of ordination is principal component analysis (PCA), whereas the unimodal methods of ordination are correspondence analysis (CA) and detrended correspondence analysis (DCA).
4. **Canonical ordination:** With the additional constraint that the ordination axis must be a linear combination of environmental variables we obtain canonical ordination. Canonical ordination is thus a particular form of constrained ordination. It has an environmental basis. In CANOCO, the linear method of canonical ordination is redundancy analysis (RDA), whereas the unimodal methods of ordination are canonical correspondence analysis (CCA) and detrended canonical correspondence analysis (DCCA).
5. **Partial ordination:** One can also apply ordination to the variation in the community data that remains after known environmental variables have been fitted by regression. Ordination of the residual variation is called **partial ordination**: the effect of particular variables is

partialled out (eliminated) from the ordination. The variables of which the effects are partialled out are called **covariables**.

6. **Partial canonical ordination**: When the axes of a partial ordination are constrained to be linear combinations of particular environmental variables, we obtain a **partial canonical ordination**.

CANOCO provides the solutions to the data analytical problems numbered 3-6. The solutions are obtained by the iterative ordination algorithm that is detailed in the Appendix of Ter Braak & Prentice (1988). For the unimodal methods this is a weighted averaging algorithm. Under particular conditions the weighted averaging methods are a close approximation to maximum likelihood methods based on unimodal, Gaussian models (Ter Braak 1986). These are more formal statistical methods which require heavy computation and which are therefore less attractive for routine use. One cannot obtain them with CANOCO. But CANOCO is useful to obtain starting values for these maximum likelihood methods.

Although CANOCO can solve some regression and calibration problems, it is not handy to use CANOCO for this. CanoDraw for Windows has some regression facilities, including the (Gaussian) logit regression and Poisson loglinear regression. These can be used to regress individual species to ordination axes, such as in Figs 5.8 and 5.10 in Jongman et al (1987) and Fig. 4 in Ter Braak & Prentice (1988). For calibration based on weighted averaging methods, see Ter Braak & Juggins (1993), the program Calibrate (Juggins & Ter Braak, 1993), or the program WACALIB (Line et al. 1994)

It may come as a surprise that canonical correlation analysis (Gittins, 1985) is missing in the above list of methods, as this is the standard linear multivariate technique for relating two sets of variables (in our case, the set of species and the set of environmental variables). In its place comes the lesser known technique of redundancy analysis, alias least-squares reduced-rank regression. The most important difference between these techniques is that redundancy analysis can analyze any number of species, whereas in canonical correlation analysis the number of species must be less than $n-q$ with n the number of samples and q the number of environmental variables. The latter restriction makes canonical correlation analysis unattractive for most studies in community ecology. More details about the difference are given in Ter Braak & Looman (1994, Unimodal Models pp. 238 - 258). Another major difference is that the two groups of variables play the same role, as in correlation analysis. In RDA and reduced-rank regression on the contrary, distinction is made between response (species) variables and explanatory (environmental) variables, as in regression analysis.

CANOCO is particularly efficient for ordination of sparse data sets (data containing many zero values compared to the number of nonzero values). It is quite common in community data that the average number of species present in a sample is in the order of 10-30, whereas the total number of species in the data set is in the order of 100-1000. By not storing zero values, a large saving of memory space and of computer time is achieved. The iterative ordination algorithm used by CANOCO is specially designed to make storage of zero values unnecessary. It uses methods of calculation that are efficient for sparse data. This design makes CANOCO efficient also for the ordination of nominal response data (section 3.10).

An ordination yields sets of scores for species, samples and, if present, environmental variables along ordination axes. Section 6.3 details how the ordination scores of species, samples and environmental variables are related. The relationships are given per ordination axis, either in the form of a simple linear regression or in the form a weighted average. It is important to note that the same relationships carry through for all ordination axes simultaneously. Multivariate equations are not needed: because the ordination axes (eigenvectors) are orthogonal, multivariate equations can be simplified to equations per ordination axis, as explained above equation (3.3) on page 40. In particular, the sample scores (or, in canonical methods, the sample scores that are linear combinations of environmental variables) of one axis

are orthogonal to those of another; they are also uncorrelated, as can be verified in the correlation matrix in the log-window (section 6.2.4). Also, the species scores of one axis are orthogonal (but not necessarily uncorrelated) to those of another axis, except in DCA with detrending-by-segments.

3.4 The two faces of (canonical) correspondence analysis

In the previous section, correspondence analysis (CA) and canonical correspondence analysis (CCA) were presented as methods for analyzing unimodal data. However, CA and CCA are chameleons: in the one context they show up as unimodal methods, whereas in another they show up as linear methods. These two faces are discussed in Ter Braak & Verdonschot (1995: pp 263-265 and p. 278; pages 153-187 in *Unimodal Models*). They conclude that the common element in all theoretical derivations is that CA and CCA model relative abundance instead of the absolute abundance. See also section 3.9.4.

The Results chapter (in particular, section 6.3) presents both faces of CA and CCA. In the unimodal context, species scores are weighted averages of sample scores, and vice versa (see equations (6.11) and (6.20) on pages 158 and 161), whereas in the linear context, the species scores are derived from a weighted linear regression of transformed species data on to the sample scores, and vice versa (see equations (6.17) and (6.25) on pages 159 and 163). Both types of formulae are given in Table 6.33 and Table 6.36 for CA and CCA, respectively. The ordination diagnostics have been developed in either context (section 6.3.11). The linear context is most useful when the gradients are short (<3 SD), and the unimodal context when the gradients are long (>4 SD). For intermediate gradient lengths either context may be useful. See also section 3.9.4.

3.5 Interpreting ordination diagrams

An ordination diagram with both samples and species can display either the relationships among samples or the relationship among species in an optimal way, but not both. The difference between the two types of diagrams is simple: the ordination axes of one type are a linear rescaling of those of the other. A compromise scaling is also possible (Table 6.2 and Table 6.3). In a diagram that optimally displays inter-sample relationships, the variance of the sample scores on each ordination axis reflects the importance of the axis as measured by the eigenvalue, whereas the variances of the species scores along the axes are equal (or, in the so-called Hill's scaling, about equal). As a consequence, the sample scores on the first axis will show a larger spread than on the second axis. This type of diagram also allows you to interpret distances between centroids of groups of samples (as specified by nominal environmental variables, see section 6.3.10). In contrast, in a diagram that optimally displays the inter-species relationships, the variance of the species scores on each ordination axis reflects the importance of the axis, whereas the variances of the sample scores along the axes are equal (or, in Hill's scaling, about equal). As a consequence, the species scores on the first axis will show a larger spread than on the second axis. See also Table 6.26 and Table 6.34.

How to interpret plots of species, samples, and environmental variables can be understood from the interrelationships between their scores along each ordination axis (section 6.3). There are two important types of interrelations, centroid relations and regression relations, leading to the centroid principle and the biplot rule, respectively.

1. Centroid relation. In a default CA or CCA, a species score is a weighted average of the sample scores. Therefore, the species' point in the CA or CCA ordination diagram is at the

centroid of the sample points where it occurs. The samples that contain the species are thus scattered around that species' point in the diagram. This way of interpreting species-sample diagrams is called the **centroid principle**. Such ordination diagrams are called joint plots.

2. Regression relation. In PCA and RDA, the species score is obtained by regression of the species data on to the samples scores. The species score is thus a slope parameter. The position of the species point, with respect to the origin (0,0) of the diagram, thus tells us the rate of change of the fitted species abundance along each of the axes. By connecting the point (0,0) with the species point we obtain an arrow: the arrow points in the direction in which the species' abundance value increases at the largest rate across the ordination diagram. The rate of change in the orthogonal direction is 0. Such plots are called biplots and are interpreted by the biplot rule. The **biplot rule** works as follows (for an illustration see page 145 of *Unimodal Models* or page 128 in Jongman et al. 1987). By connecting the origin point (0,0) with the species point, an arrow is obtained that points in the direction of increasing fitted values. The arrow can be extended on either side to form a line. By projecting the sample points on the line and ranking the projection points, a ranking can be obtained of the fitted values for that particular species. We can also start with a particular sample, draw a line through the origin and the sample, and project the species points on the line. The projection points give the rank of the fitted values for that particular sample.

Because of the ubiquity of biplots in CANOCO, we give a more formal and general description here. Let the data table have values $\{a_{ik}\}$ [$i = 1, \dots, I; k = 1, \dots, K$]. Suppose the rows have known scores $\{r_{is}\}$ [$i = 1, \dots, I$] on each axis s [$s = 1, 2$]. We now derive for the k th column its scores (c_{k1}, c_{k2}) by a weighted linear regression of the k th column of the data table on to the row scores (r_{i1}, r_{i2}) using weights $\{w_i\}$ for the rows, i.e. we find the scores (c_{k1}, c_{k2}) by minimizing the least-squares criterion

$$(3.1) \quad \sum_i w_i \{a_{ik} - (c_{k1} r_{i1} + c_{k2} r_{i2})\}^2$$

The fitted values of the model are

$$(3.2) \quad \hat{a}_{ik} = c_{k1} r_{i1} + c_{k2} r_{i2}$$

If the row scores are orthogonal (e.g. have zero mean and zero correlation), the optimal score c_{ks} ($s = 1, 2$) is given by the simple formula

$$(3.3) \quad c_{ks} = \sum_i w_i a_{ik} r_{is} / \sum_i w_i r_{is}^2$$

The direction of maximum change makes an angle of θ with the first axis, where $\theta = \arctan(c_{k2} / c_{k1})$. This direction can be indicated in the ordination diagram by an arrow running from the origin (0,0) to the point with coordinates (c_{k1}, c_{k2}) . See pages 134-135 in Jongman et al. (1987). Because (3.2) is symmetric in "c" and "r", the biplot rule can be applied both row-wise and column-wise. The biplot exactly represents the fitted values \hat{a}_{ik} , and approximately represents the original data table $\{a_{ik}\}$. The representation is optimal, given the positions of the row points, as judged by the least-squares criterion (3.1).

In the exposition above we can also interchange rows and columns, resulting in a biplot which is optimal conditionally on the positions of the column points. In many instances in CANOCO, the column points are not only the result of a regression on the row points but, also, vice versa, the row points are the result of a regression on the column points. Such biplots minimize (3.1) unconditionally and are thus optimal unconditionally.

Equations of the form (3.3) can be found on page 158, equation (6.9), and on page 161, equation (6.19), but could also be given in other places. Examples are:

- The environmental biplot scores (Table 6.29 and section 6.3.9) are the result of a weighted regression of a table of correlation coefficients on the species scores in linear methods (weights $\{w_k\}$). The equivalent of (3.3) for this case reduces to (6.32) as shown in section 17.1. The species scores and environmental biplot scores thus together form a biplot that displays the fitted correlation coefficients. In unimodal methods, the table of correlations is replaced by a table of weighted averages of the species with respect to environmental variables (Table 6.38 and section 17.1).
- The environmental centroids (Table 6.30 and section 6.3.10) are the result of a weighted regression of the table of class means on the species scores in linear methods. The equivalent of (3.3) for this case reduces to (6.34) as shown in section 17.2. The species scores and centroids thus together form a biplot that displays the fitted means. In unimodal models, the table of means is replaced by a table of relative class totals (Table 6.39 and section 17.2).
- The regression coefficients of Table 6.31 and section 6.3.6 can be obtained by a weighted regression of the table of regression coefficients $\{d_{jk}\}$ in (6.49) on to the species scores. The equivalent of (3.3) for this case reduces to (6.28) [resulting from the regression of the sample scores on the standardized environmental variables]. The proof of this result is analogous to that for the environmental biplot scores in section 17.1. The species scores and regression coefficients of section 6.3.6 thus together form a biplot that approximates the regression coefficients $\{d_{jk}\}$ (see Table 6.31). This biplot is called a “regression biplot”.

In summary, the biplot rule and the centroid principle are the key to the interpretation of ordination diagrams. In linear methods, the data table $\{a_{ik}\}$ in the above description of the biplot rule can be absolute abundances of species in samples, correlation coefficients between species and quantitative environmental variables, and mean abundances of species in classes of samples, leading to biplots of species and samples, species and environmental variables, and species and classes of nominal environmental variables, respectively. In unimodal methods, the same series of biplots represent relative abundances of species in samples, weighted averages of species with respect to quantitative environmental variables, relative total abundances of species in classes of samples, respectively. The samples and environmental variables together form a biplot of the environmental data in the default diagrams (i.e. with focus on species, scaling 2) but not in other scalings. All these biplots follow from the interrelations among scores as presented in section 6.3. The rows and columns of the tables can be the same, leading to biplots of correlations among species or among environmental variables. These biplots apply in the default scaling 2, but not in other scalings. A full list is given in Chapters 9 and 10 of *Unimodal Models* (in particular, for linear methods, Fig. 1 on page 140 and, Table 1 on page 143 and, for unimodal methods, Figure 1 on page 158 and Table 2 on page 164). The so-called regression biplots are explained in *Unimodal Models* in Chapter 15 and are further discussed in Chapter 5 (page 63), Chapter 9 (page 148 and the dashed arrows in Fig 2 on page 141) and Chapter 14 (page 235). For an ecological application see Baar & Ter Braak (1996).

In the terminology of Gower & Hand (1996), the biplots as described here are all linear predictive biplots. Gower & Hand (1996) also discuss interpolative biplots. The interpretation of the plot of the sample scores and the environmental scores below equation (6.33) on page 169 and below (6.35) on page 172 is related to interpolative biplots (Ter Braak 1997).

As discussed in the previous section, CA and CCA can be placed both in a unimodal context and in a linear context. In particular, the weighted average of species with respect to samples can also be interpreted, in a linear context, as the slope parameter of species data on sample scores. This implies that the species-sample plot of CA and CCA (if in biplot scaling) can not only be interpreted by the centroid principle but also by the biplot rule. The biplot rule is more quantitative and is more attractive when the gradient lengths are short (< 3 SD), whereas

the centroid principle is more qualitative and more attractive if gradient lengths are long (> 4 SD). These points are further discussed on pages 171-275 of *Unimodal Models* and also in section 3.9.

Finally we mention the distance rule to interpret ordination diagrams. The distance rule is an extension of the centroid principle. The distance rule says that a sample that is close to the species point is more likely to contain the species than a sample that is far from the species point. The rank order of abundance values of a species can be inferred from the distances of the samples to the species point. The distance rule can be applied to DCA diagrams with long gradients ($>3-4$ SD).

3.6 Supplementary species, samples, and environmental variables

Supplementary species, samples, and environmental variables differ from active ones in that they do not influence the definition of the ordination axes. Nevertheless, a supplementary item can be added to an existing ordination by projection, i.e. by regressing its data on to the existing ordination axes. Supplementary items are also called passive.

In particular, supplementary species and samples are added afterwards so that their relation to the other samples or species can still be judged from the ordination diagram. In CANOCO, the data on supplementary species and samples must be supplied in the species data file. You can specify which species or samples must be made supplementary in the project for that particular analysis. The scores for supplementary species are calculated from the eigenvector sample scores using equations (6.9) and (6.11). For supplementary samples, equations (6.20) and (6.21) are used.

In contrast to supplementary species and samples, supplementary environmental variables must be specified in a separate file. Supplementary environmental variables are useful to provide an alternative interpretation of the ordination diagram, the other interpretation being given by the environmental variables of the analysis. CANOCO supplies the same type of results for supplementary environmental variables as for normal environmental variables (rows 7 - 14 of Table 6.22 on page 133). The default way of displaying a supplementary environmental variable is by its environmental biplot score or its centroid. See sections 6.3.9 and 6.3.10. In an indirect gradient analysis, there is, at least in theory, no distinction between a variable in the environmental file and a variable in the supplementary environmental file.

Because a species is a response variable and an environmental variable is an explanatory variable, there is a theoretical distinction between a supplementary species and a supplementary environmental variable. See Baar & Ter Braak (1996) for an example. The distinction disappears in scaling 2.

3.7 Permutation tests

3.7.1 Introduction

The statistical significance of the relationship between the species and the whole set of environmental variables, given the covariables, can be evaluated using Monte Carlo permutation tests. A Monte Carlo permutation test is a test of statistical significance obtained by repeatedly shuffling (permuting) the samples. This section summarizes the basic ideas behind Monte Carlo permutation tests, discusses how to obtain valid tests in structured study designs, and discusses

what is actually shuffled in partial tests. The section closes with a subsection on design-based and model-based permutation methods in the analysis of variance with fixed and random factors.

Note that CANOCO cannot determine the significance of ordination axes of indirect analyses (PCA, CA, DCA). The theory that has been developed for such tests is less convincing than the theory presented here for direct gradient analyses.

3.7.2 The basic idea

To understand how a Monte Carlo tests differs from more traditional statistical tests, we first summarize the basics of the statistical hypothesis test. We take as the null hypothesis of the test that the species data are unrelated to the environmental data and as the alternative hypothesis that the species respond to the environment. The basic idea of a statistical test is then as follows.

1. Choose a test statistic that expresses how strongly the species data respond to the environmental data. Familiar examples are the correlation coefficient, the t -ratio, and F -ratio.
2. Calculate the test statistic for the data. We denote the value obtained by F_0 .
3. Determine a reference distribution for the test statistic under the null hypothesis. The reference distribution shows which values can be expected under the null hypothesis that the species are not related to the environmental data.
4. Calculate the significance level (P-value), i.e. the probability that F_0 or larger values occur in the reference distribution.

The crux of all standard statistical tests is that the reference distribution can be derived mathematically from the assumptions of the test. For example, the reference distribution of the F -ratio calculated in a regression analysis or an analysis of variance (ANOVA) is the F -distribution (with particular numbers of degrees of freedom) and holds true if the data are independent and follow a normal distribution with homogeneous variance. The procedure to carry out an F -test thus simplifies to calculating the F -ratio from the data (step 2) and reading off the significance level from a table of the F -distribution in a statistical textbook or from a computer program that can calculate percentage points of the F -distribution (step 4). In contrast, the reference distribution in a permutation test is determined from the data themselves without the assumption of normality and without mathematical derivations. Its basis lies in the observation that under the null hypothesis the samples in the species data can be randomly linked with the samples in the environmental data. In other words, under the null hypothesis, each permutation of the samples in the species data is equally likely. Each permutation leads to a new data set from which we can calculate the test statistic. The reference distribution therefore is the distribution of the test statistic in the permuted data sets. In a Monte Carlo permutation test, we do not generate all possible permutations, but just a random sample thereof. This saves time if there are very many possible permutations (with small sample sizes though, we may wish to enumerate all permutations and supply these to CANOCO, see example BACIIISPE, in section 8.3.7). In summary, the steps in a Monte Carlo test are:

1. Choose a test statistic that expresses how strongly the species data respond to the environmental data. In CANOCO, you can choose from two test statistics that both have the form of an F -ratio.
2. Calculate the test statistic for the data. We denote the obtained value by F_0 .
3. Generate K new data sets that are equally likely under the null hypothesis. In CANOCO, new data sets are generated by randomly permuting the samples in the species data (the response data) while keeping the environmental data (and covariable data, if present) fixed.
4. Calculate the test statistic for each new data set, leading to values F_1, F_2, \dots, F_K .

5. Calculate the Monte Carlo significance level, i.e. place F_0 among F_1, F_2, \dots, F_K and determine the proportion of values greater than or equal to F_0 . Said otherwise, the Monte Carlo significance level is the rank of F_0 among all values $F_0, F_1, F_2, \dots, F_K$ divided by $K+1$. Division is by $K + 1$ instead of K because the value F_0 is included in the null distribution (Hope 1968).

An example is given in Table 6.12 on page 128. Because determined from K random data sets, the significance level resulting from a Monte Carlo test is not a constant for a given data set. If K is chosen large, e.g. $K = 10000$, the remaining random variation will be negligible. However, such large numbers of permutations are not strictly necessary. Even if just 19 permutations are carried out each time the test is carried out, F_0 will be greatest with probability $1/20$ if the null hypothesis holds true. Therefore the Monte Carlo test is exact in the sense that, at the 5% significance level, the null hypothesis is falsely rejected precisely in one of the twenty cases in which it is applied. The reason for using more than 19 permutations is that the power of the test to detect deviations from the null hypothesis increases with the number of permutations. This increase comes at the cost of computer time. Because the law of diminishing returns applies, a good compromise is to carry out at least 199 permutations for a test at the 5% significance level (Unimodal Models: p 198). This is the default number in CANOCO 4.5; but the number of permutations should be increased if the extra time required is bearable.

3.7.3 Permutation type: how are samples shuffled?

The validity of the Monte Carlo test hinges on the generation of new data sets that are equally likely under the null hypothesis (step 3 of the Monte Carlo test in the previous section). If we know that the samples are independent or exchangeable under the null hypothesis, then the new data sets can be obtained by permuting the samples completely at random. However, completely random permutations yield invalid tests if the samples show additional structure in the way they are collected. For example, the data may come from a survey that uses a stratified sampling design. We may also want to account for the fact that the samples form a time series or have a particular spatial layout. In a designed experiment, the samples may have been grouped in blocks. The experiment may have more than one source of error (error stratum), as in a split-plot design, or the survey may have used a nested sampling design, which also leads to more than one error stratum. CANOCO 4.5 can account for these types of structure if,

- in a time series, the samples are taken at equal time intervals.
- in a spatial layout, the samples are at equal distances along a line transect or are arranged in a rectangular grid.
- in a study with more than one error stratum, the design is balanced, and the appropriate error term of the test can be obtained by shuffling groups of samples. The groups are called “whole-plots”. A typical example is the testing of a whole-plot factor in a split-plot design.

A study design may also consist of several blocks, each consisting of one of the above structures. We now discuss each structuring element in turn.

Blocks

Blocks are groups of related samples. Samples within blocks are permuted, whereas samples from different blocks are never exchanged. In CANOCO, blocks are defined by covariables. The variation between blocks is excluded from the statistical test. In ANOVA terminology, a block is a random factor that has no interaction with the factors that vary within blocks.

Time series and line transects

Under the null hypothesis that two stationary time series (sampled at equal time intervals) are unrelated, the starting point of the one series can be randomly linked to a time point of the other series. So, the null hypothesis is rejected if the observed correlation between the series is extreme in the reference distribution of correlations (or any other test statistic) generated by such random links. In the practical implementation of this idea, we face the problem that, after random linking, the start of the second series has no first series' points linked to it. Similarly, the end of the first series has no linked points either. Rather than using the linked points only, we use the trick of bending the time series into a circle, so that start and end meet (Besag & Clifford, 1989; section 5). This mathematical trick of using cyclic shifts works fine, provided there is no trend or cyclic structure, as is the case with stationary time series; the cyclic shift only corrupts the autocorrelation structure of each series at the beginning and the end of each series. For line transects, the dependence structure is not unidirectional as in time series. Usually, a point is related to its neighbors in both directions. Therefore, each observation series 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 in CANOCO, the test of association must use permutations which preserve marginal correlations, but change cross-correlations (B. Ripley, pers. comm.).

In CANOCO, the one series consists of the samples in the species data, whereas the other series consists of the samples in the environmental data. For the above permutation test to work, the series need to be trend-free under the null hypothesis. A series can be made trend-free by linear detrending. This is done in CANOCO by using covariables. For time series, time can be used as covariable, for line transects use position. Note, however, that a series should not be detrended a priori, if the aim of the test is to determine the significance of the trend. If the trend is cyclic or of some other nature, there are more advanced methods of detrending (Legendre & Legendre 1998: section 12.2).

Rectangular grids

The idea of random shifts (Besag & Clifford, 1989; section 5) can also be applied to data on a rectangular grid (with equal horizontal and vertical spacing). When wrapped around a torus (so that opposite sides meet), the samples can be randomly shifted (toroidal shifts). 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 (so obtaining grid D from grid A, shown 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. 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. If you disable shifts from the mirror image, CANOCO generates random shifts starting from grid A only.

For the above permutation test to work, the series need to be trend-free under the null hypothesis. The data can be made trend-free by linear or polynomial detrending (Legendre & Legendre 1998: section 13.2.1). This is done in CANOCO by using the spatial coordinates of a sample as covariables, one covariable for the horizontal position and one for the vertical position. Note, however, that the data should not be detrended a priori, if the aim of the test is to determine the significance of the spatial trend.

Whole-plots in a split-plot design

The term 'whole-plot' derives from an experimental design called the split-plot design. A split-plot design is a hierarchical design with two levels of units: whole-plots containing split-plots. Split-plots are the lowest level sampling units, i.e. the samples in the data file. Examples of two-level designs are samples-within-estuaries, plots-within-stands, plots-along-transects, relevés-within-time-series (in a study of permanent plots). In CANOCO, either the whole-plots or the split-plots can be permuted, or both. Whole-plots should be of equal size, because whole-plots with different numbers of samples cannot be permuted. Many experimental and sampling designs with more than one error stratum can be analyzed in the split-plot framework. The different permutation types available in the split-plot framework all define fewer distinct permutations than when samples are permuted completely at random.

The effect of environmental variables that vary between whole-plots (e.g. the whole-plot factors of a split-plot design) can be tested by permuting whole-plots completely at random while keeping the split-plots of each whole-plot together. This test is valid if whole-plots are exchangeable under the null hypothesis, as they are in an experiment with a balanced split-plot design. If the whole-plots form a time series, a line transect, or a spatial grid, the whole-plot permutations can be restricted to cyclic or toroidal shifts so as to account for autocorrelation among whole-plots. If your environmental variables vary little or not at all between whole-plots, the test will never show significant effects.

The effect of environmental variables that vary within whole-plots (e.g. the split-plot factors of a split-plot design) can be tested by permuting split-plots completely at random within whole-plots without permuting whole-plots. Whole-plots restrict the permutations in the same way as blocks, but without the necessity of block-defining covariables. This test is valid if split-plots are exchangeable under the null hypothesis, as they are in an experiment with a balanced split-plot design. If the split-plots form a time series, a line transect, or a spatial grid, the split-plot permutations can be restricted to cyclic or toroidal shifts so as to account for autocorrelation among split-plots. If the split-plots form parallel time series and time is an autocorrelated error component affecting all series, the same shift should be applied to all time series. In the standard split-plot design, split-plots of different whole-plots are unrelated and the permutations of split-

plots in different whole-plots should be independent. If your environmental variables vary little or not at all within whole-plots, the test will never show significant effects.

3.7.4 What is shuffled?

If there are no covariables in the analysis (or if all covariables are used to define blocks), it does not matter whether the samples in the species data or the samples in the environmental data are being permuted, and there is a wide choice of possible test statistics (of which CANOCO offers only two). The null hypothesis of the test is the overall null model: species are unrelated to the environmental data (within blocks, if defined). This is a simple hypothesis test, which can be compared with the overall F -test in a regression analysis. However, the research question is often more intricate, namely whether one variable has an effect on the species after taking into account the effect of another variable. Examples are:

- Does the management regime in the Dune meadow data have an effect after accounting for the fact that the meadows differ in moisture status and in thickness of the A1 horizon?
- Does nutrient pollution affect the species composition after taking into account the natural variation in salinity of the water?

In regression analysis, such questions are addressed by a t -test or, if the effect of more than one variable is of interest, a partial F -test. Such tests are called partial tests or conditional tests. It would be desirable to have a corresponding permutation test, which does not require the assumption of normality. The theory for such permutation tests is given in Ter Braak (1992) [Unimodal models: pp 217-223]. A multivariate form of the test is used for the multivariate methods used in CANOCO. In CANOCO, the variables of interest must be specified as the environmental variables, whereas the variables that are accounted for, are covariables.

To explain what is being permuted in a partial test, we introduce a multivariate regression model for the $n \times m$ matrix \mathbf{Y} of species data, namely,

$$(3.4) \quad \mathbf{Y} = \mathbf{X} \mathbf{B} + \mathbf{Z} \mathbf{C} + \mathbf{E}$$

where \mathbf{X} and \mathbf{Z} are fixed, known $n \times p$ and $n \times q$ matrices with covariable data and environmental data, respectively, \mathbf{B} and \mathbf{C} are $p \times m$ and $q \times m$ matrices of unknown and fixed regression coefficients and \mathbf{E} contains random errors with zero mean and constant, but unknown variance. Apart from the covariable data, \mathbf{X} is also assumed to contain a column with ones to take account of the intercepts of the regressions for each species. The value of p is thus "1 + the number of covariables". Notice that the usage of the letters \mathbf{X} , \mathbf{B} , and \mathbf{C} in this section differs from that in the remainder of this manual. Our interest focuses on the effects of the environmental variables in \mathbf{Z} on the species data in the presence of the covariables in \mathbf{X} , i.e. we want to test the null hypothesis that all elements of \mathbf{C} are 0, when the elements of \mathbf{B} are unknown. Model (3.4) is the basis of a RDA of \mathbf{Y} on environmental variables \mathbf{Z} with covariables \mathbf{X} .

To test $H_0: \mathbf{C} = \mathbf{0}$ (i.e. the effect of \mathbf{Z}) in this model it has been proposed to

1. Permute the rows of the species data \mathbf{Y} (Manly, 1991, 1997).
2. Permute the rows of the environmental data \mathbf{Z} (CANOCO 2.x, Collins, 1987).
3. Permute the residuals E_r of the regression of \mathbf{Y} on \mathbf{X} (CANOCO 3.x & 4.x, Freedman & Lane, 1983).
4. Permute the residuals E_r of the regression of \mathbf{Y} on \mathbf{X} and \mathbf{Z} (CANOCO 3.x & 4.x, Ter Braak, 1992).

Proposal 1 is attractive in that it is simple and stays close to the data and the study design. However, even if the data stem from a randomized experimental design, Y -values obtained at different values of X are not exchangeable under the null hypothesis, if X has an effect (B not equal to 0). The resulting test has the wrong type I error and also a low power if the nuisance parameters are important (Stapel & Ter Braak, 1994; Kennedy & Cade, 1996).

Proposal 2 stems from the idea that the values of Z are arbitrary under the null hypothesis and can thus be permuted. However, when permuting the rows of Z , the correlations between X and Z change, so that variables that were originally highly correlated become, on average, uncorrelated in the permutation test. In a fixed regression-design context, the resulting type I error is inflated if X and Z are correlated (Kennedy & Cade, 1996). With proposal 2 there is no logical basis for the testing of interaction effects. The problems in both these proposals can be alleviated by taking the F -ratio as the test-statistic instead of the regression mean square (Manly, 1991; CANOCO 2.x) as shown by Manly (1997), Kennedy & Cade (1996), and Anderson & Legendre (1999).

Proposals 3 and 4 explicitly use the regression model because residuals cannot be calculated without a model. Proposals 3 and 4 are therefore sometimes called **model-based permutation** methods. Note that the residuals are the best estimates of the random errors E . Independent and identically distributed random errors are exchangeable, but the residuals of a regression analysis are, strictly speaking, not exchangeable, except in some simple ANOVA models; in particular, the variance of the residuals may vary across units. Nevertheless, these methods produce type I errors that are close to the desired nominal value if a t - or F -ratio is used as test statistic and if the number of data points or, rather, the number of degrees of freedom, is large enough (say, $n - p - q > 10$). In technical terms, if an asymptotic pivotal test statistic is used and the number of degrees of freedom is large enough (> 10 , say) the methods have good level-accuracy (i.e. the reported P -value is accurate). Proposal 3 does slightly better than proposal 4 in this respect (Cade & Richards, 1996; Anderson & Legendre 1999; Anderson & Robinson, 2001). Proposals 3 and 4 are available in CANOCO 4.x under the names of “permutation under the reduced model” and “permutation under the full model”, respectively. In CANOCO 3.x the reduced model method was termed the “null model method”. Following Cade & Richards (1996), the name has been changed to avoid confusion of the null model of the test ($C=0$) with the overall null model ($B=C=0$).

Because CANOCO implements proposals 3 and 4, all sentences in the previous sections implying that the samples of the species data are permuted should be qualified to mean that the samples of the residualized species data are permuted. The residualization is with respect to X in the reduced-model method (proposal 3) and with respect to X and Z in the full-model method (proposal 4).

3.7.5 Model-based permutations

In this section, details are given of the permutation test using residuals from the reduced model (proposal 3 of the previous section). The method using residuals from the full method is described in Unimodal Models (pages 217-223). For linear methods (RDA), the steps of a Monte Carlo test outlined in section 3.7.2 work out as follows.

The permutation test of the effect of Z , adjusted for the possible effects of X , in RDA

Step 1. Choose a test statistic.

As one test statistic, we choose the F -ratio of the partial F -test to test the null hypothesis $\mathbf{C} = \mathbf{0}$. This F -ratio is calculated by the following procedure:

Regress the species data \mathbf{Y} on the covariable data \mathbf{X} , then add the environmental variables \mathbf{Z} to the regression, giving residual sums of squares RSS_X and RSS_{X+Z} , respectively. Calculate the F -ratio for testing the null hypothesis $\mathbf{C} = \mathbf{0}$ from RSS_X and RSS_{X+Z} by the formula

$$(3.5) \quad F = \{ (RSS_X - RSS_{X+Z}) / q \} / (RSS_{X+Z} / (n-p-q)) \}$$

The F -ratio in (3.5) differs from the usual partial F -statistic in (univariate) regression analysis only in that the residual sums of squares, from which it is calculated, are totaled across all species. It could be called the pseudo- F . In CANOCO, $RSS_X - RSS_{X+Z}$ is equal to the sum of all canonical eigenvalues (after controlling for the effects of the covariables). Recall also that p is 1 + the number of covariables.

Step 2. Calculate the test statistic for the data, yielding F_0 .

Step 3. Generate K new data sets that are equally likely under the null hypothesis.

Each new data set is generated by the following two substeps, the first of which gives the same results for each new data set and therefore needs to be carried out only once. The two substeps are:

1. Regress \mathbf{Y} on \mathbf{X} , yielding fitted values $\hat{\mathbf{Y}}$ and residuals \mathbf{E} , with $\mathbf{E} = \mathbf{Y} - \hat{\mathbf{Y}}$.
2. Permute the rows of \mathbf{E} to yield \mathbf{E}^* and calculate the new data set $\mathbf{Y}^* = \hat{\mathbf{Y}} + \mathbf{E}^*$.

The first substep yields residuals that are not correlated with the covariables from the data. Under the null hypothesis, these residuals can be permuted (even though they are not strictly exchangeable).

Step 4. Calculate the test statistic for each new data set \mathbf{Y}^* .

This step is carried out as step 1, with \mathbf{Y}^* replacing \mathbf{Y} , leading to the F -ratios F_1, F_2, \dots, F_K .

Step 5. Calculate the Monte Carlo significance level.

Place F_0 among F_1, F_2, \dots, F_K and determine the proportion of values greater than or equal to F_0 .

As remarked in Ter Braak (1992), the F -ratio (3.5) in step 4 can also be obtained by regressing the permuted residuals \mathbf{E}^* directly on \mathbf{X} and then adding \mathbf{Z} to the regression. Kennedy & Cade (1996) claimed that this procedure could be further simplified to a regression on residualized \mathbf{X} , but this claim was shown to be false by Anderson & Legendre (1999). CANOCO implements the correct Freedman & Lane (1983) procedure.

The permutation test in CCA

The above permutation test finds its rationale in linear theory. We now address the question whether and how the test can be applied to unimodal methods, in particular CCA. As noted in section 3.4, CCA has two faces, a unimodal one and a linear one. Its unimodal face is only visible with strong gradients (> 3 SD) whereas the linear face is particularly useful for short gradients (< 3 SD). The null hypothesis of no additional environmental effects on the species implies that there are no meaningful environmental gradients after accounting for the covariables: their true gradient lengths are all 0. The statistical test aims at detecting small,

systematic deviations from the null hypothesis. These deviations express themselves in small true gradient lengths (even if the observed gradient lengths are large). The linear face is therefore most pertinent for developing statistical tests, and is used in CANOCO for permutations tests in CCA.

In terms of a linear model, CCA is based on a weighted multivariate regression of transformed species data \mathbf{Y}' on the covariable data \mathbf{X} and environmental data \mathbf{Z} (Sabatier et al. 1989; Box 3 of Unimodal models: p 162). Denote the original species data by $\mathbf{Y} = \{y_{ik}\}$ and the transformed data by $\mathbf{Y}' = \{y'_{ik}\}$ [$i = 1, \dots, n$; $k = 1, \dots, m$]. Then, the data transformation can be written as

$$(3.6) \quad y'_{ik} = (y_{ik} y_{i+}) / (y_{i+} y_{+k})$$

where the subscript + replacing an index indicates the sum over the subscript. The row and column weights of the multivariate regression are denoted by $\{w_i\}$ [$i = 1, \dots, n$] and $\{w_k\}$ [$k = 1, \dots, m$] with w_i and w_k proportional to y_{i+} and y_{+k} , respectively (see the text surrounding equations (6.4) and (6.5) on page 156 for details when there are also user-defined weights). The theory of the previous two sections can be applied by transforming the weighted regression to an unweighted regression by pre-multiplying both sides of the regression by the square root of the row weights (e.g. Seber, 1977) and post-multiplying the left hand-side by the square root of the column weights. The post-multiplication by the column weights results in regression coefficients that need to be backtransformed (i.e. divided by the column weights) to obtain the same coefficients as in the weighted regression (cf. Unimodal models: p 182), but this is unimportant here, as the regression coefficients are not needed themselves.

It is of some interest to understand what happens without the mathematical trick of transformation to an unweighted problem. We only discuss the role of the row weights, because the rows are permuted, with all column values kept in the same order. Assume, as usual, that the weights of a weighted regression are inversely proportional to the variance of the data y'_{ik} . Then, the residuals of a row tend to be larger, when the row weights are smaller. Such residuals cannot be permuted. Therefore, it makes sense to standardize the residuals before permutation by multiplication by the square root of the row weight, i.e.

$$(3.7) \quad e'_{ik} = w_i^{1/2} e_{ik}$$

with e_{ik} the residual of species k in sample i obtained from the weighted regression. The standardized rows are now permuted. After permutation the standardized residuals are denoted by e'_{ik}^* . Before the standardized residuals can be added to the fitted values of a particular row, they need to be scaled to the variance of that row by division by the square root of the row weights, i.e.

$$(3.8) \quad e_{ik}^* = w_i^{-1/2} e'_{ik}^*$$

The new data set \mathbf{Y}^* is obtained by adding the values $\{e_{ik}^*\}$ to the fitted values $\hat{\mathbf{Y}}$. The new data \mathbf{Y}^* is then analyzed by weighted regression, i.e. by CCA.

Note that the sums of squares that define the F -ratio in CCA are weighted sums of squares. As shown in the section on Ordination diagnostics, in particular in subsection 6.3.11.2 (page 174), these weighted sums of squares can be interpreted as chi-square statistics. An early application of chi-square statistics in permutation tests appeared in Manly (1983). From this point of view, CANOCO extends Manly's (1983) permutation test to quantitative explanatory variables and partial tests.

The assumption that the weights that are used in CCA are inversely proportional to the variance of y'_{ik} is not essential for the rationale of the test. In any weighted regression, rows with large weights tends to fit more closely to the data than rows with small weights. Consequently, the residuals of rows with large weights tend to be smaller than the residuals of rows with small weights. This practical observation sufficiently motivates the weighting scheme in (3.7) and (3.8). Nevertheless it is of interest to note that the assumption holds true if, for example, the raw data $\{y_{ik}\}$ are Poissonian counts in a contingency table in which the rows and columns are independent. Under the Poisson model, the variance is equal to the mean, i.e.

$$(3.9) \quad \text{var}(y_{ik}) = \mu_{ik}$$

so that the variance of the transformed data is

$$(3.10) \quad \text{var}(y'_{ik}) = \text{var}(y_{ik} / \mu_{ik}) = \mu_{ik}^{-2} \text{var}(y_{ik}) = 1 / \mu_{ik}$$

so that the weight is

$$(3.11) \quad w_{ik} = \mu_{ik} = y_{i+}y_{+k} / y_{++}$$

under the independence model.

Test statistics

As described above, the permutation test uses the F -ratio (3.5) as test statistic. This test statistic is available in CANOCO under the name “Test of significance of all canonical axes” or “Test based on the trace statistic”. An alternative test statistic available in CANOCO is the first canonical eigenvalue λ_1 , expressed as the F -ratio

$$(3.12) \quad F_\lambda = \lambda_1 / (\text{RSS}_{X+1} / (n-p-q))$$

where RSS_{X+1} is the residual sum of squares of the model with all covariables and the first ordination axis of the environmental data. In other words, RSS_{X+1} is the sum of squares of the residuals of the regression in which a rank 1 restriction is imposed on the matrix of regression coefficients C in (3.4) (see, for example, Ter Braak & Looman, 1994). Recall also that p is 1 + the number of covariables. The test statistic F_Σ is calculated for the original data Y in step 2 and for each new data set Y^* in step 4. This test statistic has maximum power against the alternative hypothesis that there is a single dominating gradient that determines the relation between species and environment.

3.7.6 Multifactorial analysis of variance

In this section we show how to test individual factors and interactions in experiments with an orthogonal design. A guideline is given for constructing valid permutation tests in complex analyses with fixed and random factors by using the split-plot framework of CANOCO.

Consider a randomized factorial experiment with two fixed crossed factors A and B with levels a and b , with r replicate samples per combination of levels of A and B . The usual main effects model for this experiment can be written as

$$(3.13) \quad y_{ijk} = \mu + \alpha_i + \beta_j + e_{ijk}$$

where y_{ijk} is the response at replicate k of the i th level of factor A and the j th level of factor B. Suppose we wish to test the main effect of A. If factor A does not have an effect on the response, samples within each level of B are exchangeable and can thus be permuted (Edgington, 1995). In CANOCO, this can be achieved (using permutations under the reduced model) by treating each level of B as a block, so that permutations are restricted within levels of B. The same test results can be obtained by defining each level of B as a whole-plot and by requesting that whole-plots are not permuted, whereas split-plots are randomly permuted. If r is small (e.g. $r = 1$), then this exact test will have low power, merely by the limited number of permutations that are possible. This method of permutation is sometimes called “design-based”, but note that the permutations for the test differ from the randomization scheme of the experimental design. In the experimental design, the levels of B have no special status: the units are randomized over A.B treatment combinations.

In the model-based permutation approach, the possible effect of the factor B is eliminated by residualizing y with respect to the factor B, i.e. the mean of each B-level is subtracted from the observations. The permuted residuals are then analyzed by ANOVA in the same way as the original data, yielding for each permutation a permutation F -ratio. Each permutation is completely random, corresponding to the randomization in the experimental design. However, there is no exact permutational argument for this test, because residuals of different levels of B do not have identical distributions. The test is asymptotically exact (i.e. for large r) and almost exact, if the number of degrees of freedom for the residual exceeds 10, say, as follows from the arguments in Hall & Titterton (1989) and as verified by simulations by various authors (e.g. Fisher & Hall, 1990; Manly, 1997; Anderson & Legendre, 1999). In CANOCO, the test is obtained by entering dummy variables indicating the levels of B as covariables, and dummy variables indicating the levels of A as environmental data and by asking for unrestricted permutation under the reduced model. A real data example is experiment E40 in section 8.3.3.

Now suppose the model includes an interaction effect allowing the effect of A to depend on the level of B, written as follows:

$$(3.14) \quad y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + e_{ijk}$$

There is no permutation test of the interaction effect in the design-based approach (Edgington, 1995), except in special cases (Welch, 1990; section 8.3.13). In the model-based approach, the test of interaction presents no problem: in CANOCO, specify indicators for the levels of A and B as covariables and indicators for A.B combinations as environmental data, and ask for unrestricted permutations. If the interaction effect is significant, it does not make sense usually to test for the main effect of A. If the interaction effect is not significant, one may proceed as in model (3.13) to test the main effects. The corresponding approach in classical ANOVA consists of pooling the A.B interaction term with the error term. A permutation test equivalent to “not pooling” can be obtained by entering the main effect of B and the interaction contrasts as covariables. For completeness, we add that this last permutation test also allows the testing of the main effect of A if the interaction effect is judged significant.

Now suppose that factor A is a fixed factor and factor B is a random factor. For example, in section 8.3.6 the effect of sod-cutting is investigated in each of a number of different forests. Then, sod-cutting is a fixed factor and forest is a random factor. If the effect of sod-cutting is the same in all forests, i.e. if there is no interaction effect, then the experiment can be analyzed as a randomized block design with forest as block. The possible methods to test the effect of sod-cutting do not differ from the methods used for model (3.13) with B a fixed factor. In this example, the design-based approach is perhaps most appealing, as plots from different forests

cannot be randomized. If we expect that the effect of sod-cutting may vary between forests, then the model becomes

$$(3.15) \quad y_{ijk} = \mu + \alpha_i + b_j + ab_{ij} + e_{ijk}$$

with Roman letters indicating random terms. Table 3.2 shows the classical ANOVA table for this model. The denominator for the classical F -test shows that the random factor B and the random interaction must be tested against the lowest stratum error variance, but that the fixed factor A must be tested against the A.B interaction (Table 3.2). This interaction has less degrees of freedom than the residual error. A permutation test that permutes the individual replicates would be too liberal as a test for the main effect of A. A valid permutation test can be obtained by averaging over the replicates per A.B combination, leading to the model

$$(3.16) \quad y_{ij.} = \mu + \alpha_i + b_j + e'_{ij}$$

where $e'_{ij} = ab_{ij} + e_{ij}$ and where a dot replacing an index means averaging over the index. In this model, the A.B interaction and the lowest level error are indistinguishable. Model (3.16) is the same as the model of a randomized block design (the dot as index of y_{ij} just being notation). There are thus two ways to test the main effect of factor A:

1. The design-based test of permuting the levels of A within each level of B.
2. The model-based test of permuting the residuals (estimates of e'_{ij}), under either the reduced model or the full model, by using factor B as covariable.

The split-plot design framework of CANOCO 4.5 makes it possible to carry out this permutation test without the need to reduce the data to averages per A.B combination yourself. For this test,

- Define the A.B combinations as whole-plots and r individual replicates per combination as split-plots. The number of whole-plots is thus ab .
- Define the levels of factor B as covariables and, for the design-based method, also as blocks.
- Exchange whole-plots freely and do not permute split-plots.

The design-based method yields an exact test (it is not a partial test because the nuisance factor B is conditioned upon). The model-based method has good level-accuracy, provided the test-statistic is of the correct form (asymptotic pivotal). In the notation of Table 3.2, the correct test statistic is the F -ratio $MS(A)/MS(AB)$. Unfortunately, CANOCO 4.5 continues to use the pseudo- F of (3.5) which amounts to $MS(A)/MS(AB+Error)$. The resulting test is too liberal (Anderson & Ter Braak, 2002). A way around this is to use the exact, design-based version of the test, if that makes sense, by specifying all covariables as blocks. Alternatively, calculate totals (or means) per whole-plot yourself, as in (3.16), and enter these as "species data" in Canoco. See also section 8.3.6.2.

Note that the permutation test on whole-plots uses the correct number of degrees of freedom. In CCA it would be more logical to take sums over y_{ijk} rather than means, and that is the equivalent of what happens by using the split-plot framework of CANOCO in a CCA.

Table 3.2 Expected mean squares.

Factor A (with a levels) is fixed, and factor B (with b levels) is random. The factors A and B are crossed. At each combination of levels of A and B there are r replicates.

Source	degrees of freedom	expected mean square	F -denominator
B stratum			
B	$b-1$	$\sigma^2 + ar \sigma_b^2$	MS(error)
AB stratum			
A	$a-1$	$\sigma^2 + r \sigma_{ab}^2 + br \sigma_a^2$	MS(AB)
AB	$(a-1)(b-1)$	$\sigma^2 + r \sigma_{ab}^2$	MS(error)
Units stratum			
Error	$ab(r-1)$	σ^2	*

This example serves to demonstrate that the conventional analysis-of-variance table with associated expected mean squares is as useful to derive a valid permutation test for a particular factor or interaction as it is to derive the correct F -ratio in the usual normal theory-based tests (Anderson & Ter Braak 2002). In particular, if the denominator of the F -ratio is not the residual mean square at the lowest stratum, then the samples themselves are not the appropriate permutable units and, with a model-based permutation, the test statistic must use the correct F -ratio. The source term of denominator of the F -ratio indicates which are the appropriate permutable units. With this guideline, many experimental and sampling designs with more than one error stratum can be analyzed in the split-plot framework of CANOCO.

Note that in repeated measurement designs (such as Before-After-Control-Impact (BACI) designs) the quantities of interest are often an interaction with time and are thus on the lowest stratum. The usage of the whole-plot set-up for testing such interaction effects is needed for another reason, namely autocorrelation in time. To avoid that terms from higher strata enter the denominator of the CANOCO F -statistics, one must define each site that is being measured over time, as covariable, as is done in the BACI examples in section 8.3.7 - 8.3.10. In another example, the Principal Response Curves (PRC) method (section 8.3.11 and 8.3.12), the significance of the model “treatment + treatment.time” is tested. The permutation test works in PRC because it is design-based; it is not model-based because the contribution of the covariables, which code for the main effect of “time”, is constant in the test.

Further research is needed to fully explore the possibilities and pitfalls of the split-plot framework of CANOCO.

3.8 RDA and CCA as regression procedures: reduced-rank regression

Chapter 14 and 15 of Unimodal Models present the theory of redundancy analysis (RDA) in terms of reduced-rank regression, resulting in so-called regression biplots. These chapters are rather mathematical. This section explains the basic theory in more simple terms without the use of matrix algebra.

We first show that one-dimensional PCA and RDA (with a single ordination axis) are in-between simple regression and multiple regression. In simple linear regression, the abundance of each species separately is regressed on a single explanatory variable x , a known variable, e.g. pH. The model can be written as

$$(3.17) \quad y_{ik} = a_k + b_k x_i + \text{error}$$

where y_{ik} is the abundance of species k in sample i , x_i is the known value of the explanatory variable in sample i , and a_k and b_k are unknown regression coefficients that are to be estimated. This model specifies a straight-line relationship between the species' abundance and x , with a_k and b_k the intercept and slope parameter of the straight line, respectively. Now suppose, as in Jongman et al. (1987: pp 116-118), that the values $\{x_i\}$ are unknown. With data from m species, we can then try to find a theoretical explanatory variable that gives the best fitting straight lines. PCA does this: it finds the optimal values $\{x_i\}$, i.e. the sample scores $\{x_i\}$ for which model (3.17) fits best. These values do not need to have a relation to any measured environmental variable. The idea of RDA is to constrain the values by requiring that x is a linear combination of measured environmental variables (a weighted aggregate). With two environmental variables, the constraint is

$$(3.18) \quad x_i = c_1 z_{i1} + c_2 z_{i2}$$

RDA gives the best weighted aggregate, i.e. the optimal weights c_1 and c_2 . The weights are called canonical coefficients. On inserting (3.18) into (3.17) we obtain the model

$$(3.19) \quad y_{ik} = a_k + b_k c_1 z_{i1} + b_k c_2 z_{i2} + \text{error}$$

RDA estimates the unknowns in this model, i.e. the species parameters a_k and b_k ($k = 1, \dots, m$) and the weight parameters c_1 and c_2 , from the species data $\{y_{ik}\}$ and environmental data $\{z_{ij}\}$. By defining

$$(3.20) \quad d_{1k} = b_k c_1 \text{ and } d_{2k} = b_k c_2$$

the RDA-model (3.19) can be written as a multiple regression model, namely

$$(3.21) \quad y_{ik} = a_k + d_{1k} z_{i1} + d_{2k} z_{i2} + \text{error}$$

RDA is thus a multiple regression for all species simultaneously (i.e. a multivariate regression) with linear constraints on the regression coefficients. Because of the constraints, RDA uses less parameters than multivariate multiple regression. In summary, RDA is thus both a constrained form of PCA and a constrained form of multivariate multiple regression.

Apart from notational difficulties, it is straightforward to extend the above formulae from one dimension to more dimensions. In two dimensions, the model for the species data becomes

$$(3.22) \quad y_{ik} = a_k + b_{k1} x_{i1} + b_{k2} x_{i2} + \text{error}$$

where b_{ks} is the species score of the k th species and x_{is} is the sample score of the i th sample on the s th ordination axis ($s = 1, 2$). In RDA, the sample scores are constrained by

$$(3.23) \quad x_{is} = c_{1s}z_{i1} + c_{2s}z_{i2}$$

with c_{js} the canonical coefficient of the j th environmental variable on the s th ordination axis. By inserting (3.23) into (3.22) we obtain a multiple regression model with constraints on the regression coefficients. If the regression coefficients are denoted by $\{d_{jk}\}$ as in equation (3.21), the constraints are

$$(3.24) \quad d_{jk} = b_{k1}c_{j1} + b_{k2}c_{j2}$$

This equation forms the basis of the so-called regression biplot: the species scores $\{b_{ks}\}$ plotted together with the canonical coefficients $\{c_{js}\}$ represent the regression coefficients $\{d_{jk}\}$ by means of the biplot rule. Similarly, we see from (3.22) that the sample scores $\{x_{is}\}$ together with the species scores $\{b_{ks}\}$ form a biplot of the fitted species values.

With m and q the number of species and environmental variables, respectively, an r -dimensional RDA-model uses $m + r(q+m) - r^2$ parameters¹ (Robinson, 1973), whereas multivariate multiple regression uses $m + mq$ parameters. For example, if $m=100$, $q=10$ and $r=2$, RDA uses 316 parameters, whereas multivariate multiple regression uses 1100 parameters. If $r=q$, there are no constraints left and the numbers of parameters coincide. In multivariate multiple regression, the $q \times m$ matrix \mathbf{D} of regression coefficients $\{d_{jk}\}$ has rank $\min(m,q)$. With the linear constraints (3.20) the rank of the matrix \mathbf{D} is reduced to 1. In an r -dimensional RDA, the rank of \mathbf{D} is r , hence the name reduced rank regression.

The above theory for RDA can be extended to CCA by using weighted regression on transformed data as specified around equation (3.6) on page 50.

3.9 Compositional data

3.9.1 Introduction

Compositional data *sensu stricto*, also called percentage data, are obtained when, for example, for each sampling unit a fixed number of individuals is counted and each individual is identified to belong to one of m species. This sampling method is common in palynology and diatom research. Information from such a sample resides in the fraction of individuals belonging to each of the species. Compositional data also frequently arise in chemistry and geology where a sample is analyzed into its constituents. In this section we present two alternative methods of analysis of such data. The first method is log-ratio analysis and is based on a series of papers by Aitchison (1982-90). Log-ratio analysis amounts to applying linear methods to log-percentage data which are centered both by samples and by species (see also Aitchison, 1986). Because the logarithm of the percentages is analyzed, the method is attractive only when the data contain few zero values. The interpretation of the resulting biplots is discussed in Aitchison (1990) and summarized in Unimodal models (pp 144-145). The second method derives from a generalized linear model and amounts to applying unimodal methods to the untransformed percentage data. It is appropriate when the data contain many zeroes.

¹ The number of parameters given in Ter Braak & Prentice (1988) [Unimodal Models: pp 102] for $r = 2$ is not quite correct, as was pointed out by J. Van der Meer.

3.9.2 Compositional data not containing zeroes: log-ratio analysis

Let, for this section only, p_{ik} be the fraction of species k in sample i ($\sum_k p_{ik} = 1$; $p_{ik} > 0$). Because the fractions are positive, it is not acceptable to model them by a linear model such as

$$(3.25) \quad \eta_{ik} = a_k + b_k x_i + \varepsilon_{ik}$$

because there is nothing to prevent the right-hand side from resulting in a negative value. Equation (3.25) is the familiar straight line regression model with a_k the intercept, b_k the slope parameter or regression coefficient, x_i the value of an explanatory variable x and ε_{ik} an error term with mean zero and variance σ_k^2 . The problem that the equation can result in negative values could be solved by modeling the fractions by $\exp(\eta_{ik})$, but then the model values still do not need to sum to 1. This problem is solved by dividing the $\exp(\eta_{ik})$ by their sum, yielding

$$(3.26) \quad p_{ik} = \exp(\eta_{ik}) / \sum_j \exp(\eta_{ij})$$

The fractions $\{p_{ik}\}$ are said to follow a logistic normal distribution if the error ε_{ik} in (3.25) follows a normal distribution with mean 0 and covariance matrix Σ (Aitchison, 1982: p. 162). Now we have posed a model for fractions, we derive a method of analysis. Retracing the steps of the preceding argument, we take logarithms of the fractions and obtain from (3.25) and (3.26)

$$(3.27) \quad \log p_{ik} = \gamma_i + a_k + b_k x_i + \varepsilon_{ik}$$

where $\gamma_i = -\log(\sum_j \exp(\eta_{ij}))$ is an incidental parameter. Fortunately, the incidental parameters $\{\gamma_i\}$ can be removed by centering the log-fractions by samples. When the data are also centered by species we obtain quantities y_{ik}

$$(3.28) \quad y_{ik} = q_{ik} - q_{i.} - q_{.k} + q_{..}$$

where $q_{ik} = \log(p_{ik})$ and a dot replacing an index denotes that the average is taken over the index. On inserting (3.27) into (3.28) we obtain a linear model for the quantities $\{y_{ik}\}$,

$$(3.29) \quad y_{ik} = b_k^* x_i^* + \varepsilon_{ik}^*$$

where $b_k^* = b_k - b_{.}$, $x_i^* = x_i - x_{.}$ and $\varepsilon_{ik}^* = \varepsilon_{ik} - \varepsilon_{i.} - \varepsilon_{.k} + \varepsilon_{..}$ is still an error term with mean 0. Note that there is nothing in the above derivation which prevents us from using more than one explanatory variable in (3.25). Percentage data without zero values can thus be analyzed with the linear methods available in CANOCO by using the log-transformation and centering both by samples and by species. For this, the data do not need to be transformed *a priori* to percentages or fractions. The log-transformation and the double centering automatically take account of this. When using principal components analysis (PCA) in this way, one obtains what Aitchison (1984b: p. 622) calls loglinear-contrast principal components. The interpretation of the resulting biplots is discussed in Aitchison (1990) and summarized in Unimodal models (pp 144-145). Use of redundancy analysis (RDA) opens up the possibility of applying regression analysis to percentage data (cf. section 3.8). The Monte Carlo permutation test is then useful to test the effect of particular environmental variables. Data transformation (3.28) is equivalent with the centered logratio transformation of a composition of Aitchison (1986).

As an example we use the boxite and coxite data sets presented by Aitchison (1984a: pp. 535-536) which each consist of the percentages of five chemical constituents in 25 samples of rock taken at different depths. We test the hypothesis whether the chemical composition of

the rock samples depends on depth. For the boxite and coxite data we obtain P-values of 0.59 and 0.21, respectively. Using a different test statistic, Aitchison (1984: p. 553) obtained P-values of ca. 0.35 and 0.001, respectively. There is a discrepancy for the coxite data which is caused by the large residual correlations among the constituents in these data. The type I error of both tests is the same, but the type II error of the test in CANOCO is larger than for Aitchison's test statistic. Aitchison's test which is based on the standard multivariate linear hypothesis is more powerful when there are large residual correlations. Although Aitchison's test detects that the composition of coxite is significantly related to depth, depth explains only 6% of the variance (because $\lambda_1 = 0.06$). The other variable given by Aitchison, porosity of the rock, is much more strongly related to composition: it explains 39% of the variability and is significant ($P = 0.01$ in 99 Monte Carlo permutations). The coxite data are the example data of section 8.4.2.

3.9.3 Compositional data containing zeroes: CA, DCA and CCA

Zero values present a problem in the preceding approach because the logarithm of 0 is $-\infty$. When the data contain few zeroes the problem may be circumvented by replacing zeroes by an arbitrary small value or by adding a small value to all numbers², but this is unattractive when there are many zeroes because the result may depend considerably on the choice of the value replacing zero.

An alternative approach can be based on a generalized linear model (McCullagh & Nelder, 1989) for percentage data. Instead of defining a model for observed fractions as is done in (3.27) we define a model for expected fractions. Let y_{ik} from now on be the fraction of species k in sample i and Ey_{ik} the expected fraction. By analogy with (3.25) and (3.26), we define the multinomial logit model (e.g. McCullagh & Nelder, 1989: p. 159; Anderson, 1984: p. 5), also called the generalised logit model,

$$(3.30) \quad Ey_{ik} = \exp(\eta_{ik}) / \sum_j \exp(\eta_{ij})$$

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

$$(3.31) \quad \eta_{ik} = a_k + b_k x_i$$

In comparison with (3.25), the error term has been dropped in (3.31). As McCullagh & Nelder (1989: p. 212) note, the regression coefficients in (3.31) can be estimated from data $\{y_{ik}\}$ and known $\{x_i\}$ by using standard computer packages by transforming (3.30) to a loglinear model (see equation (3.36)).

When both the $\{b_k\}$ and the $\{x_i\}$ are unknown, Eqs. (3.30) and (3.31) define an ordination model for percentage data. There are then two routes which show that approximate estimates of the unknown parameters can be obtained by applying correspondence analysis to the fractions $\{y_{ik}\}$.

The first route begins by rewriting (3.30) and (3.31) and using a first order Taylor approximation (Ihm & van Groenewoud, 1984: p. 49)

² In CANOCO, the value 0.1 cannot be added directly. Instead of $\log(y+0.1)$ use the transformation $\log(10y+1)$. See sections 5.6.2 and 8.4.2.

$$(3.32) \quad E y_{ik} = \gamma_i^* \alpha_k^* \exp(b_k x_i) \approx \gamma_i^* \alpha_k^* (1 + b_k x_i)$$

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

$$(3.33) \quad y_{ik} = (y_{i+y+k}/y_{++}) (1 + b_k x_i)$$

This is the reconstitution formula of correspondence analysis (Greenacre, 1984: p. 93; Ter Braak, 1985: p. 861). This similarity was noted also by Goodman (1981); the equality in (3.32) defines Goodman's RC-model that can be written as

$$(3.34) \quad \log E y_{ik} = r_i + c_k + b_k x_i$$

with $r_i = \log(\gamma_i)$ and $c_k = \log(\alpha_k)$.

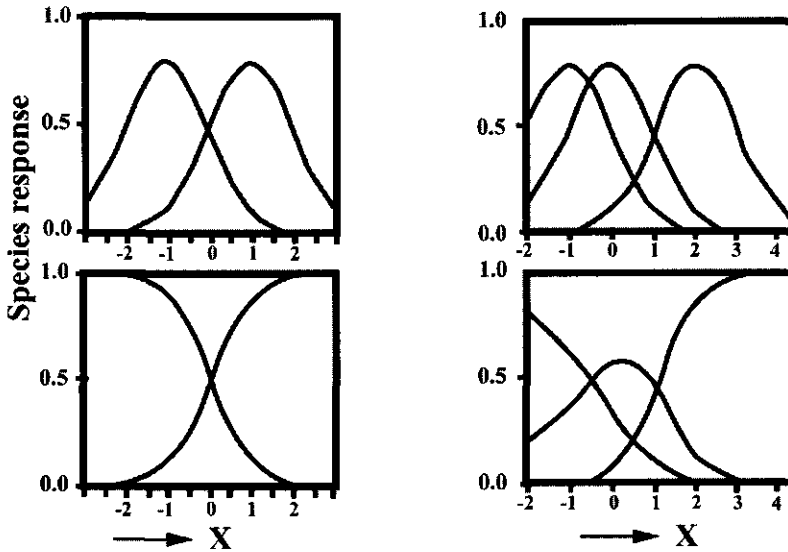


Figure 3-3 Comparison of Gaussian and multinomial logit models.

The top figures display the Gaussian model $E y_{ik} = \exp(\eta_{ik})$ for the abundance of two or three species along a gradient X ($m=2$, left; $m=3$, right) and the bottom figures display the corresponding model for percentage data, i.e. the multinomial logit model. After Ihm & Van Groenewoud (1984).

The second route begins by noting that (3.30) does not change when (3.31) is replaced by (see Figure 3-3)

$$(3.35) \quad \eta_{ik} = a_k^* - \frac{1}{2}(x_i - u_k)^2$$

with $a_k^* = a_k + \frac{1}{2} u_k^2$ and $u_k = b_k$; the missing term in (3.31), $\frac{1}{2} x_i^2$, cancels out because it occurs in both the numerator and denominator of (3.30). The top panels of Figure 3-3 show equi-tolerance Gaussian curves. The lower panels are derived from the top panels by division by the sample total and show the multinomial logit model of equation (3.30) with (3.31) or, equivalently, with (3.35) for two (left column) and three (right column) species. Further, (3.30) and (3.35) can be written as the general loglinear model (McCullagh & Nelder, 1989: p. 159, p. 212)

$$(3.36) \quad \log E y_{ik} = \gamma_i + a_k^* - \frac{1}{2}(x_i - u_k)^2$$

where $\gamma_i = -\log [\sum_j \exp(\eta_{ij})]$ is an incidental parameter (c.f. (3.27)). (We can take the logarithm of $E y_{ik}$ because $E y$ is a positive value even if some of the $\{y_{ik}\}$ are 0). Except for the incidental parameter, model (3.36) was taken as the starting point by Ter Braak (1985) in showing that correspondence analysis gives under particular conditions an approximate solution to the fitting of a unimodal model by maximum likelihood. It turns out that the derivation carries through also with the incidental parameter included (Unimodal models: pp 84-85). Moreover, the derivation of canonical correspondence analysis given in Ter Braak (1986) carries through equally for percentage data using model (3.30) with (3.35) and the constraint that x_i is a linear combination of environmental variables. Note that (3.30) and (3.35) together define Ihm and Van Groenewoud's model B.

In section 3.5 and the later sections 5.5 and 6.3.4 it is noted that there is some arbitrariness of how to scale the sample scores $\{x_i\}$ with respect to the species scores $\{u_k\}$. The scaling is governed by the value of α in the section 6.3.4. For incidence and abundance data there exists a best fitting value of α (which is unfortunately unknown in general). But for percentage data α is completely arbitrary: the model (3.30) with (3.35) does not change if we take, instead of u_k , x_i and a_k^* , the terms βu_k , x_i / β and $a_k^* + \frac{1}{2}(\beta-1)u_k^2$, respectively. Moreover, the optima $\{u_k\}$ may be shifted arbitrarily with respect to the sample scores $\{x_i\}$; with a shift from u_k to $u_k + d$ for a constant d , just change the value of a_k^* to $a_k^* + du_k$.

The model can be extended to two-dimensions by taking

$$(3.37) \quad \eta_{ik} = a_k^* - \frac{1}{2}[(x_{i1} - u_{k1})^2 + (x_{i2} - u_{k2})^2]$$

Again correspondence analysis can be used to obtain approximate estimates, except that detrending may be required to remove the arch effect when it occurs (see Unimodal Models, p. 51-52).

3.9.4 The two faces of compositional data revisited

It is rather paradoxical that the unimodal model (3.36) can be transformed to the linear model (3.34) with $b_k = u_k$. The intrinsic unimodality of (3.34) is proved by Anderson (1984). This apparent paradox was already noted to occur in correspondence analysis in the discussion of Ter Braak (1985). In compositional data without zeroes there even exists an explicit linearizing transformation (Kooijman, 1977). The linearizing transformation is given in equation (3.28), in words: "take logarithms and double center". The linearization can be seen as follows. Assume that the response curve of each species with respect to variable x is an equi-width Gaussian curve of the form

$$(3.38) \quad f_k(x) = a_k \exp\{-\frac{1}{2}(x - u_k)^2\}$$

where a_k is the maximum expected response attained at $x = u_k$, the optimum. Let $q_{ik} = \log(f_k(x_i))$ [$i = 1 \dots n$; $k = 1 \dots m$] and derive y_{ik} from the $\{q_{ik}\}$ by the data transformation (3.28). After this data transformation, the squared terms in x_i and u_k drop out so that the response model after transformation is linear, without an intercept. For data with error, a linear model with an additive error term is obtained, as in (3.29). Note that (3.28) is identical to the centered logratio transformation of a composition, thus showing an important link between data analysis of compositional data and unimodal data. It emerges that models for compositional data have both a linear and a unimodal face.

If the data contain zeroes, there does not exist any explicit linearizing data transformation, because we can not take logarithms. The previous section shows two solutions to this problem: turning to a generalized linear model in which the expected proportions can be transformed to linearity or applying a transformation that is close to the exact transformation. The latter is done in (canonical) correspondence analysis. For comparison with the transformation (3.6) that is implicit in correspondence analysis, the exact transformation can be written as $\log(y'_{ik})$ with

$$(3.39) \quad y'_{ik} = (y_{ik} g_{++}) / (g_{i+} g_{+k})$$

where g_{i+} and g_{+k} denote the geometric averages of the abundance data y_{ik} across rows and columns, respectively, and g_{++} the overall geometric average. Correspondence analysis inherits its two faces (section 3.4) from models for compositional data.

3.10 Nominal response data

Outside ecology, correspondence analysis is most frequently applied to nominal data (Gifi, 1990, Greenacre, 1984). Nominal data arise when each response variable consists of a series of mutually exclusive categories or classes. For example, vegetation type and soil type are nominal variables. Correspondence analysis can be applied to nominal variables to investigate their interrelations. It is termed multiple correspondence analysis (Greenacre, 1984) or homogeneity analysis (Gifi, 1990) if there are more than two such variables. When interest focuses on how the nominal response variables depend on external explanatory variables, one can use canonical correspondence analysis. Applied to such data, it is equivalent to redundancy analysis of qualitative variables (Israels, 1984). Torgerson (1958: p. 338) already described the types of data which can be analyzed by what is now called correspondence analysis. In biology, nominal data are encountered frequently in numerical taxonomy and genetics.

To analyze nominal response variables with CANOCO, each nominal variable must be represented by a series of dummy variables each representing a category: $y_{ik} = 1$ or 0 depending on whether sampling unit i belongs or does not belong to category k . Each **category** is thus a **species** in the terminology of CANOCO and each individual a sample. For the analysis by CANOCO the categories of different nominal variables must be assigned different numbers. One can number them consecutively from 1 to m with m the total number of categories. Alternatively, if the maximum number of categories per variable is less than 10, one can reserve the number 11-19 to the categories of nominal variable 1, the numbers 21-29 to the categories of nominal variable 2, etc. Nominal response data are best supplied to CANOCO in Cornell condensed format (sections 4.4.2). If one has 3 nominal variables one needs to specify 3 couplets per sampling unit (= individual). With nominal data, the species scores are category quantifications in the sense of Gifi (1990). For each nominal variable, the weighted mean of its category scores is equal to 0.

The theory of section 3.9.3 can be applied to nominal data. With such data, (3.30) models the probability that sample i belongs to category k . In applying multiple correspondence analysis conditional independence is assumed, i.e. the joint probability that a sample belongs to the categories k_1, k_2, k_3, \dots of nominal variables 1, 2, 3, ... is simply the product of each of the category probabilities given by (3.30). The logarithm of the joint probability can therefore be expressed as $\sum_l \phi_{il} + \eta_{ik(l)}$ where l indexes the nominal variables. By modeling $\eta_{ik(l)}$ by (3.37) and using the approach of Ter Braak (1985, 1988) we obtain an alternative derivation of multiple correspondence analysis. This approach shows that the category quantification can equally well be considered optima of response curves (Figure 3-3) with respect to the ordination axes. See Unimodal models pages 84-89.

In some applications a sampling unit may belong partly to one category and partly to another one. For such data, **fuzzy coding** has been proposed: for example, the sample is assigned the value 0.5 for both categories (see Greenacre, 1984: p. 159, *codage flou* in French). Obviously fuzzy coding is allowed in CANOCO. Percentage data are of this form (section 3.9.3).

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

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

3.11 Canonical Variates Analysis (CVA) and discriminant analysis

A canonical variates analysis (CVA), alias Fisher's linear discriminant analysis, can be obtained with CANOCO, because CCA is a generalization of CVA (Chessel et al. 1987, Lebreton et al. 1988; Unimodal Models: pp 175-177). Fisher's Iris data are analyzed as an example in section 8.4.3.

Suppose you want a CVA to see which linear combinations of environmental variables discriminate best between clusters of samples, e.g. obtained by a cluster analysis on species data. For this, specify the clusters as dummy variables in a file, for example CLUSTERS.DAT. This is perhaps most easily done in condensed format.

To obtain a CVA, use the following options in CANOCO:

1. CCA
2. CLUSTERS.DAT as species data
3. the environmental variables as environmental data
4. Hill's scaling with focus on inter-species distances (scaling -2)

In the solution file, the species scores are the cluster means in the CVA ordination diagram. Distances between cluster means (species points) represent Mahalanobis distances, as they should in CVA. These two properties are the result of Hill's scaling focusing on inter-species distances. The sample scores that are linear combinations of environmental variables are the individual points in the diagram.

The biplot scores for the environmental variables form with the species scores a biplot of the cluster means of each of the environmental variables and with the individual points a biplot of the environmental data (both are least-squares approximations).

The sample scores that are linear combinations of environmental variables are scaled so that the within-cluster variance equals 1 (in this variance the divisor is n and not $n-g$ with g the number of clusters). See equation (6.15) on page 159.

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

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

We close this subsection with some technical remarks. The eigenvalues reported by CANOCO are those of the eigenvalue equation:

$$(3.40) \quad (\mathbf{B} - \lambda \mathbf{T}) \mathbf{c} = \mathbf{0}$$

where λ is the eigenvalue, \mathbf{c} the vector of canonical coefficients (weights), \mathbf{B} the matrix of between-cluster sums of squares and products, and \mathbf{T} the matrix of total sums of squares and products. Most computer programs calculate the eigenvalue of

$$(3.41) \quad (\mathbf{B} - \theta \mathbf{W}) \mathbf{c} = \mathbf{0}$$

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

$$(3.42) \quad \theta = \lambda / (1 - \lambda)$$

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

$$(3.43) \quad F = [(n-g)/(g-1)] \theta$$

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

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

$$(3.44) \quad V = (\theta_1 + \theta_2) / (\text{sum of all } \theta\text{'s})$$

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

$$(3.45) \quad C = (\lambda_1 + \lambda_2) / (\text{sum of all } \lambda\text{'s})$$

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

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

$$(3.46) \quad \text{lower bound for } V = (\theta_1 + \theta_2)/(a+b)$$

$$(3.47) \quad \text{upper bound for } V = (\theta_1 + \theta_2)/(a+d)$$

where $a = \theta_1 + \theta_2 + \theta_3 + \theta_4$, $b = (\text{trace} - a)/(1 - \lambda_4)$, $d = \text{trace} - a$, with trace the sum of all λ 's, reported by CANOCO and θ is calculated from λ by formula (3.42). With six clusters the actual V is precisely equal to the lower bound given by (3.46).

3.12 Principal coordinates analysis

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

The easiest way to obtain a principal coordinate analysis (PCO) with the Canoco for Window package is by using the program PrCoord. See Chapter 9 of this manual for details. Chapter 9 also describes how to obtain the constrained form of PCO, called distance-based RDA by Legendre & Anderson (1999).

To stress the relation between PCO and PCA, the remainder of this section shows how to obtain a PCO by using CANOCO only. A PCO can be obtained with CANOCO by taking as "species data" a square table of similarities or a square table with elements $-\delta_{ij}$ where δ_{ij} is the dissimilarity between sample i and sample j ($i = 1, \dots, n; j = 1, \dots, n$). In the "species data" there are thus as many species as there are samples, the j -th species corresponding to the j -th sample. To obtain a PCO, use the following options in CANOCO:

- principal components analysis (PCA)
- centered by samples
- centered by species
- symmetric scaling of ordination scores, do not post-transform species scores

If one has a data file with the values of δ_{ij} , one can use the piecewise linear transformation of the console version of CANOCO to obtain the values of $-\delta_{ij}$ by specifying 0 0 followed on the next lines: by 100 -100 and -1 0 (assuming that all δ_{ij} are smaller than 100).

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

Unfortunately CANOCO cannot be used to obtain directly the solution of a constrained PCO. When choosing the RDA-option in the above setting, CANOCO solves the wrong eigenvalue equation. The correct eigenvalue equation is given in Ter Braak (1992). Partial PCO is not directly available either in CANOCO. See Chapter 9 for work-arounds using the program PrCoord.

The PrCoord program can be also used alternatively to produce an unconstrained PCO solution.

3.13 The stability of ordination axes

Tausch et al. (1995) observed that changing the order of species or samples in the input data file of the program DECORANA (Hill 1979) can sometimes cause relatively large changes in the sample scores on the ordination axes. Oksanen and Minchin (1997) showed that CANOCO 3.12 suffered from the same type of instability. They investigated the role of the convergence criteria in the power algorithm used to extract the ordination axes (see Step A8 of the iteration algorithm on page 137 of Unimodal Models). By comparison with another algorithm to extract

ordination axes, they showed that the use of more stringent convergence criteria gives results that are acceptably stable. In line with their proposals, CANOCO 4.5 uses a maximum number of iterations of 999 and a tolerance of 10^{-6} , which is in-between their strict and superstrict tolerance criteria. CANOCO issues a warning in the log window if these criteria could not be met. See the example in section 8.2.2.2. For some data sets instability is inevitable: if two eigenvalues are exactly equal, there is more than one solution to the eigenvalue equations, and these solutions form a plane. It is then the plane that is stable. Thus, if two eigenvalues are close in value, the extracted ordination axes are numerically unstable and there is little to be gained by more stringent convergence criteria. Such ordination axes should always be plotted together.

In DCA with detrending by segments, Oksanen and Minchin (1997) detected a bug in the subroutine SMOOTH that contributed to the instability. This bug has been fixed. Both line 7 and line 17 needed to be changed to

```
IF (AZ3 .LE. 0.0) ISTOP = 0
```

to make the subroutine order invariant. In addition a small change has been made in the subroutine SEGMNT. The line

```
IF (SQCORR .GT. 0.9999) SQCORR = 0.9999
```

has been changed to

```
IF (SQCORR .GT. 0.9999) GOTO 50
```

A problem in DCA is that convergence problems may go undetected. The eigenvalue problem solved for the first axis, for example, differs from the eigenvalue problem solved for the second axis (cf. ter Braak & de Jong 1998). The second eigenvalue in the eigenvalue problem for axis 1 may be close to the first eigenvalue, whereas the second DCA axis is much lower. Unstable axes in DCA may therefore go unnoticed. In other words, the reason why there is slow convergence in DCA is not obvious from the CANOCO output. The likely reason is close eigenvalues at a particular stage of the algorithm. The Shuffle software available from Jari Oksanen at WWW page <<http://cc.oulu.fi/~jarioksa/>> may help detecting such unstable eigenvalues.

4. Data input

The data files in Canoco for Windows must be in one of three strict formats. The three formats (full format, condensed format, and free format) are described in this chapter. The CANOCO formats can be used for any of the input data files (species, environmental, covariable, or supplementary environmental data files). A CANOCO solution file can also be used as input for a new analysis, but not as new Species Data (example section 8.2.2).

Data in spreadsheets or databases must first be converted to a CANOCO format. The utility programs CanoImp and WCanolmp can help you to import data from a spreadsheet, program CanoMerge can be used to merging data tables. The utility WinTran (Juggins, 1998) is very useful for converting ecological and palaeoecological data between Access, Excel, Lotus, Paradox, dBase, and CANOCO formats.

4.1 Importing from spreadsheets: CanoImp, WCanolmp

The CanoImp utility converts input data in TAB-separated format to CANOCO format. The utility exists in two forms, a Windows-based form and a console form.

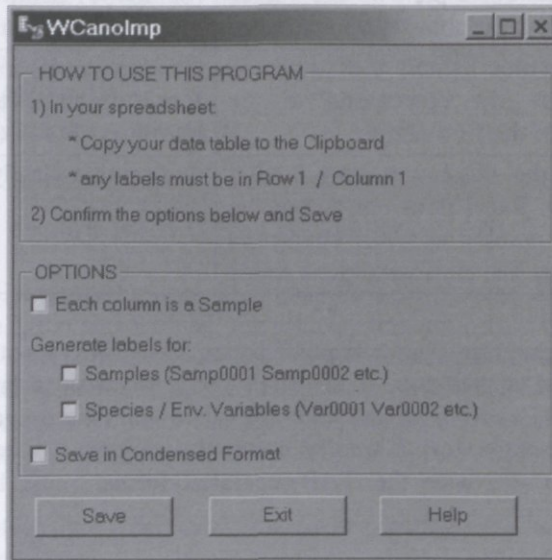


Figure 4-1 WCanolmp program window.

WCanolmp is the Windows-based form and its icon is available in the Canoco for Windows submenu of the **Start / Programs** menu. When launched, WCanolmp displays a window (Figure 4-1) which tells how to use the program. WCanolmp reads the input data from the Windows Clipboard when you click the **Save** button. So you need to place the data there first by a Copy command (**Ctrl-C**) in your spreadsheet (all Windows-based spreadsheet applications use the TAB-separated format for this). After you have done this and clicked the **Save** button, WCanolmp displays a file-selection dialog box to get the name of the output data file where the dataset has to be saved in the Canoco-compatible format, and also asks for a title. The WCanolmp window contains a range of options which are self-explanatory.

The CANOIMP.EXE program is the console form. The names of the input and the output files are specified on the command line, together with one or more optional command line parameters. The syntax of the CANOIMP.EXE command line is:

CANOIMP [options] <inp-file-name> <outp-file-name>

where the options are described in the Table 4.1, **inp-file-name** is the path to the input data file (which must be in TAB-separated format), and **outp-file-name** is the path of the data file to be created (in Canoco format). Table 4.1 also summarizes the correspondence of the command-line options of the CANOIMP.EXE program with the options of WCanolmp.

Table 4.1 Command-line options of CANOIMP.EXE compared with WCanolmp.

command-line option	its meaning	corresponding option in WCanolmp
-C	the output file in Condensed format	Save in Condensed format
-P	transpose the input data matrix	Each column is a Sample
-Q	work quietly, no messages	not available
-R	no sample names; the first fields in each row correspond to the values of the first variable	Generate labels for: Samples (Samp0001 Samp0002 etc.)
-S	no variables names; the values in the first row correspond to the values of the first sample	Generate labels for: Species / Env.Variables (Var0001 Var0002 etc.)
-T	ignore the missing trailing TABs; program Excel does not output the correct number of TAB characters when the line ends with blank fields	done automatically

As the Canolmp program uses already existing data files, most often created by your spreadsheet or database application, it is usually able to process larger data sets than the alternative form - the WCanolmp program. On the other hand, the program has no user-friendly interface and also some extra work is needed before the program can be used to transform your data sets, because the file with the TAB-separated format must be exported from your spreadsheet application.

4.1.1 Processing capacity

Both the console and Windows-based forms of Canolmp can process input data lines up to 80 000 characters in length. Note that the single data-line contains the textual representation of a single data row (being a single sample in the default case). Both forms have no fixed limit for the number of rows in the input data file. Both forms are limited by the availability of free memory, as they need to allocate arrays for the names of samples and the names of variables. Additionally, the Windows-based form of the program is restricted by the limitation imposed on the size of the data that can be passed through the Windows Clipboard. This size was not published by the Microsoft Corp., but it can be expected to depend on the amount of the virtual memory available on your system.

4.1.2 Properties of the output files

This section describes the output files of CanoImp and WCanoImp. Both programs produce identical output files (when given identical input files and options) except that the Windows-based form asks for the title line of the output file and uses it as the first line of the data file. CanoImp, on the other hand, always uses titles such as:

CANOCO full format export from mydata.txt by CanoImp

CANOCO condensed format export from mydata.txt by CanoImp

where mydata.txt is a shortened version of the input file path (first 32 characters only).

CanoImp restricts the characters that might appear in the labels of samples and variables: only the lowercase and uppercase ASCII letters, digits, colon (':'), dash ('-'), plus ('+'), asterisk ('*'), underscore ('_'), left and right parentheses ('()'), left and right square brackets ('[]'), and the space character (' ') are supported. Any other characters are changed to the dot character ('.'). Also, labels have no more than 8 characters.

If variable names are not in the spreadsheet section being exported, CanoImp generates labels of the form **VarIII**, where **III** is the sequential number running from **0001** to the number of variables in the data set. Similarly, if the sample names are missing, the generated labels use either the form **SampIII**, if the number of samples is lower than 10 000, or the form **SamIIIII**, if the number of samples is greater than 9999.

CanoImp determines the range of values contained in the data table and adjusts the formatting of the output accordingly. Consequently, each variable gets the same output field specification (in the terms of its total width and number of decimal digits) to simplify the format line specification in the output file.

For technical details see the Appendix C.

4.2 Merging data tables with CanoMerge program

If you use the WCanoImp program to import data containing many samples and many variables (species) from Microsoft Excel®, the primary limitation is in the maximum number of 256 columns per single Excel sheet. If both the number of samples and the number of variables is larger than 255, the data table must be split over two or more Excel sheets. To create a single data file in Canoco format from such multiple sheets, the data must be exported using the WCanoImp program (as described in the previous section) and merged using the CanoMerge program.

To start the CanoMerge program, select the corresponding item in the **Start / Programs / Canoco for Windows** submenu.

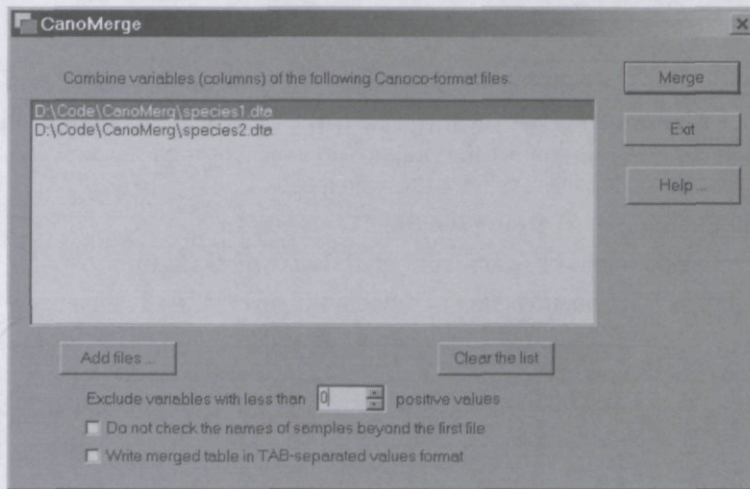


Figure 4-2 CanoMerge program window

The largest part of the program window is taken by the list where you must place the names of all input data files which you want to be merged together. To create the list, click the **Add files** button. This action opens an **Open** dialog box where you can navigate to a particular file folder and select one or more input files. If the input files are spread across several folders, you must use the **Add files** button several times. Use the **Clear the list** button to remove all input file references from the list. Use the **Merge** button to perform the actual merging. The merging is concluded with a dialog box providing information about the number of tables which were merged and the resulting number of variables.

All the files selected for merging must have the same number of rows (samples), because they are merged column-wise. This strategy is based on the fact that there is hardly any limitation on the number of rows (samples) in an Excel spreadsheet. A particular row in the individual input files must correspond to the same sample entity. CanoMerge checks this correspondence using the sample names, unless you select the option **Do not check the names of samples beyond the first file**. If you merge data files which were not produced by the WCanolmp program, you must make sure not only that there is the same total number of samples, but also that the samples identification numbers are contiguous, running from 1 to N, where N is the total number of samples. This condition is always met in WCanolmp produced files. The individual input data tables, as well as the resulting merged table, cannot have more than 25000 samples and 5000 variables.

The data columns from the individual source files are added to the merged table in the order, in which the input files appear in the list. You can change this ordering by selecting a single line of the list and dragging it to a new position. Note that each file name can be dragged to almost any position, except the end of the list.

CanoMerge allows you to filter the set of input species (variables), merging only the species with number of occurrences (non-zero values) larger or equal to a pre-specified limit. The default value 0 (which is passed by all the columns) can be increased in the CanoMerge window.

☞ **CanoMerge can also be used for reducing the number of species in a single data file. To do so, simply specify this input data file as the only one in the list and increase the value in the *Exclude variables with less than option*, as required.**

When the option at the bottom of CanoMerge program window is checked (**Write merged table in TAB-separated values format**), CanoMerge produces the resulting file in a text

(ASCII) format where the individual columns are separated by TAB characters. This can be useful both for exporting your datasets into formats accepted by other statistical packages and for transforming Canoco-formatted sets into file format readily accepted by the spreadsheet software.

CanMerge can be used for transforming Canoco-formatted datasets into TAB-separated text format, acceptable by Excel™ program.

4.3 Linking samples in different data files

Important: Species data, environmental data, and covariable data are commonly stored in different files. CANOCO determines which samples in different files correspond to the same physical sampling unit on the basis of the sample identification number (not the sample name). See the example in section 8.2.1. The sample names in the species data file serve to label the samples in the project setup and in the output. The sample names in the environmental data and covariable data are used merely to check whether the sample numbers in different files have the same name. If differences in names are detected, Canoco for Windows issues an error message and asks whether differences in names are permitted. If you answer yes, the program disregards the differences, except for a reminder warning in the log-window. If you answer that no differences in names are allowed, you probably need to cancel the project-setup and repair the error first.

4.4 Native CANOCO formats

If you use a utility program such as WCanolmp to produce input data files for Canoco for Windows, you may wish to skip this section. This section is useful when you need to prepare files with the correct data format yourself and in troubleshooting.

4.4.1 Full format

In full format, each sample is represented by an identification number followed by the values of all variables for that particular sample, in a fixed order and a fixed format. Each sample starts on a new line; its data may occupy one or more lines. Table 4.2 shows a small example with three samples and eleven variables, in which each sample takes two lines.

The first four lines of a full format file must look like this:

Line 1 is a title. The first part of the title is reproduced in the output to remind the user which data were used in the analysis.

Line 2 contains a FORTRAN format which specifies how the data are stored for a sample. The FORTRAN format in Table 4.2 is:

```
(I6, (T10, 6F7.0))
```

- “I6” means that the sample identification number is in the first 6 positions of the first line for a sample (right-justified, i.e. with the last digit in position 6).
- “T10” means that the next value is to be read from position 10 onwards.
- “6F7.0” reads six values, each of seven positions. Note that each value contains a decimal point, but this is not a necessity. Whole numbers are also allowed. Note also that some values are negative.

- The brackets around “T10,6F7.0” indicate that the subsequent values are read from the next line, again starting at position 10.

Line 3 contains the number of variables in the data file (the position of the number is arbitrary). In Table 4.2 the number of variables is 11. Because the FORTRAN format on line 2 specifies that 6 (or less) values are being read per line, each sample takes two lines in the data: one line with 6 values and one line with 5 values

Line 4 is the beginning of the data. The data end with a notional sample with identification number 0, which occupies as many lines as a normal sample. The data values of the notional sample are arbitrary. For clarity the values are 0.0 in Table 4.2.

After the notional sample 0, the code names of the eleven variables follow, 10 per line, and then the code names of the three samples. There are at most 10 code names per line. Each code name takes 8 positions. The names of the variables in Table 4.2 are thus Column01, Column02, . . . , Column11 and the names of the samples are Row 0001, Row 0002, Row 0003. If the last line would have been

```
Row 0003Row 0002Row 0001
```

then the samples with identification numbers 1 and 3 would receive the code names Row 0003 and Row 0001, respectively. The linkage of numbers to names is thus by the order in the sequence of names, each 8 positions representing one variable or sample.

Table 4.2 Full format data file with 3 samples and 11 variables (species or environmental variables).

Example data in CANOCO full format: 3 Samples and 11 Variables
(I6, (T10, 6F7.0))

```

11
 1 | 42.0  21.0  12.0  67.0  32.0 -12.0
   | -0.1   3.0  -9.0   5.0   8.0
 2 | 52.0  27.0  15.0  80.0   9.0  40.0
   |  0.5   8.0   8.0  -7.0   6.0
 3 | 70.0  18.0  17.0  21.0   .0  17.0
   |  0.9   2.0  11.0  11.0   2.0
 0 |  0.0   0.0   0.0   0.0   0.0   0.0
   |  0.0   0.0   0.0   0.0   0.0
Column01Column02Column03Column04Column05Column06Column07Column08Column09Column10
Column11
Row 0001Row 0002Row 0003

```

The FORTRAN format in Table 4.2 says that six values are read from the second line (and further lines, if present) of a sample, whereas in Table 4.2 the second line of a sample contains five values only. This is not a problem, because we specified on line 2 that only 11 values should be read for each sample. In the example, the data values of a sample would easily fit on a single line, but in practice you will often need more than one data line per sample, because the maximum allowed length of each line of the data file is 127 positions.

Any data table (with any number of samples and variables) in which the values in the columns take a fixed number of positions can be specified by a FORTRAN format similar to that used Table 4.2. Simply adapt the number of values per line and the number of positions per value to your own needs. The format can also be specified as (I6, 3X, 6F7.0 / (9X, 6F7.0)), in which

- “3X” means skip the next three positions,
- “/” means go to the next line.

This way of writing the format has the advantage that it allows you to separately specify the format for the first line and the format for the further lines for the data of a sample. For example, if the data on the second line for a sample do not start in position 10, you must change the "9X" to the appropriate number.

As a second example, Table 4.3 shows the environmental data of the extended Dune Meadow data (Table 16.2 on page 477) in full format. Five "environmental" variables were recorded at each site, two of which are nominal. The first column of numbers in Table 4.3 lists the sample identification numbers. The next three columns list the data on the three quantitative variables A1, Moisture, and Manure. The data on the nominal variable agricultural use are given in the next three columns. Each of its three classes (Hayfield, Haypasture, and Pasture) has a column of 0/1 values in Table 4.3. For example, sample number 1 is a haypasture because the corresponding column has the value 1. The data on the nominal variable management regime, with the four classes standard farming (SF), bio-dynamical farming (BF), hobby farming (HF), and nature management (NM) are given in the last four columns. So, for CANOCO there are, in total, 10 variables, the names of which you can recognize at the bottom of Table 4.3. Variable 1 is A1, variable 2 is MOISTURE, ..., and variable 10 is NM (for Nature Management). The values of the 10 variables of sample 3 are given on line 6. In sample 3 the A1 horizon was 4.3 mm, its moisture content was scored the value 2, its manure score was 4. Sample 3 is a haypasture as the haypasture-column has a 1. Sample 3 is a standard farm: its SF-value is 1 whereas the other management classes have the value 0 in sample 3. The FORTRAN format in Table 4.3 is:

(I5, F5.0, 1X, 2F3.0, 3X, 3F2.0, 3X, 4F2.0)

- "I5" means that the sample number is in the first 5 positions of the first line for a sample (with the last digit in position 5).
- "F5.0" reads a value from the next five positions (A1).
- "1X" means that position 11 is skipped.
- "2F3.0" reads two values of three positions each (Moisture and Manure).
- "3X, 3F2.0" skips the next three positions and reads three values of two positions each (Hayfield, Haypasture, and Pasture).
- "3X,4F2.0" skips the next three positions and reads four values of two positions each (SF, BF, HF, NM).

In Table 4.3 the sample identification numbers are increasing, but not consecutive: the sample numbers 18-27 are missing. The sample with identification number 1 has name "Sample 1". The sample with identification number 30 must have its name on positions 73-80 of the third line of sample names. Its name is thus "Sample20". Note that the samples with name "SupplSAM" has sample identification number 20, but does not occur in these data. The sample does occur in the corresponding species data file. The data values for A1 contain a decimal point, but the other values do not.

The requirements of the full format data file are:

- Each sample starts on a new line and is represented by an identification number followed by the values of all variables for that particular sample, in a fixed order and a fixed format. Each sample has the same format.
- The sample identification numbers are increasing but do not need to be consecutive.
- The number of variables is specified on line 3. It is also allowed to specify the number of variables in the positions 69-70 of line 2 (or, if it is a number of one digit, in column 70). In this case line 3 is the beginning of the data.
- The maximum allowed length of each line of the data file is 127 positions.

- The layout of the data values of each sample is specified by a FORTRAN format on line 2. It must be within the first 80 positions.
- The identification number of a sample must be read with an I-format. In the example of Table 4.2 this is "16" (line 2). It should be the first item read.
- The data values must be read with one or more F-formats, even the values are whole numbers. In the example of Table 4.2 this is "6F7.0" (line 2).
- The T-format item must not be used to jump back in a line to read the data values in another order than in the data file.
- All numbers and values must be separated by at least one space. No other characters are allowed between values. Special characters like tabs should be absent. In the data section, the only allowed characters are digits (0-9), the period (.), the minus sign (-) and the space.
- Missing values are not allowed. For missing values one may wish to insert a best possible guess, perhaps the mean value of the corresponding variable.
- The data values end with a notional sample with identification number 0 followed by the same number of data values in the same layout as for a normal sample.
- The names of the variables start on a new line after the notional sample 0. The names are listed in lines of 80 positions, with 10 names per line, each name taking 8 positions. Trailing blanks are allowed unless they extend the line beyond the maximum line length of 127 positions.
- The names of the samples start on a new line after those of the species. The names are listed in lines of 80 positions, with 10 names per line, each name taking 8 positions. Trailing blanks are allowed, unless they extend the line beyond the maximum line length of 127 positions.

Table 4.3 The environmental data of the Dune meadow data in full format. The file is named 'DUNEENV.DTA'.

ENVIRONMENTAL DATA IN FULL FORMAT - DUNE MEADOW DATA
(I5, F5.0, 1X, 2F3.0, 3X, 3F2.0, 3X, 4F2.0)

```

10
1 2.8 1 4 0 1 0 1 0 0 0
2 3.5 1 2 0 1 0 0 1 0 0
3 4.3 2 4 0 1 0 1 0 0 0
4 4.2 2 4 0 1 0 1 0 0 0
5 6.3 1 2 1 0 0 0 0 1 0
6 4.3 1 2 0 1 0 0 0 1 0
7 2.8 1 3 0 0 1 0 0 1 0
8 4.2 5 3 0 0 1 0 0 1 0
9 3.7 4 1 1 0 0 0 0 1 0
10 3.3 2 1 1 0 0 0 1 0 0
11 3.5 1 1 0 0 1 0 1 0 0
12 5.8 4 2 0 1 0 1 0 0 0
13 6.0 5 3 0 1 0 1 0 0 0
14 9.3 5 0 0 0 1 0 0 0 1
15 11.5 5 0 0 1 0 0 0 0 1
16 5.7 5 3 0 0 1 1 0 0 0
17 4.0 2 0 1 0 0 0 0 0 1
28 4.6 1 0 1 0 0 0 0 0 1
29 3.7 5 0 1 0 0 0 0 0 1
30 3.5 5 0 1 0 0 0 0 0 1
0 0.0 0 0 0 0 0 0 0 0
A1 MoistureManure HayfieldHaypastuPasture SF BF HF NM
Sample 1Sample 2Sample 3Sample 4Sample 5Sample 6Sample 7Sample 8Sample 9Sample10
Sample11Sample12Sample13Sample14Sample15Sample16Sample17 Supp1SAM
Duplic17 Sample18Sample19Sample20

```

4.4.2 Condensed format

A data file in condensed format differs from a full format data file (section 4.4.1) in that the data values that are zero are not stored. In a condensed data file, samples and species are both indicated by an identification number. To identify which variable a particular non-zero value belongs to, the value is preceded by the identification number of the variable. Table 4.4 shows the example data of Table 4.2 in condensed format. Each data line begins with a sample identification number and continues with a number of “couplets”, each consisting of a species identification number and a data value. For example, line 7 of Table 4.4 begins with the number 2, which means that this line gives the data of the sample with identification number 2 (shortly sample 2). This number is followed by the couplet “1 52.0”, which says that variable with identification number 1 has value 52.0 in sample 2. The next couplet “2 27.0 ” says that species 2 has value 27.0, etc. There are five couplets on this line. The next two lines also begin with a 2 indicating that these lines also contain data values for sample 2. The last line that begins with a 2 (line 9) has only one couplet “11 6.0” as an example that the number of couplets may vary among lines. The data of the samples 1 and 2 are on three lines each. The data of sample 3 just take two lines: there is no couplet for variable 5, because its value is 0 in sample 3. The identification numbers of the variables are in increasing order within samples in this example, but this is not necessary. After sample 3 there is a line for the notional sample 0, which indicates the end of the data. This line does not need to have any couplets. Thereafter are the code names for the eleven variables (10 per line), followed by the code names for the samples.

The first four lines of a condensed format file must look like this:

Line 1 is a title. The first part of the title is reproduced in the output to remind the user which data were used in the analysis.

Line 2 contains a FORTRAN format which specifies how the data are stored on each data line. The FORTRAN format in Table 4.4 is:

```
(I6,5(I7,F6.0))
```

- “I6” means that the sample identification number is in the first 6 positions of each line (right-justified, i.e. with the last digit in position 6).
- “5(I7,F6.0)” means that there are a maximum of five couplets on a line, each with 7 positions for the species identification number (I7) and 6 positions for its data value (F6.0). Note that each data value contains a decimal point, but this is not a necessity. Whole numbers are also allowed. Note also that some values are negative.

Line 3 contains the maximum number of couplets on a line (the position of the number is arbitrary). In Table 4.4 the maximum number of couplets per line is 5.

Line 4 is the beginning of the data. The data ends with a notional sample with identification number 0, which occupies a single line. No couplets are required for the notional sample.

After the notional sample 0, the code names of the eleven variables follow, 10 per line, and then the code names of the three samples. As in full format, there are at most 10 code names per line. Each code name takes 8 positions. The names of the variables in Table 4.4 are thus Column01, Column02, .. , Column11 and the names of the samples are Row 0001, Row 0002, Row 0003.

Table 4.4 Condensed format data with 3 samples and 11 variables. Same data as Table 4.2.

Example data in CANOCO condensed format: 3 Samples and 11 Variables
(I6,5(I7,F6.0))

```

5
1 1 42.0 2 21.0 3 12.0 4 67.0 5 32.0
1 6 -12.0 7 -.1 8 3.0 9 -9.0 10 5.0
1 11 8.0
2 1 52.0 2 27.0 3 15.0 4 80.0 5 9.0
2 6 40.0 7 .5 8 8.0 9 8.0 10 -7.0
2 11 6.0
3 1 70.0 2 18.0 3 17.0 4 21.0 6 17.0
3 7 .9 8 2.0 9 11.0 10 11.0 11 2.0
0
Column01Column02Column03Column04Column05Column06Column07Column08Column09Column10
Column11
Row 0001Row 0002Row 0003

```

Table 4.5 The species data of the Dune Meadow data in condensed format. The file is named 'DUNE_SPE.DTA'.

SPECIES - DUNE MEADOW DATA (M. BATTERINK AND G. WIJFFELS, 1983)
(I5,9(I5,F2.0))

```

9
1 1 1 11 4 17 7 19 4 20 2 32 3
2 1 3 4 2 6 3 7 4 11 4 16 5 17 5 19 4 20 7
2 27 5 32 3
3 2 4 4 7 6 2 11 4 16 2 17 6 19 5 20 6 27 2
3 29 2 32 6
4 2 8 4 2 6 2 7 3 9 2 11 4 16 2 17 5 19 4
4 32 4
4 20 5 24 5 27 1 29 2
5 1 2 5 4 6 2 7 2 11 4 16 3 17 2 18 5 19 2
5 33 2 20 6 23 5 26 2 27 2 29 2
6 1 2 5 3 16 3 17 6 18 5 19 3 20 4 23 6 26 5
6 27 5 29 6 33 3
7 1 2 5 2 7 2 15 2 16 3 17 6 18 5 19 4 20 5
7 23 3 26 2 27 2 29 2 32 2 33 2
8 2 4 4 5 10 4 14 4 16 3 17 4 19 4 20 4 22 2
8 24 2 27 2 29 2
9 2 3 4 3 11 6 14 4 15 4 16 2 17 2 19 4 20 5
9 23 2 24 2 27 3 29 2 31 1 32 2 33 2
10 1 4 5 4 6 2 7 4 16 3 17 6 18 3 19 4 20 4
10 27 6 28 1 29 2 32 3
11 13 2 16 5 17 7 18 3 19 4 24 2 27 3 28 2 29 4
11 32 2
12 2 4 4 8 15 4 16 2 20 4 23 2 24 4 27 3 29 4
13 2 5 4 5 8 1 15 3 16 2 19 2 20 9 22 2 24 2
13 27 2 32 3
14 2 4 10 4 16 2 21 2 22 2 27 6 30 4 33 1
15 2 4 10 5 14 3 16 2 21 2 22 2 27 1 29 4 33 1
16 2 7 4 4 10 8 14 3 20 2 22 2 29 4 30 3
17 1 2 3 2 5 4 13 2 16 2 18 2 19 1
20 6 5 19 4 23 3
21 1 2 3 2 5 4 13 2 16 2 18 2 19 1
28 6 2 16 5 17 2 18 3 19 3 25 3 27 2 28 1 29 6
28 31 2 32 4
29 3 3 5 4 12 2 13 5 16 6 24 3 25 3 27 2 29 3
29 31 1
30 2 5 10 4 14 4 16 2 22 4 25 5 29 4 30 3
0
Ach mil Agr sto Air pra Alo gen Ant odo Bel per Bro hor Che alb Cir arv Ele pal
Ely rep Emp nig Hyp rad Jun art Jun buf Leo aut Lol per Pla lan Poa pra Poa tri
Pot pal Ran fla Rum ace Sag pro Sal rep Tri pra Tri rep Vic lat Bra rut Cal cus
Hip rha Poa ann Ran acr
Sample 1Sample 2Sample 3Sample 4Sample 5Sample 6Sample 7Sample 8Sample 9Sample10
Sample11Sample12Sample13Sample14Sample15Sample16Sample17 SupplSAM
Duplic17 Sample18Sample19Sample20

```

The condensed format is a compact way of representing species data matrices with many species but where the number of species per sample is relatively small. Table 4.5 shows the species matrix of the extended Dune Meadow data (Table 16.1) in condensed format with 9 couplets per line. Notice that the samples are rows in Table 4.5, whereas they are columns in Table 16.1. The code names of the samples reflect the original numbering. The first sample contains six species (with code names Ach mil, Ely rep, Lol per, Poa pra, Poa tri, and Poa ann) with abundance values 1, 4, 7, 4, 2, and 3, respectively. The other species are absent and thus have abundance value 0.

The condensed format is also a handy way to represent nominal data. In condensed format, each nominal variable takes one couplet, irrespective of the number of classes. This is illustrated in Table 4.6, which is the extended Dune meadow environmental data in the condensed format. The data on the nominal variables Use and Management regime (in the last two couplets in Table 16.2) are represented in the last two couplets of each data line in Table 4.6. For example, sample 1 has the value 1 for variables 5 and 7 which represent the class Haypasture of the variable Use and the class Standard Farm (SF) of the variable Management regime.

Table 4.6 The environmental data of the Dune meadow data in condensed format.

The variables are numbered as follows: 1 = Thickness of A1 horizon; 2 = moisture; 3 = quantity of manure; 4 = hayfield; 5 = haypasture; 6 = pasture; 7 = Standard Farm; 8 = Biodynamic Farm; 9 = Hobby Farm; 10 = Nature Management. The sample numbers and names are as in Table 4.5, except that samples 20 and 21 are missing. For explanation see text.

ENVIRONMENTAL DATA IN CONDENSED FORMAT - DUNE MEADOW DATA

(I3, I2, F5.0, 3 (I3, F2.0), I4, F2.0)

```

5
1 1 2.8 2 1 3 4 5 1 7 1
2 1 3.5 2 1 3 2 5 1 8 1
3 1 4.3 2 2 3 4 5 1 7 1
4 1 4.2 2 2 3 4 5 1 7 1
5 1 6.3 2 1 3 2 4 1 9 1
6 1 4.3 2 1 3 2 5 1 9 1
7 1 2.8 2 1 3 3 6 1 9 1
8 1 4.2 2 5 3 3 6 1 9 1
9 1 3.7 2 4 3 1 4 1 9 1
10 1 3.3 2 2 3 1 4 1 8 1
11 1 3.5 2 1 3 1 6 1 8 1
12 1 5.8 2 4 3 2 5 1 7 1
13 1 6.0 2 5 3 3 5 1 7 1
14 1 9.3 2 5 3 0 6 1 10 1
15 1 11.5 2 5 3 0 5 1 10 1
16 1 5.7 2 5 3 3 6 1 7 1
17 1 4.0 2 2 3 0 4 1 10 1
28 1 4.6 2 1 3 0 4 1 10 1
29 1 3.7 2 5 3 0 4 1 10 1
30 1 3.5 2 5 3 0 4 1 10 1
0
A1      MoistureManure HayfieldHaypastuPasture SF      BF      HF      NM
Sample 1Sample 2Sample 3Sample 4Sample 5Sample 6Sample 7Sample 8Sample 9Sample10
Sample11Sample12Sample13Sample14Sample15Sample16Sample17      SupplSAM
Duplic17      Sample18Sample19Sample20

```

The condensed format has no means to deal with missing data, because a variable that is not listed for a sample automatically receives the value zero. For the results of the ordination, it does not matter whether the data file is presented in full format or in condensed format.

The requirements of the condensed format data file are:

- Each line starts with a sample identification number followed by a number of couplets, each consisting of a variable identification number and a data value. Each line has the same format.
- The maximum number of couplets on a line is specified on line 3. It is also allowed to specify the number of couplets in the positions 69-70 of line 2 (or, if it is a number of one digit, in column 70). In this case line 3 is the beginning of the data.
- The maximum allowed length of each line of the data file is 127 positions.
- The layout of the data values of each sample is specified by a FORTRAN format on line 2. It must be within the first 80 positions.
- The identification numbers of samples and variables must be read with an I-format. In the example of Table 4.4 (line 2) this is "I6" for the sample number and "I7" for each variable number. The FORTRAN-format in a condensed format file thus contains at least two I's.
- The data values must be read with one or more F-formats, even the values are whole numbers. In the example of Table 4.4 this is "F6.0" (line 2).
- The number of couplets specified on the format line must be (greater than or) equal to the maximum number of couplets specified on line 3. In Table 4.4 this requirement is met by the 5 in front of "(I7,F6.0)". In Table 4.6 the FORTRAN-format also specifies 5 couplets, which is the number on line 3.
- The next-line format item "/" should not be used. The T-format item is not needed.
- All numbers and values must be separated by at least one space. No other characters are allowed between values. Special characters like tabs should be absent. In the data section, the only allowed characters are digits (0-9), the period (.), the minus sign (-) and the space.
- The sample identification numbers are non-decreasing. They do not need to be consecutive.
- The order of variable identification numbers within a sample is arbitrary.
- If, for a particular sample, there is no couplet that assigns a value to a particular variable, the variable is assigned the value 0.
- Missing values are not allowed. For missing values one may wish to insert a best possible guess, perhaps the mean value of the corresponding variable.
- The data values end with a notional sample with identification number 0 without couplets.
- The names of the variables start on a new line after the notional sample 0. The names are listed in lines of 80 positions, with 10 names per line, each name taking 8 positions. Trailing blanks are allowed unless they extend the line beyond the maximum line length of 127 positions.
- The names of the samples start on a new line after those of the species. The names are listed in lines of 80 positions, with 10 names per line, each name taking 8 positions. Trailing blanks are allowed unless they extend the line beyond the maximum line length of 127 positions.

If, for a particular variable, a variable occurs in more than one couplets, the last assigned value is used, except when the file is used as a species data file in CANOCO. When used as a species data file, the sum of the assigned values is used.

4.4.3 Free format

Free format is an easy form of full format. In free format, the data values of all variables are given, sample after sample, in a fixed order, but without the need to adhere to fixed column positions for the variables or to add sample identification numbers. Each sample starts on a new

line and the first value on that line is that of the first variable. Table 4.7 shows an example with three samples and eleven variables with the same data as Table 4.2.

Table 4.7 Free format data file with 3 samples and 11 variables (species or environmental variables).

```

Example data in CANOCO free format: 3 Samples and 11 Variables
free
 11 3
42.0  21.0  12.0  67.0  32.0 -12.0
      -0.1   3.0  -9.0   5.0   8.0
52.0  27.0  15.0  80.0   9.0  40.0
      0.5   8.0   8.0  -7.0   6.0
70.0  18.0  17.0  21.0   .0  17.0
      0.9   2.0  11.0  11.0   2.0
Column01Column02Column03Column04Column05Column06Column07Column08Column09Column10
Column11
Row 0001Row 0002Row 0003

```

The first four lines of a free format file must look like this:

Line 1 is a title. The first part of the title is reproduced in the output to remind the user which data were used in the analysis.

Line 2 contains the keyword “free” or “FREE” starting in position 1 (or alternatively in position 2, 3, ... or 7).

Line 3 contains the number of variables and the number of samples in the data file (the position of the numbers is arbitrary) . In Table 4.7 the number of variables is 11 and the number of samples is 3.

Line 4 is the beginning of the data. Each sample starts on a new line. The data of a sample may cover as many lines as needed. In Table 4.7 each sample takes two lines, but, in general, the number of lines per sample may vary. The data do not end with a notional sample 0.

After the data, the code names of the eleven variables follow, 10 per line, and then the code names of the three samples. There are at most 10 code names per line. Each code name takes 8 positions. The names of the variables in Table 4.7 are the same as those in Table 4.2. The code names are optional.

Table 4.8 shows the same data in another layout.

If you want to use free format but your file contains a sample identification number, you must count the identification number as an extra variable, add an extra name to the names of the variables and delete that variable in the project set-up for the analysis.

The most important additional requirements of the free format data file are:

- The maximum allowed length of each line of the data file is 127 positions.
- All numbers and values must be separated by at least one space. No other characters are allowed between values. Special characters like tabs should be absent. In the data section, the only allowed characters are digits (0-9), the period (.), the minus sign (-) and the space.
- Missing values are not allowed. For missing values one may wish to insert a best possible guess, perhaps the mean value of the corresponding variable.
- There should be no blank lines between the last data value and the beginning of the optional code names for variables.

Table 4.8 Another data file in free format (same data as Table 4.7).

Example data in CANOCO free format: 3 Samples and 11 Variables

```
FREE
11      3
42.0  21.0  12.0  67.0  32.0  -12.0  -0.1  3.0  -9.0  5.0  8.0
52.0  27.0  15.0  80.0  9.0  40.0  0.5
8.0   8.0  -7.0  6.0
70.0  18.0  17.0  21.0  .0  17.0  0.9  2.0  11.0  11.0  2.0
Column01Column02Column03Column04Column05Column06Column07Column08Column09Column10
Column11
Row 0001Row 0002Row 0003
```

5. Project setup and analysis in Canoco for Windows

5.1 Introduction

This chapter describes Canoco for Windows and is best read after the Getting Started chapter.

The first step in an ordination analysis is to specify which data you want to analyze and which ordination method you want to apply. These details together form a “project”. The project is the central unit of action in Canoco for Windows. The second step is to carry out the ordination analysis, i.e. to analyze the data by the computational method specified in the project. The third step is to create ordination diagrams. These three steps (specification, analysis, and plotting) are separated in Canoco for Windows. CanoDraw for Windows, which is used to accomplish the third step maintains its own project files. CanoDraw project files are based on individual Canoco project files, when they are created.

For the first step (to specify a project) Canoco for Windows has a **Project Setup Wizard**, which starts automatically when you create a new project via the **File** submenu or by clicking the **New project** button on the toolbar. The Project Setup Wizard guides you through all available choices for the ordination analysis. When you finish the wizard sequence, the project is completed and can be saved to a file (the Canoco project file with extension CON) so that it can be re-opened and modified on a later occasion. This ends the first step.

The second step is to analyze the data by the ordination method specified in the project and the third step is to plot the data using the program CanoDraw 4.0. These two steps are performed by clicking buttons on the Project View, which is a window with a summary of the project and with buttons for common actions, or alternatively, via the **Project** menu of Canoco for Windows. The actions that are carried out on a project are logged in the log-window of the project, together with some summary results of the analyses. The ordination scores of an ordination analysis are stored in a “solution file”, the name of which you must specify in the Project Setup Wizard. The ordination diagrams can be stored, modified, printed, or exported to a graph file using one of the popular formats (BMP, WMF, AI, PNG) - see section 12.1.

You can also open a project that you saved in an earlier run of Canoco for Windows. This is done via the menu item **File | Open**. This brings you in the project view, from which you can choose

- to modify the specifications of the project by clicking the **Options...** button,
- to carry out the ordination analysis by clicking the **Analyze...** button
- to plot ordination diagrams of the analysis by clicking the **CanoDraw...** button

In summary, a project consists of all what is needed to carry out and modify an ordination analysis in Canoco for Windows. A project can be defined from scratch or modified from an existing project. In both cases the Project Setup Wizard helps you to define or modify the ordination method.

Most of this chapter is also available online. Help is accessible via the **Help** submenu on main menu, but also via **Help** buttons on the project view and on each page of the wizard sequence. The latter is often more useful as the Help is directly on the choices you must make to define a good ordination method for your data. Short context sensitive help is also available via the question mark button on each page.

5.1.1 How to use Canoco for Windows

The first step after launching Canoco for Windows is to create a new project. This is done by clicking the **New project** button on the toolbar or by selecting the menu item **File | New project** (or using the **Ctrl-N** keyboard short-cut) in the Canoco for Windows workspace (section 5.1.2). This starts the Project Setup Wizard which asks you to specify the data for analysis and the ordination method to apply. When completed, all details that you specified (the project settings) are saved in a file (the CANOCO project file) before the second step, the actual ordination analysis, can be carried out. A CANOCO project file is a text-only file with an extension CON. After the CANOCO project file has been saved you are back in the CANOCO workspace, which now contains two windows (section 5.1.3):

- the Project View with a summary of the project settings and buttons for common actions and
- the Log View which logs the creation dates of the project (and later on the results of the analyses)

The Log View may hardly be visible because the project view is the active window after leaving the project setup wizard. You may wish to resize and rearrange each window. The actual ordination is carried out after clicking the **Analyze...** button on the Project View or selecting the menu item **Project | Analyze (Ctrl-A)**. A box labeled "RUNNING CANOCO" reports on the progress of the analysis. When the box disappears, the analysis is completed and you may consult the output of the analysis in the log-window. The ordination scores are stored in the solution file that you specified in the Project Setup. To make ordination diagrams, click the **CanoDraw...** button on project-view or the menu item **Project | Run CanoDraw (Ctrl-C)**. You can return to Canoco for Windows by clicking Exit to DOS in CanoDraw. If you wish to change an ordination option, start the project setup wizard again by clicking the **Options...** button on the project-view or use the menu item **Project | Setup wizard (Ctrl-S)**. After completing the wizard sequence, you may wish to save the changed project under a new name so as to retain the original project. For this, use the **Save as** toolbar button or the menu item **File | Save as...** You are asked to specify a name for the new Canoco project file. A saved project can be opened on a later occasion by clicking the **Open** button on the toolbar or the menu item **File | Open (Ctrl-O)**. To close a project, click the menu item **File | Close** or the close button on the project view.

5.1.2 The Canoco for Windows workspace on startup

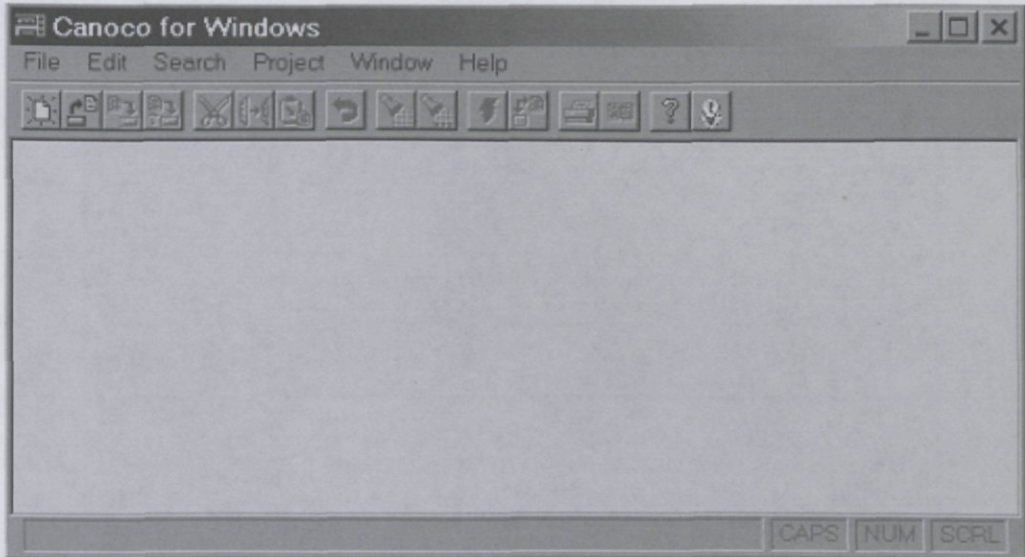


Figure 5-1 Canoco for Windows empty workspace.

Figure 5-1 shows the Canoco for Windows workspace after you close the Tip of Day window. Below the title bar is the main menu and the toolbar row with buttons representing shortcuts to frequently used menu commands. The white space is not for typing; it holds windows with information on projects that are created or opened. The toolbar can be positioned on any side of the Canoco for Windows workspace or it can be completely undocked and floating outside the Canoco for Windows main window. If you position the mouse pointer over one of the toolbar buttons (for example the one on the far left), you can see the tooltip text “New project” and a more detailed description of the command in the status line at the bottom of the Canoco for Windows workspace.

You can now create a new project or open an existing project either via the **File** menu or by clicking one of the first two toolbar buttons.

5.1.3 The Canoco for Windows workspace with project and log views

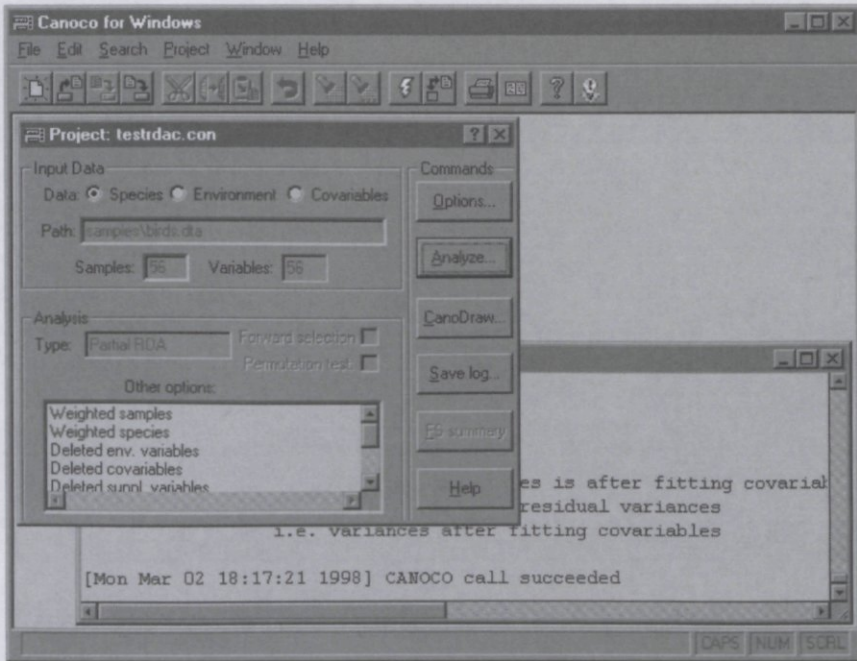



Figure 5-2 Canoco for Windows workspace with Project and Log views.

After you completed the Project Setup for a new project, opened an existing project or carried out an analysis, the Canoco for Windows workspace contains two windows per project (see Figure 5-2):

- the Project View with a summary of the project settings and buttons for common actions and
- the Log View which logs the creation dates of the project and, if available, results of the analyses

The Figure 5-2 shows the workspace after the action of the **Analyze...** button has been completed. The Log View shows the text “CANOCO call succeeded” preceded by the summary of a partial ordination analysis. The windows can be rearranged and resized. A window can be activated by clicking in it, by the commands in the Window submenu in the main menu or by switching between the Project View and the Log View with the toolbar button  or using the **F3** keyboard shortcut.

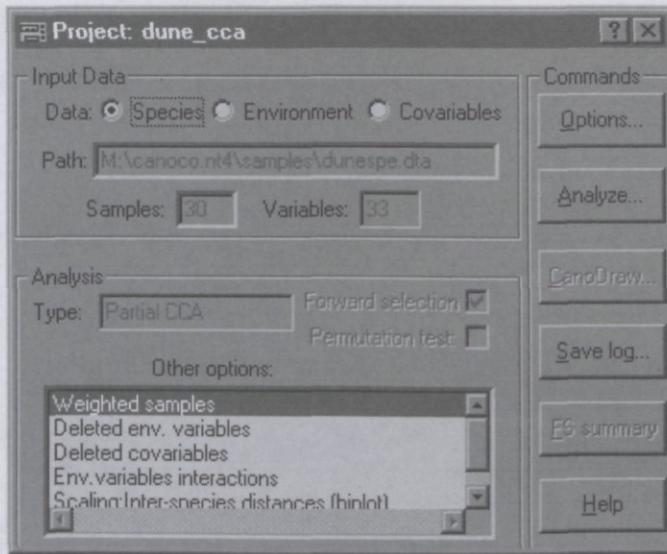


Figure 5-3 Project View window.

5.2 Project View

The Project View window (Figure 5-3) summarizes the project in the terms of the input data and the type of analysis. To the right is a column of buttons to tools that can be applied to a project. When you have just finished the Project Setup Wizard sequence (invoked by the **Options...** button) and saved the specifications to a project-file ("CON-file"), a natural sequence of actions is to click the **Analyze...** button first to calculate the ordination specified in the project, then to click in the Log View window to look for the ordination summary (or you can use **F3** to switch to the Log View), to activate the Project View again (by the **F3** key, by clicking the view or via the **Window** submenu) and, finally, to click **CanoDraw** button to plot the ordination diagrams. You may also wish to modify the option settings of the project by clicking **Options**, or to save the contents of the project log-window by clicking **Save log**, or to display the results of an automatic forward selection, if calculated, by clicking **FS summary**.

5.3 Selecting data sets and analysis type

The **Project Setup Wizard** starts with the three pages of this section. They specify the most important aspects of the ordination analysis, namely which data the analysis is to be applied to and which ordination method is used.

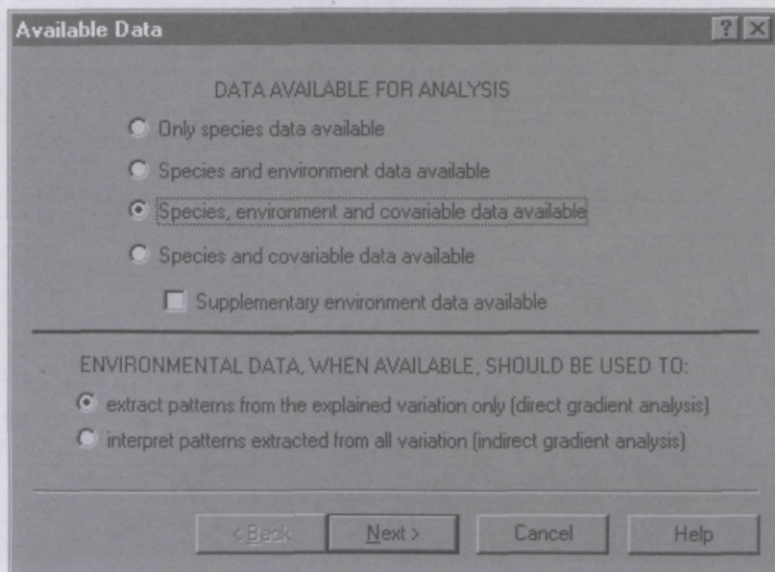


Figure 5-4 Available Data wizard page.

5.3.1 Available Data

Figure 5-4 shows the first page of the Project Setup wizard. You can specify 1 - 4 data sets for analysis. In Canoco terminology, the term *species* is used throughout for variables to be explained (response variables). The term *environmental variables* is used throughout for explanatory variables (predictors). A single available data set thus consists of 'species data', even if the data may actually be environmental measurements. The variables in this data set are to be explained by the ordination axes. Other data, in Canoco terminology 'environmental data', can be used to help interpret the ordination axes or to define them (indirect vs direct gradients). The ordination can be adjusted for effects of 'covariable data' (concomitant or nuisance variables). 'Supplementary environmental data' are a means of obtaining an alternative interpretation of already extracted ordination axes. In the current (1997) version of Canoco for Windows, CanoDraw cannot display supplementary environmental variables in the ordination diagram.

Direct gradient analysis gives an ordination with an optimal environmental basis. It does show only those patterns in the species data that can be explained by the available environmental data. The ordination axes are aggregates of the environmental variables that best explain the species data (constrained or canonical ordination). This is a form of regression analysis: species are explained by the environment via a small number of ordination axes.

Indirect gradient analysis gives an ordination that is calculated from the species data only. It shows the major patterns in the species data, irrespective of any environmental data. Environmental data, if available, are subsequently used to interpret the ordination. The ordination axes are theoretical gradients that best explain the species data. The axes are not constrained to be aggregates of the available variables (unconstrained ordination).

To proceed to the next page of the wizard, click the **Next>** button or press the **Enter** key. If you press the **Cancel** button all the current changes in the project are canceled and you return into the Project View of the project you started from.

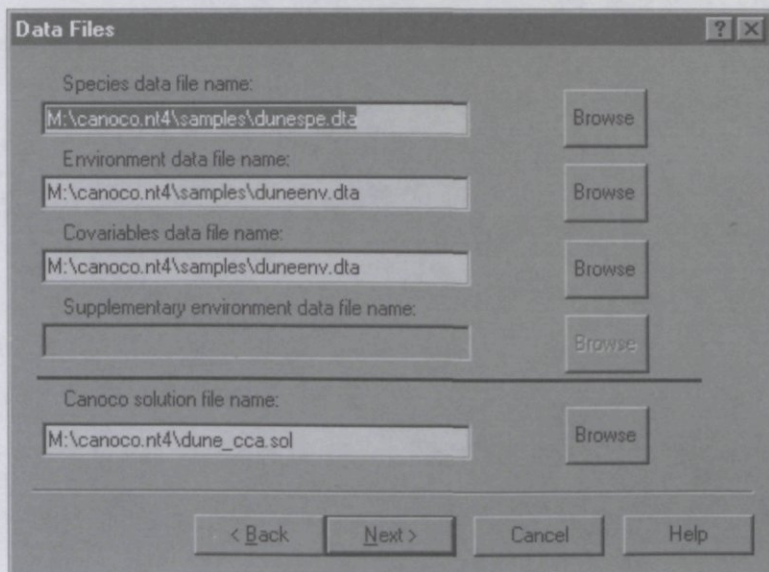


Figure 5-5 Data Files wizard page.

5.3.2 Data Files

Figure 5-5 shows the second page of the Project Setup wizard. Specify here the names of the data files. Either fill-in the names or click on **Browse** to select your file. Data files must conform to one of three strict data formats. Data in a spreadsheet can be converted to a correct format using the utility program WCanoImp. It is permitted to use a single file for environmental, covariable and supplementary environmental data. Later you must indicate which category each variable belongs to. The species data must be in a separate file. Here you must also give a name for a (new) file, the solution file, where the output is to be stored, e.g. ordination scores required for preparing ordination diagrams.

How to obtain a PCA on a correlation matrix. Ask for Only species data available, enter the data file with the variables as Species Data, center and standardize the data by species, choose scaling with focus on species correlations.

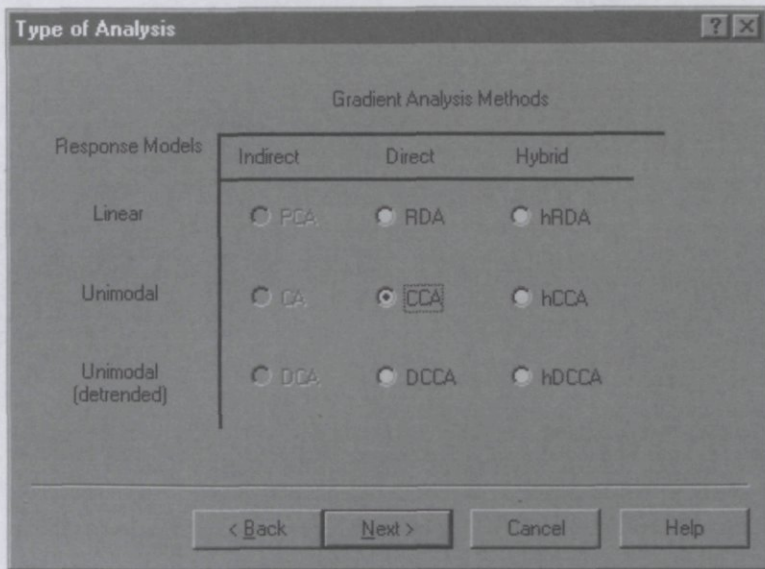


Figure 5-6 Type of Analysis wizard page.

5.3.3 Type of Analysis

Figure 5-6 show the third page of the Project Setup wizard. Specify here the ordination model. The appropriate model depends on whether you believe the species are responding roughly linearly to gradients (linear response) or have the best performance around some environmental optima (unimodal response). If you are not sure about this, analyze the species data first by a DCA and look at the length of gradient in the log-window. If the maximum gradient length exceeds 4 SD, your data show a strong unimodal response. Linear and unimodal methods stress patterns in absolute and relative abundance, respectively. Species data with many zeroes are often best analyzed with an unimodal method. With strong unimodal responses, correspondence analysis (and CCA with many environmental variables) tends to show an arch effect in the ordination diagram. This can be counteracted by choosing the detrended forms. Hybrid methods extract indirect gradients after direct ones. They are rarely used. If you wish to reconsider your initial choice for direct or indirect gradient analysis, go back to the first page.

The abbreviations of ordination methods, used in this page are:

- PCA Principal Components Analysis
- CA Correspondence Analysis
- DCA Detrended Correspondence Analysis
- RDA Redundancy Analysis (alias Reduced Rank Regression)
- CCA Canonical Correspondence Analysis
- DCCA Detrended Canonical Correspondence Analysis

☞ For a Canonical Variates Analysis (CVA), choose CCA and select in section 5.5.2 Hill's scaling with a focus on inter-species distances. See also section 8.4.3.

☞ For a multiple regression or ANOVA, choose RDA here. See also section 8.4.4.

5.4 Number of canonical axes and detrending method

The following two wizard pages do not appear in all analyses.

5.4.1 Canonical Axes

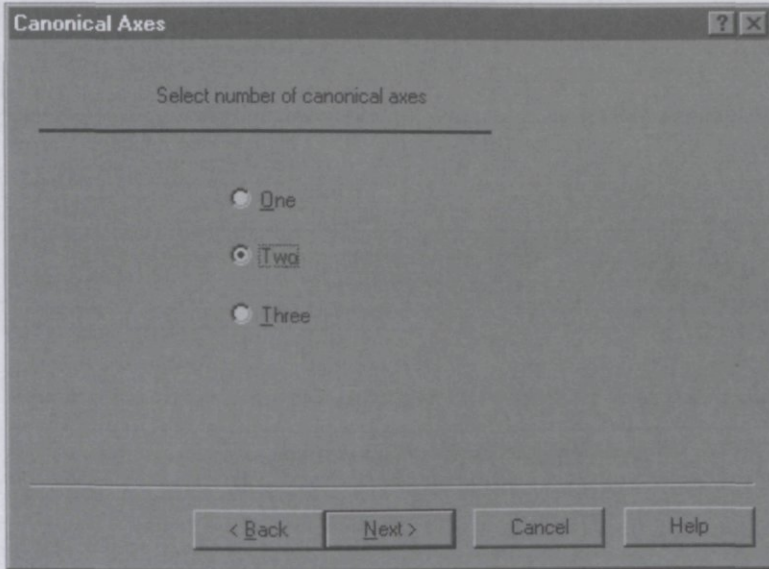


Figure 5-7 Canonical Axes wizard page.

The page shown in the Figure 5-7 appears only if you chose a hybrid method in the previous wizard page (see section 5.3.3). Canoco extracts four ordination axes at a time. In this page, you must select how many of these axes should represent direct gradients (i.e. canonical axes).

5.4.2 Detrending Method

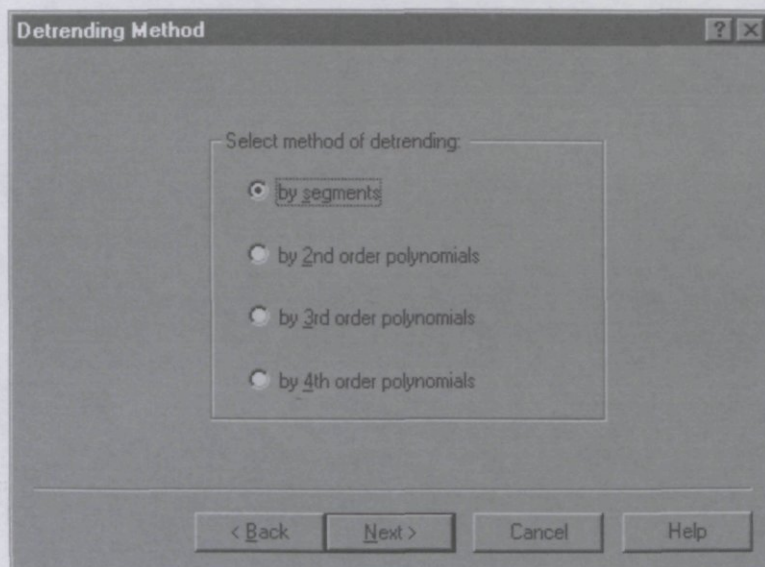


Figure 5-8 Detrending Method wizard page.

The page in Figure 5-8 only appears if you choose DCA or DCCA in section 0. With DCA, choose detrending-by-segments for DECORANA's default detrended correspondence analysis (Hill et Gauch, 1980) and for obtaining estimates of gradient lengths in standard deviation units of species turnover (SD). A length greater than 4 SD indicates a strong unimodal response. A less drastic method of detrending is by polynomials. In this method, non-linear dependence of axis scores on lower-order axes is fitted with a polynomial of degree 2, 3 or 4, and then the original scores are replaced by the residuals from this polynomial regression model. This is the recommended detrending method in DCCA.

5.5 Scaling of ordination scores

The options in this section determine how the sample scores are scaled. The sample scores can be linearly rescaled so that their mean square is equal to (or, in Hill's scaling, related to):

1. the eigenvalue (λ) of the ordination axis, or
2. the value 1.0, or
3. the square-root of the eigenvalue

The scaling of the species scores (and of all the other scores) follows from that of the sample scores by use of the transition equations - see section 6.3.2. In that section, we give an advice which scaling is best in which case, by focusing on the consequences of the scaling for the interpretation of ordination diagrams.

5.5.1 Scaling: Linear Methods

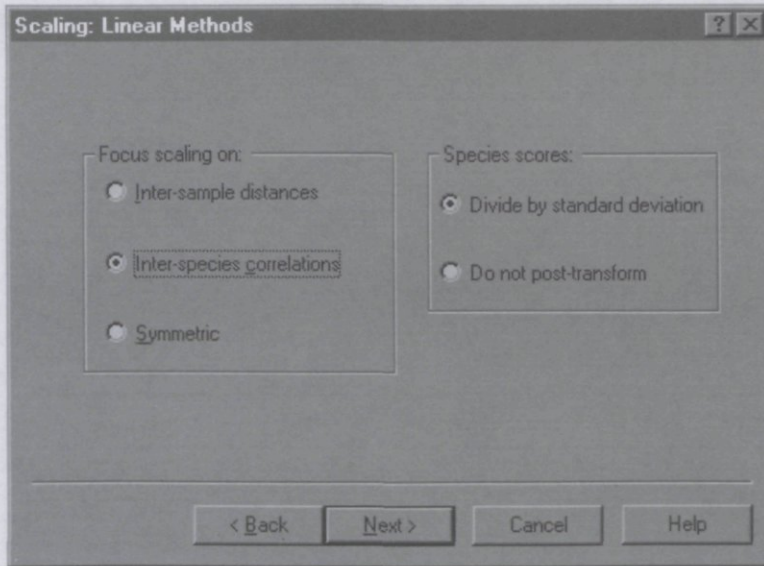


Figure 5-9 Scaling: Linear Methods wizard page.

In this page (Figure 5-9), displayed only for linear ordination methods, you should specify whether you predominantly want to interpret relationships among samples or among species from the ordination diagram. There is also an option to use for a compromise scaling. The only effect of your choice in this page is that the ranges of the samples and species scores on one ordination axis with respect to another are either shrunk or blown up. Your choice is unimportant if the eigenvalues of the axes of interest are similar.

Nominal environmental data define groups of samples. The sample scaling then allows you to interpret the distances between the groups. With quantitative environmental data, the species scaling results in an ordination diagram that reflects the environmental data and the correlations among the environmental variables. However, environmental effect sizes are best inferred from diagrams in sample scaling. With both nominal and quantitative environmental data, either scaling may be appropriate. Irrespective of your choice of scaling here, the ordination diagram displays the major patterns in the species data table, the table of correlations between species and quantitative environmental variables and, for nominal environmental data, the tables of class means per species (all interpreted by the biplot rule).

Untransformed, a species' score is proportional to the standard deviation of the species. Thus, species with a large variance (often the dominant species) lie far from the center of the ordination diagram and so unduly dominate the diagram. To counteract this effect and to make the species scores more comparable, you can opt here to divide them (after extraction of the axes) by their standard deviation. Then, the ordination diagram displays standardized species data, and correlations instead of covariances. In conjunction with species scaling, a correlation biplot is obtained; the length of a species' arrow is then a measure of fit (R) with the ordination diagram.

More information on the scaling of ordination scores is provided in the sections 3.5 and 6.3.2.

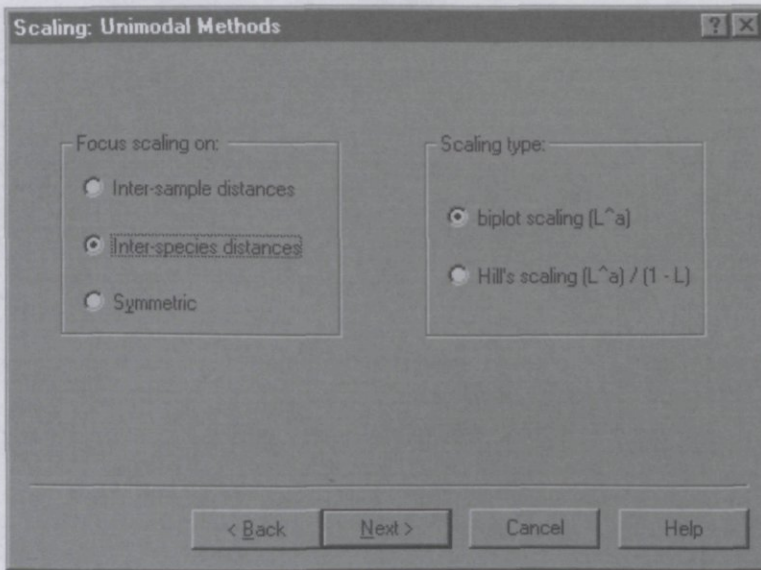


Figure 5-10 Scaling: Unimodal Methods wizard page.

In this page (Figure 5-10), displayed only for unimodal ordination methods, you should specify 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.

Nominal environmental data define groups of samples. The sample scaling then allows you to interpret the distances between the groups. With quantitative environmental data, the species scaling results in an ordination diagram that reflects the environmental data and the correlations among the environmental variables. However, environmental effect sizes are best inferred from diagrams in sample scaling. With both nominal and quantitative environmental data, either scaling may be appropriate. Irrespective of your choice of scaling here, the ordination diagram displays the major patterns in the species data table, the table of optima (weighted averages) of the species with respect to quantitative environmental variables and the relative abundances of species across environmental classes.

In sample scaling, the (species-derived) sample scores are weighted averages of species scores, i.e. species that occur in a sample lie around that sample's point in the ordination diagram. In species scaling, the species scores are weighted averages of sample scores, i.e. each species' point is at the center of its niche in the ordination diagram; samples in which a species occurs are scattered around it. These interpretations of weighted averages form the centroid principle.

If you select scaling with a focus on species distances, the resulting ordination diagram displays most accurately the dissimilarities between the occurrence patterns of different species. The measure of dissimilarity is, with the biplot scaling, the χ^2 distance and, with the Hill's scaling, the generalized Mahalanobis distance. Check this option if you wish to carry out a canonical variates analysis (linear discriminant analysis) using CCA.

Scaling type (biplot vs. Hill) addresses the issue how to infer the species data from the species-sample plot, other than by the centroid principle. The biplot scaling gives a more quantitative interpretation by the biplot rule and is most suited for short gradients. Hill's scaling

equalizes the average niche breadth for all axes and thus allows, for long gradients (strong unimodal response), the distance rule. This rule extends the centroid principle by taking a species' point as the optimum of its unimodal response.

In Hill's scaling the mean square of the sample scores is equal to $\lambda/(1-\lambda)$ when the focus is on the inter-sample distances, $1/(1-\lambda)$ when the focus is on the inter-species distances and $\lambda^{1/2}/(1-\lambda)$ when the scaling is symmetric.

More information on the scaling of ordination scores is provided in the sections 3.5 and 6.3.2.

5.6 Transformation of the species data

5.6.1 Centering and Standardization

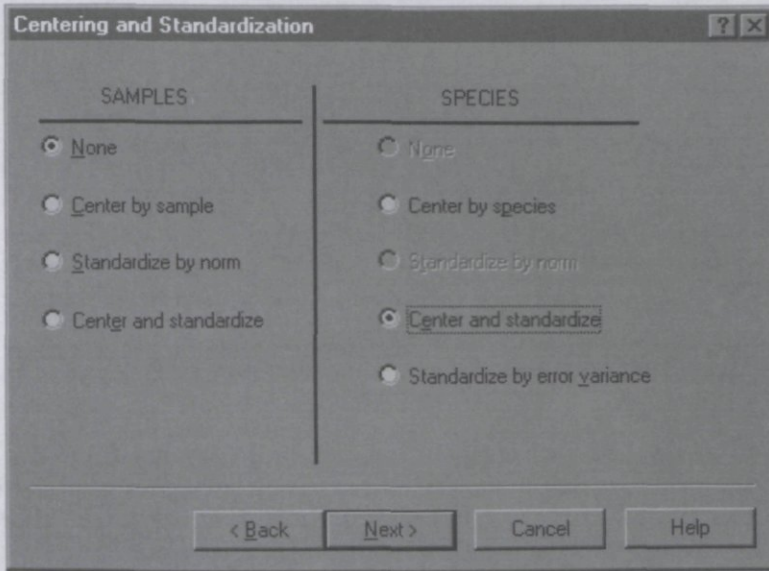


Figure 5-11 Centering and Standardization wizard page.

This page (Figure 5-11) is displayed only for linear ordination methods. 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 PCA/RDA (based on a covariance matrix) is obtained by centering by species only. Each species is then implicitly weighted by its variance. Standardized PCA/RDA (based on a correlation matrix) is obtained by centering and standardization by species. This choice is particularly suited if the 'species' are measured in different units, e.g. when the data are actually environmental variables such as pH, organic matter (g) or water depth (m). Aitchison's (1990) log-ratio analysis of compositional data is obtained by centering log-transformed 'species' data by samples as well as by species. With environmental data it is possible to weight species inversely to the error variance that remains after fitting the species to the environmental and covariable data.

Centering by sample **or** by species is achieved by subtracting sample (row) **or** species (column) means from the value in the species data. The resulting data matrix then has zero row **or** column means.

When standardizing by samples **or** by species, the species data values are divided by the corresponding sample **or** species norm (root mean square of values).

5.6.2 Transformation of Species Data

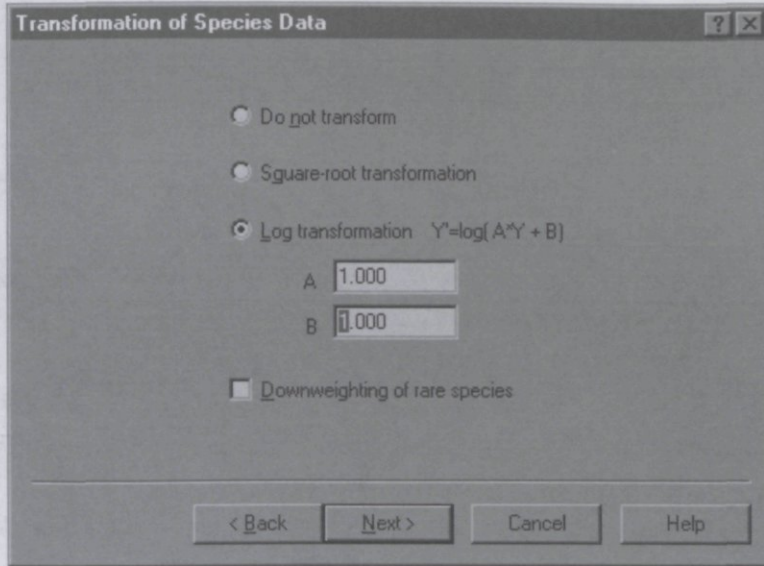


Figure 5-12 Transformation of Species Data wizard page.

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. Taking logarithms turns linear models into ecologically more plausible multiplicative models. If the data contain zero values, a small value (**B**) must be added, because **log(0)** is undefined. For technical reasons, **B** must then be equal to or greater than 1 in Canoco, but this limitation can be circumvented by specifying a value for **A>1**. For example, if you would like to add **0.1** to the original data, specify **A=10** and **B=1**.

In unimodal methods, rare species may have an unduly large influence on the analysis. Their influence can be reduced by checking the **Downweighting of rare species** box.

5.7 Data editing options

5.7.1 Data Editing Choices

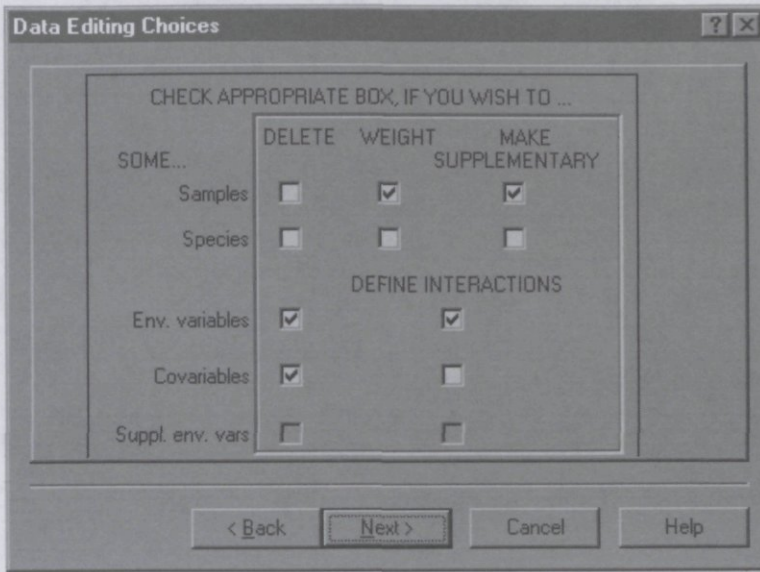


Figure 5-13 Data Editing Choices wizard page.

Checking a box in this page (Figure 5-13) allows you to specify later which samples or species you wish to delete, weight, or make supplementary (i.e. "passive" in older CANOCO terminology). For example, you may wish to downweight an unreliable sample or to upweight a target species. Exotic species may be made supplementary in a study of native species. Supplementary (passive) samples or species do not influence the ordination axes, but are added afterwards so that their relation to the other samples or species can still be judged from the ordination diagram.

You may also opt to delete some explanatory variables. If you are using a single file for all explanatory variables in a partial direct gradient analysis, the corresponding delete boxes are checked by default. You must then indicate later which variables are covariables and which are environmental variables.

The effect of one explanatory variable may depend on another. You can explicitly model such interaction effects. Check the interaction box also to define polynomials.

If you are modifying a project from a previous analysis, unchecking a box undoes the existing specification. For example, if you open a project where some samples were deleted, the checkbox in the **DELETE** column at the first (**Samples**) row is checked. If you un-check it and store the modified project, no samples will be deleted in the new analysis.

To see which species, samples, environmental variables and/or covariables are available for the analysis, check the appropriate Delete box. Canoco for Windows then lists the available species, samples, environmental variables and/or covariables. You need not to actually delete any items nor you need to uncheck the Delete box.

5.7.2 Delete Items

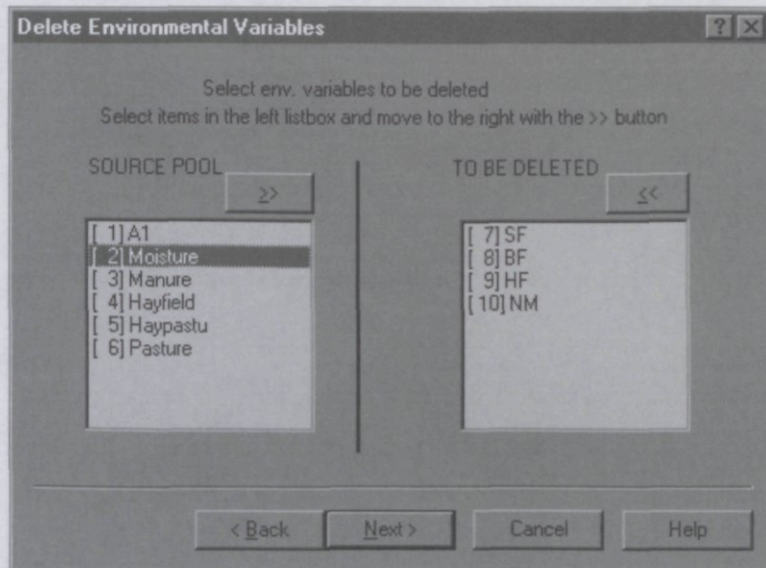


Figure 5-14 Delete Items wizard page.

The Delete Items page (Figure 5-14) allows you to specify the items to be deleted prior to the analysis (samples, species, environmental variables, covariables, or supplementary environmental variables). Highlight items to be deleted and click on the >> button. You can also undo deletions by highlighting items in the list on the right and clicking on the << button. To highlight (select) an item in any of the lists, click the item with the left mouse button. To add new individual items to an existing selection, click each of them with the left mouse button, while holding down the **Ctrl** key. To select a contiguous range of items, select the first item and then click the last one, while holding down the **Shift** key.

5.7.3 Set Weights for Samples / Species

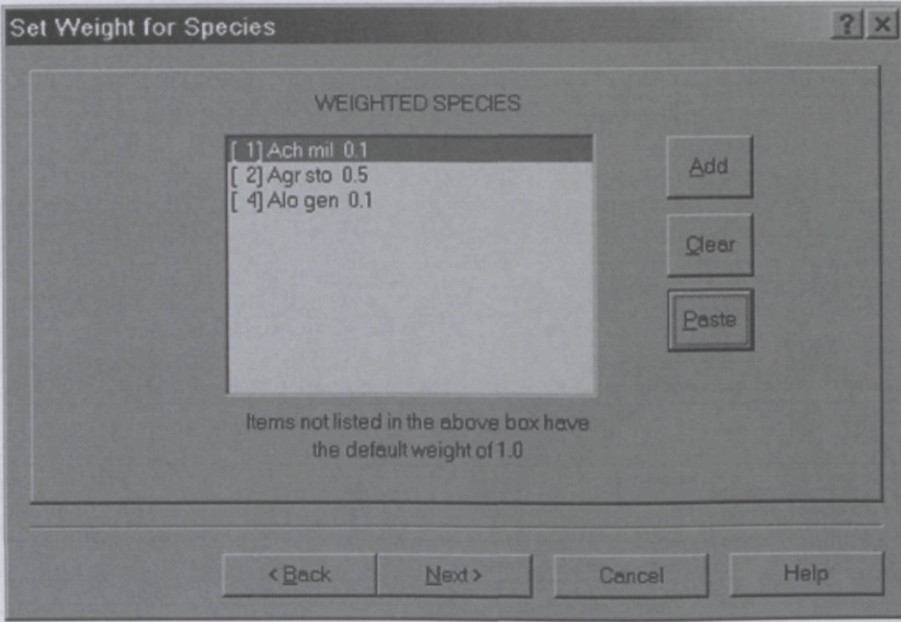


Figure 5-15 Set Weights wizard page.

This page (Figure 5-15) appears when you asked to weight some species or samples. Items (samples or species) with a non-default weight are listed on the page. If the list is empty (e.g. on its first appearance), all items have an implicit weight **1**. You can modify the list with the **Add** and **Clear** buttons. You can remove items from the list by highlighting them and clicking on the **Clear** button. Alternatively, you can populate the list using the **Paste** button. This action imports the weights of samples or species from the Windows Clipboard. You can copy the weights onto the Clipboard from Microsoft Excel spreadsheet either in the form of one column (the number of rows must match number of items in your data) or as two columns, the first one giving the identification numbers of individual samples (species) and the other one specifying the actual weight value. In the later case, only the weights of samples (species) explicitly listed in the first column are modified. The other retain the values they had before this action.

If you click the **Add** button, another dialog box is displayed and you are asked to specify a weight value and items with that weight there (Figure 5-16).

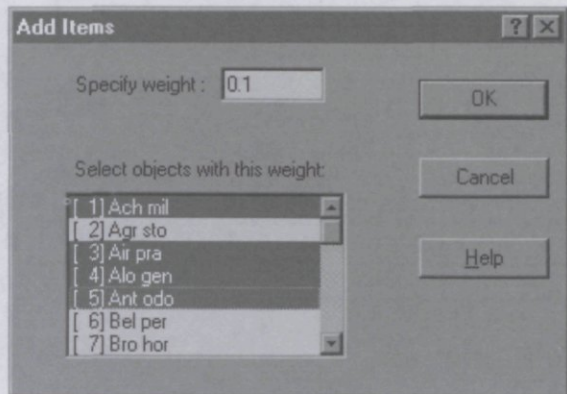


Figure 5-16 Add Items dialog.

First specify a weight, then highlight the items (samples or species) to which this weight must be assigned and, once selected, click **OK**. Weights greater than 1.0 upweight items and weights less than 1.0 downweight them. If a sample receives a weight of 2.0, the same ordination could also have been obtained from an unweighted analysis by including that particular sample twice in the data file(s). Weights should exceed the value **0.01** and cannot be larger than **100.0**.

5.7.4 Supplementary Samples / Species

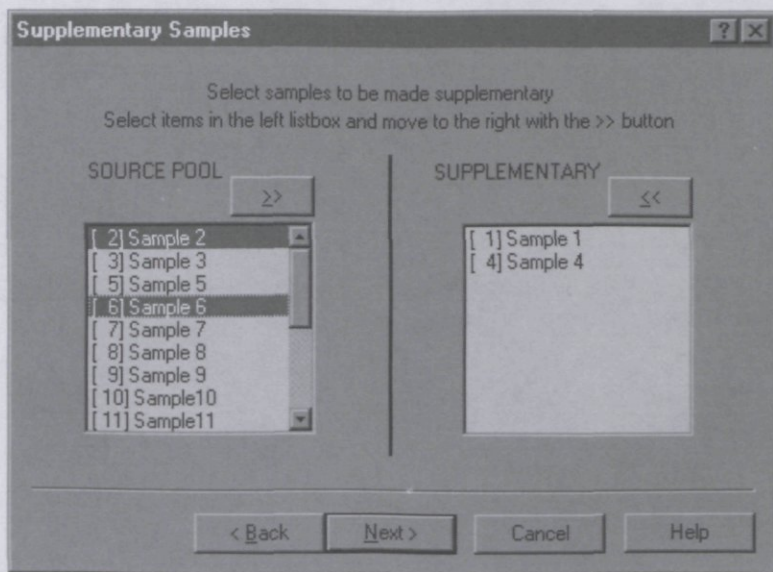


Figure 5-17 Supplementary Samples / Species wizard page.

This page (Figure 5-17) appears when you asked to make some species or samples supplementary. Supplementary (passive) samples or species do not influence the ordination axes, but are added afterwards so that their relation to the other samples or species can still be judged from the ordination diagram. Highlight items to be made supplementary and click on the

>> button. You can also re-activate supplementary items by highlighting items in the list on the right and clicking on the << button. To highlight (select) an item in any of the lists, click the item with the left mouse button. To add new individual items to an existing selection, click each of them with the left mouse button, while holding down the **Ctrl** key. To select a contiguous range of items, select the first item and then click the last one, while holding down the **Shift** key.

5.7.5 Interactions of Variables

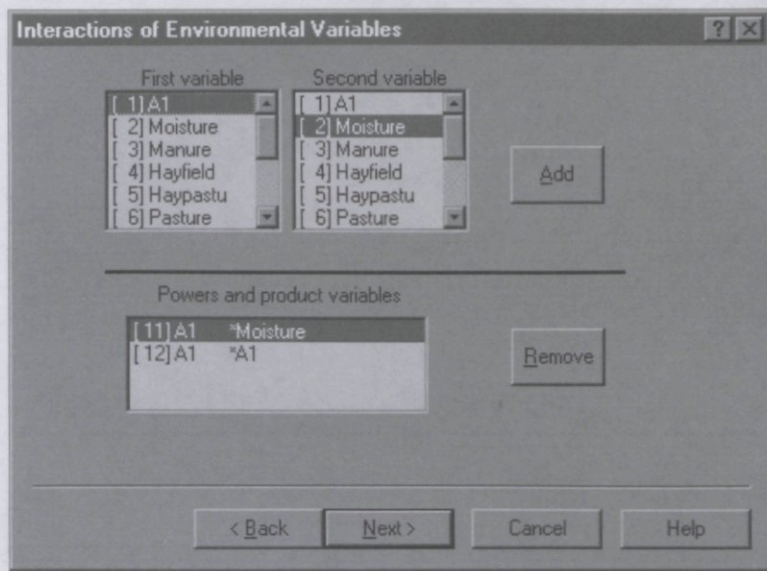


Figure 5-18 Interactions of Variables wizard page.

This wizard page (Figure 5-18) allows you to specify which products and powers of variables you wish to add to the set of explanatory variables (environmental variables, covariables or supplementary environmental variables). As in multiple regression analysis, products and powers are a means of studying interaction and polynomial effects.

For statistical testing of the pure interaction effect between P and N, say, the variables P and N must be in the covariables data and the product variable P*N must form the environmental data. The product variable P*N can be defined on this page, even if you already deleted P and N from the environmental data. If P and N are factors, the dummy variables coding their levels must be in the covariable data and all the products of these dummies in the environmental data. See Example E40 in section 8.3.3.

The product P*N*K can be formed by first defining P*N and the defining the product of the results with K. If P*N is subsequently removed from the list, the variable P*N*K is removed automatically as well. Analogously, N³ (N*N*N) requires the existence of N² (N*N).

To create a new product variable, highlight one item in the **First variable** list and one item in the **Second variable** list and click the **Add** button. The created product variable is placed in the bottom list, as well as at the end of both lists at the top, so it can be used to create more complicated terms. If you want to remove an already defined product variable, highlight it in the lower list (named **Powers and product variables**) and click the **Remove** button. Note that all the higher-order interaction terms, that were created using the removed term are automatically removed as well.

5.8 Forward selection

5.8.1 Forward Selection of Environmental variables

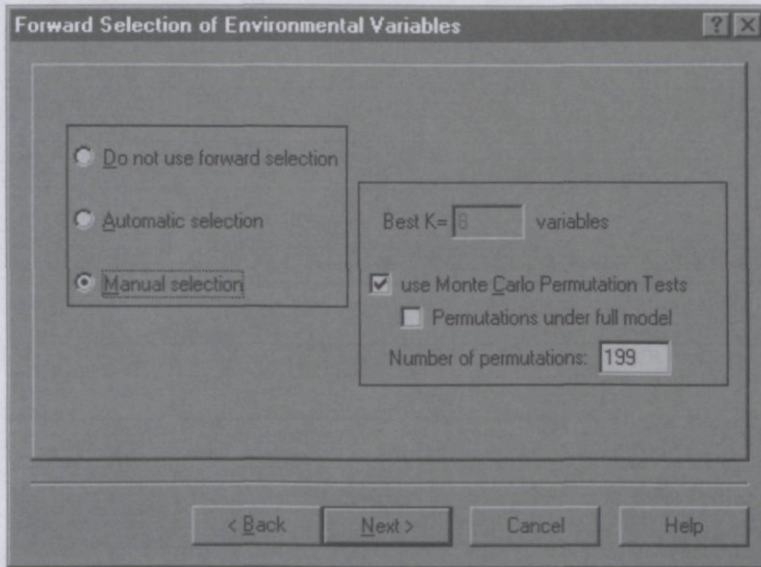


Figure 5-19 Forward Selection wizard page.

This wizard page (see Figure 5-19) allows you to specify forward selection options to be used when analyzing the current project. Forward selection is useful for ranking environmental variables in their 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. You can alter the number of permutations to be carried out for each test by specifying value less than **10000**. By default, residuals from the reduced model ('null model') are permuted. Alternatively, residuals from the full model are permuted. The reduced-model method better maintains the type I error in small data set. Without covariables, the method yields the exact Monte-Carlo significance level. The full-model method gives slightly lower type II error.

If you select the **Manual selection** option, you may choose in each step of the selection process which environmental variable is tested or included in the model. If you plan to test for variable significance, check also the Monte Carlo permutation test checkbox.

5.8.2 Forward Selection report

Marginal Effects		
Variable	Var.Num.	lambda-1
A1 *Mois	11	0.31
Moisture	2	0.28
A1 *A1	12	0.22
A1	1	0.20

Conditional effects				
Variable	Var.Num.	lambda-A	P-value	F-value
A1 *Mois	11	0.31	0.005	3.71
Moisture	2	0.17	0.010	2.20
Manure	3	0.08	0.435	0.98
Pasture	6	0.06	0.570	0.86

Figure 5-20 Forward Selection report dialog.

This dialog box (Figure 5-20) summarizes the results of the automatic forward selection procedure and can be displayed after the analysis was run by clicking the **FS Summary** button in the project view.

The table at the top of the dialog box, headed **Marginal effects**, lists the individual environmental variables in order of the variance they explain singly i.e. when that particular variable is used as the only environmental variable (**lambda-1** column). The variance is in addition to the variance explained by covariables, if present, but ignores the other environmental variables.

The table at the bottom, headed **Conditional effects**, shows the environmental variables in order of their inclusion in the model, together with the additional variance each variable explains at the time it was included (**lambda-A**) and, if Monte Carlo tests were asked for, the significance of the variable at that time (**P-value**) together with its test statistics (**F-value**). A variable contributes significantly (at the 5% significance level) to the model of already included variables if the P-value is less than or equal to 0.05.

Both tables can be copied to the Clipboard by clicking the **Copy** button. Click **OK** if you are ready to proceed.

5.8.3 Manual Forward Selection

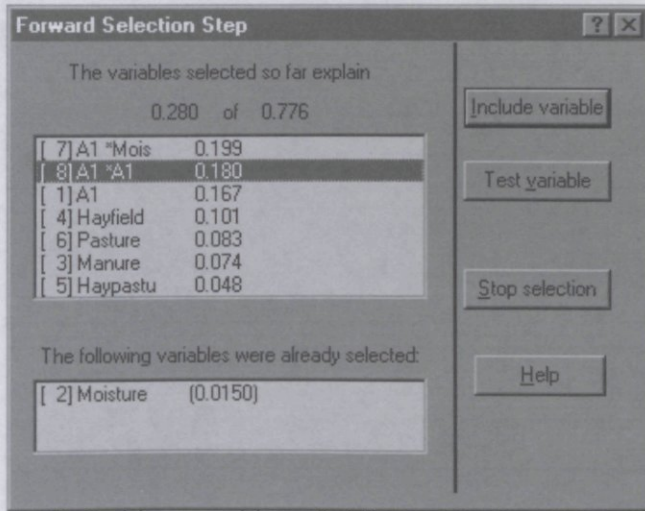


Figure 5-21 Forward Selection Step dialog.

Manual forward selection is a step-wise process of building a model for the species data. Starting from an empty model, you can select environmental variables, one after the other, for inclusion in the model. This dialog box (Figure 5-21), which appears at each selection step during the analysis, allows you to select an environmental variable for inclusion, to test the statistical significance of a variable, if it would be included in the current model, and to stop the selection process. The top panel lists environmental variables that are available for selection in order of the extra variance each of them would explain when included in the current model. The bottom panel shows the environmental variables already selected. In the first step, the bottom panel is empty and the explained variance is 0.000. Also displayed is the maximum amount of variance that can be explained by including all variables in the model.

If you specified Monte Carlo permutation tests in the project setup, you can use the **Test variable** button and the highlighted variable in the upper list is tested. At the end of the test, the results of the Monte Carlo permutation test are displayed in a dialog box (Figure 5-22).

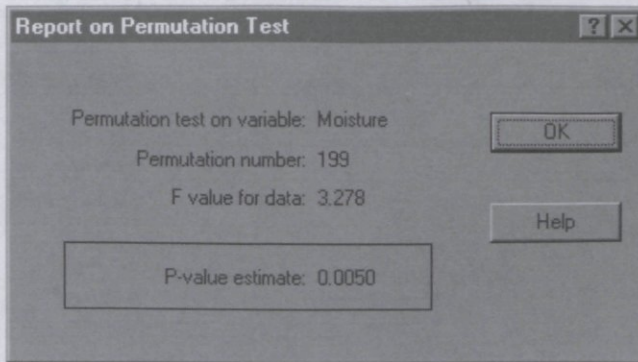


Figure 5-22 Permutation Test results dialog.

This dialog box identifies the variable on test, the number of Monte Carlo permutations carried out in determining the P-value, the value of the test statistic for the data (F-value), and the resulting significance level (P-value). Click **OK** if you are ready to proceed to the next step in manual forward selection.

5.9 Global significance tests

5.9.1 Global Permutation Test

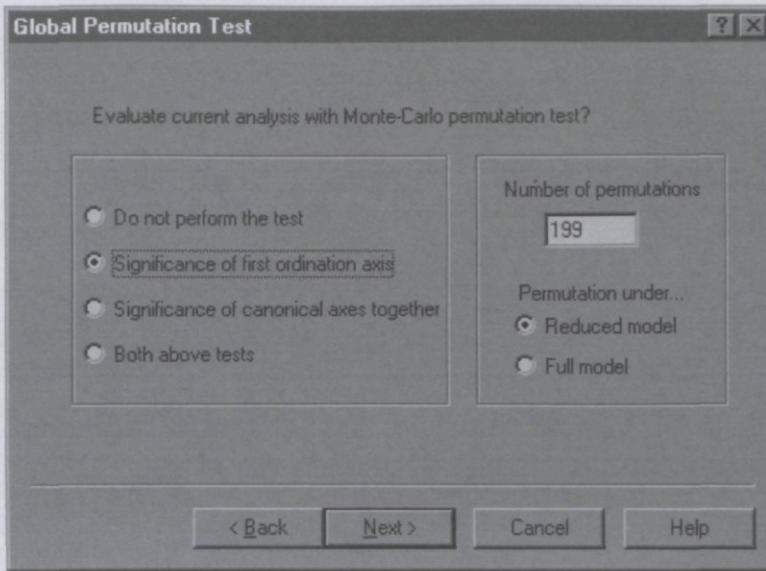


Figure 5-23 Global Permutation Test wizard page.

In this wizard page (Figure 5-23), you should specify whether you want to determine the statistical significance of the relation between the species and the whole set of environmental variables, given the covariables (if these are present in the project). 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. Test based on the first canonical axis has maximum power against the alternative hypothesis that there is a single dominating gradient that determines the relation between the species and environment. This test also requires more computer time than the test based on the all canonical axes. In this alternative test, the test statistic used is an F-ratio of the sum of all canonical eigenvalues (which takes the role of the regression sum of squares) and the residual sum of squares. This statistic yields an omnibus test, i.e. a test which is sensitive to all kinds of deviations from the null hypothesis.

The test is carried out by a Monte Carlo permutation test. You can alter the number of permutations to be carried out, specifying value up to **10000**. For a test at the 5% significance level, a minimum of 19 permutations is required. The power of the test increases with the number of permutations, but only slightly so beyond 199 permutations.

By default, residuals from the reduced model ('null model') are permuted. Alternatively, residuals from the full model are permuted. The reduced model method better maintains the

Type I error in small data sets. Without covariables, the method yields the exact Monte Carlo significance level. The full-model method gives lower type II error, but only so slightly that it is best to stick to the default (Anderson & Legendre, 1999).

5.10 Specifying the randomization model

The following wizard pages appear if you asked for a Monte Carlo permutation test in either the Forward selection page (section 5.8.1) or in the Global permutation test page (section 5.9.1).

5.10.1 Permutation Type

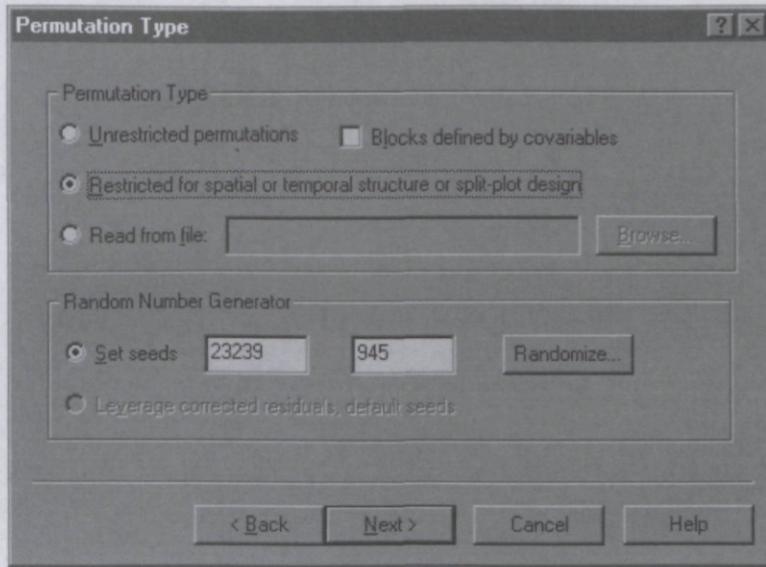


Figure 5-24 Permutation Type wizard page.

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, exchange 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.

Restricted permutation types are appropriate for line transects, time series and rectangular grids, if recorded at equal intervals, and for balanced split-plot designs and related designs such as Before-After-Control-Impact (BACI) designs, repeated measurement designs, and many ANOVA designs with random (nested or crossed) factors. If there are more line transects (or time series or grids), each one should form a block. This allows you to test for within-transect variation. In contrast, if you want to test for between-transect variation, each transect must not form a block, but a whole-plot in a split-plot design (to be specified later). Whole-plots must be of equal size. If the whole-plots themselves are arranged in blocks, you must check the **Blocks** option here.

For blocks of equal size, the required permutations can sometimes be obtained without block-defining covariables. Try the options of the split-plot design for this.

If your data require yet another permutation type, you can specify here a file with your own permutations (permutation file). With **n** active samples in the analysis, the permutation file (with ASCII format) should contain a sufficient number of permutations of the numbers **1, 2, ..., n**, with one permutation after the other. See Table 7.3 and the example BACIIISPE in section 8.3.7.

The generation of Monte Carlo permutations requires initial seeds. You may alter the seeds manually or by clicking on the **Randomize** button: the performed randomization is based on the computer system time. If more than one test is applied to the same data, it is prudent to specify different seeds for each test.

Finally, you can set here an option that has not yet been thoroughly studied. If set, leverage corrected residuals are permuted rather than ordinary residuals. For technical reasons, the method can only be used with the default seeds.

5.10.2 Definition of Blocks

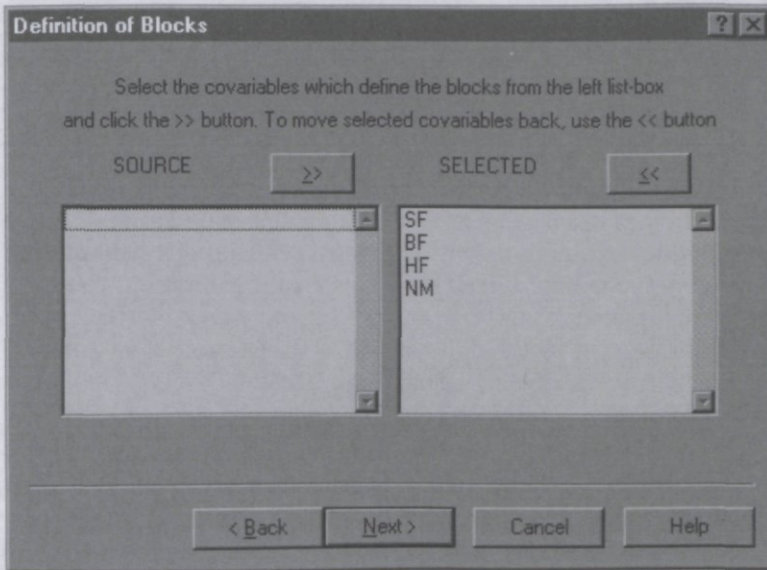


Figure 5-25 Definition of Blocks wizard page.

Specify in this wizard page (Figure 5-25) which of your covariables indicate the blocks (one dummy variable for each block). Samples with the same value for the selected covariables will end up in the same block. Blocks are usually indicated by a set of dummy (0/1) variables, but it is possible to use multiple-valued variables. Highlight these dummy covariables and click on the >> button. You can also undo selections by highlighting variables in the list on the right and clicking on the << button.

5.10.3 Permutation Restrictions

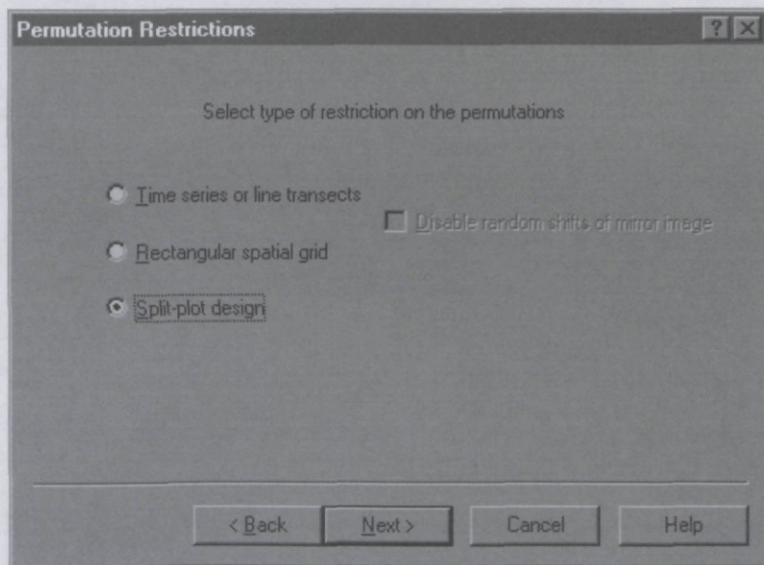


Figure 5-26 Permutation Restrictions wizard page.

In this wizard page (Figure 5-26), you should specify whether your data are from a line transect / time series, from a rectangular grid, or from a split-plot design. If you have multiple transects (or series or grid) of equal dimension, you may also select here the split-plot design option, which encompasses the other options. With the split-plot design you can test for split-plot factors (e.g. within-transect variation) as well as for whole-plot factors (e.g. between-transect variation). With the split-plot design, you can also analyze Before-After-Control-Impact (BACI) designs, repeated measurement designs, and many ANOVA designs with random (nested or crossed) factors.

If your samples are sampled in a regular time sequence or on a line transect with equal intersample distances, select the **Time series or line transects** option. With blocks in the analysis, each block must contain a single time series or line transect. Series or transects may differ in size between blocks. Between-series or between-transect variation is excluded from the test. The permutations for series / transects or grids are cyclic or toroidal shifts. It is rarely needed, but you may disable shifts from the mirror image of the series / transect or grid.

If your samples are arranged on a rectangular grid in space, select the **Rectangular spatial grid** option. With blocks in the analysis, each block must contain single grid. Between-grid variation is excluded from the test. The permutations generated for grids are independent toroidal shifts.

5.10.4 Grid Dimensions

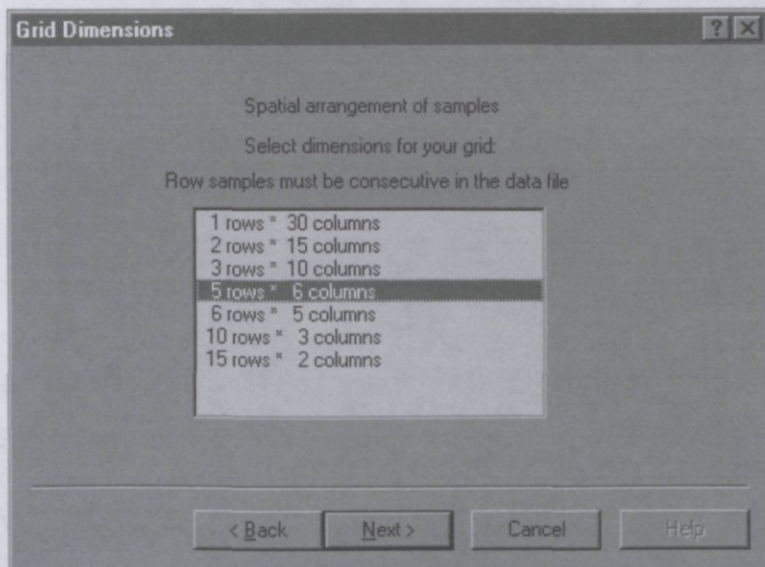


Figure 5-27 Grid Dimensions wizard page.

Specify in this wizard page the dimensions of the grid(s) you have. The units of the grid can be samples or whole-plots (sets of samples). Rows and columns are not arbitrary entities here: a row consists of samples that are consecutive in the data file. For example, if you specify that your grid has 3 rows and 10 columns, then Canoco for Windows assumes that the first 10 units (samples or whole-plots) in the data file form the first row, the next 10 the second row, etc. In contrast, if you specify that your grid has 10 rows and 3 columns, then Canoco for Windows assumes that the first 3 units (samples or whole-plots) in the data file form the first row, the next 3 the second row, etc.

5.10.5 Split-Plot Design I

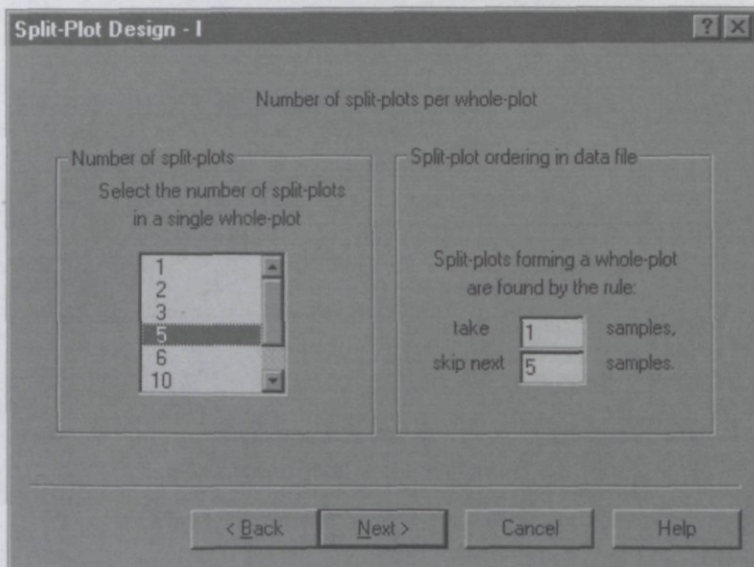


Figure 5-28 Split-Plot Design I wizard page.

This is the first wizard page (Figure 5-28) of the two pages used for specifying a split-plot design. Split-plot design is a hierarchical design with two levels of units: whole-plots containing split-plots. Split-plots are the lowest level sampling units, i.e. the samples in your data file. Examples are samples-within estuaries, plots-within-stands, plots-along-transects, relevés-within-time-series (in a permanent plots study). Specify here the number of samples per whole-plot and how the samples comprising a whole-plot are arranged in the data file. If the samples of a whole-plot are consecutive in the data file, you can apply the default rule (take 1 sample, skip next 0 samples), because no samples need to be skipped. For example, if in a permanent plot study, the vegetation of 50 locations is monitored at 20 points in time, locations are whole-plots and relevés (samples) split-plots. In the data file, the samples may be arranged by locations or by times. Arrangement by locations means that all data of a single location is consecutive in the data file, so that the default rule applies (the rule 'take 20 samples, skip 0' would work, as well). Arrangement by times means that all data of a single time point are consecutive in the data file. With a standard order of locations within times, the data of each location are found by the rule 'take 1 sample, skip next 19 samples'.

Here is an example which would require you to specify a take number other than the default. Let the whole-plots A, B, and C consist of 6 samples each and let the samples happen to be arranged as AABBC AABBC AABBC in the data file. Then the rule 'take 2, skip 4' correctly specifies the whole blocks. Such data arrangements occur naturally in ANOVA designs with random crossed factors.

5.10.6 Split-Plot Design II

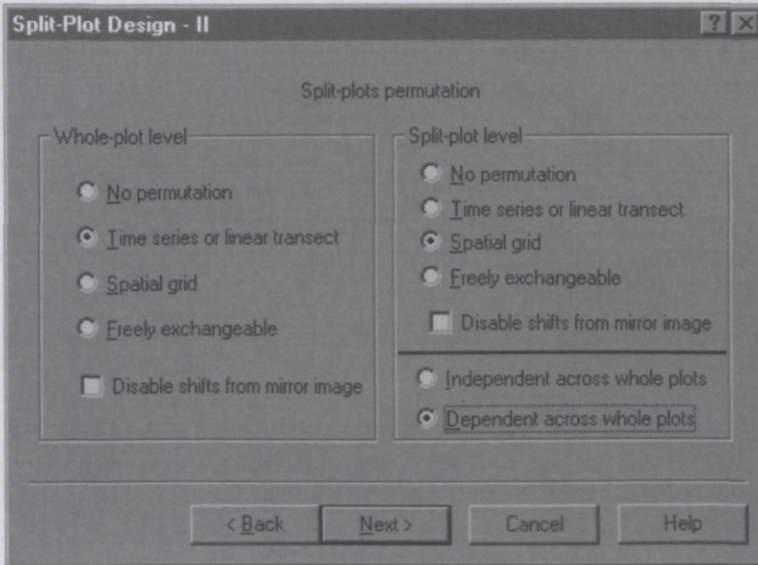


Figure 5-29 Split-Plot Design II wizard page.

This is the second Setup Wizard page (Figure 5-29) used for specifying a split-plot design. The split-plot framework allows many permutation types, all of which define fewer distinct permutations than the unrestricted permutation type. This framework is applicable when the samples of your data file can be grouped into whole-plots. The samples (split-plots) of a whole-plot are related because they share an error term or a random factor. Whole-plots should be of equal size, because whole-plots with different numbers of samples can not be exchanged.

The effect of environmental variables that vary between whole-plots (e.g. the whole-plot factors of a split-plot design) can be tested by permuting whole-plots while keeping the split-plots of each whole-plot together. This test is obtained by specifying that whole-plots are freely exchangeable whereas split-plots within whole-plots are not to be permuted. If the whole-plots form a time series, a line transect, or a spatial grid, the whole-plot permutations can be restricted to cyclic or toroidal shifts so as to account for autocorrelation among whole-plots. If your environmental variables vary little or not at all between whole-plots, the test will never show significant effects.

The effect of environmental variables that vary within whole-plots (e.g. the split-plot factors of a split-plot design) can be tested by permuting split-plots within whole-plots without permuting whole-plots. Whole-plots restrict the permutations in the same way as blocks, but without the necessity of block-defining covariables. If the split-plots form a time series, a line transect, or a spatial grid, the split-plot permutations can be restricted to cyclic or toroidal shifts so as to account for autocorrelation among split-plots. If the split-plots form parallel time series and time is an autocorrelated error component affecting all series, the same shift should be applied to all time series. This is specified by checking dependent split-plot permutations across whole-plots. In the standard split-plot design, split-plots of different whole-plots are unrelated and the permutations of split-plots in different whole-plots should be independent. If your environmental variables vary little or not at all within whole-plots, the test will never show significant effects.

It is rarely needed, but for whole-plots and for split-plots you may disable shifts from the mirror image of the series / transect or grid.

Many designs can be analyzed in the split-plot framework. The manual gives examples of specifying tests in repeated measurement designs, ANOVA with random nested and crossed factors and the BACI design. The specification for the Before-After-Control-Impact (BACI) design is summarized below.

In a BACI design, the data are parallel time series, one per site; sites are the whole-plots and the samples of the time series are the split-plots. Sites and times must be dummy covariables and the impact itself must be coded by one or more environmental variables (e.g. by the product variables Impact-site*After-times).

With a single Control and a single Impact site, the data consist of two time series with, preferably, many Before-impact and many After-impact times. A test of the impact can be obtained by permuting the samples in the time series in the same way, i.e. by checking the box Dependent [split-plot permutations] across whole-plots. If there is no autocorrelation in the time series of Control-Impact differences, the permutation must be restricted to the cyclic shifts of time series. For the whole-plots, check freely exchangeable (or, to the same effect, no permutation). See section 8.3.7 for an example (BACI1SPE).

With replicated Control and Impact sites with just a single Before-impact and a single After-impact sample, the test must be based on permuting the whole-plots (sites) without permuting the split-plots (samples). This test is not powerful with only a moderate number of Control and Impact sites. If more Before-impact and After-impact samples are available, additional power is obtained by also permuting the split-plots by identical cyclic shifts. See section 8.3.9 for an example (BACI3SIT).

The options in the **Whole-plot level** group define how the whole-plots are permuted in this analysis. All samples belonging to the same whole-plot are kept together in the permutation. The permutations allow you to test the effect of the environmental variables that vary between whole-plots:

- select the **No permutations** option if you want to test the effect of environmental variables that vary within whole-plots (e.g. the split-plot factors of a split-plot design). Unpermuted whole-plots restrict the permutations in the same way as blocks, but without the necessity of block-defining covariables.
- select the **Time series or line transect** option if the whole-plots are sampled in a regular time sequence or on a line transect with equal interpoint distances.
- select the **Spatial grid** option if the whole-plots are arranged on a rectangular grid in space.
- select the **Freely exchangeable** option if, under the null hypothesis, the whole-plots are exchangeable, e.g. if you are testing whole-plot factors in a split-plot design.

The options in the **Split-plot level** group define how the split-plots are permuted within each of the whole-plots. The permutations allow you to test the effect of environmental variables that vary within whole-plots:

- select the **No permutations** option if you want to test the effect of environmental variables that vary between whole-plots (e.g. the whole-plot factors of a split-plot design).
- select the **Time series or line transect** option if the samples of each whole-plot are sampled in a regular time sequence or on a line transect with equal intersample distances.
- select the **Spatial grid** option if the samples (split-plots) of each whole-plot are arranged on a rectangular grid in space.
- select the **Freely exchangeable** option if, under the null hypothesis, the split-plots are exchangeable within the whole-plots.

For the permutation of split-plots you must also decide whether the samples in different whole-plots are unrelated or dependent across the whole-plots. If you selected independence, then the permutation of samples within one whole-plot is independent of the permutation of the samples within the other whole-plots. If you, on the other hand, select the **Dependent across whole plots** option, the permutation scheme at each permutation iteration is identical across all available whole-plots. For example, if a number of locations is sampled at the same points in time, the samples of different locations (whole-plots) are related by time. The samples within different whole-plots must be in the same meaningful order in the data file (e.g. in order of time).

5.11 Saving the project

After you defined or modified options in the CANOCO project with the Project Setup Wizard, you can save them in a Canoco project file (with the .CON file extension). You can save an already named project by using the **Save** button on the toolbar, by selecting the **Save** command from the **File** submenu, or by using the **Ctrl-V** keyboard shortcut. To save an existing CANOCO project under a different name, you can invoke the **File Save As** dialog box by clicking the **Save As** button on the toolbar or by selecting the **Save as...** command from the **File** submenu. For new projects, the **Save As** dialog box appears automatically.

5.12 Running the analysis and saving the log

After you specified the analysis settings or after opening a fully-defined CANOCO project file in the Canoco for Windows workspace, you can analyze the project by:

- clicking the **Analyze** button in the Project View window.
- clicking the **Analyze** button in the Canoco for Windows toolbar.
- selecting the **Analyze** command from the **Project** submenu or.
- using the **Ctrl-A** keyboard shortcut.

Note that any of these methods functions only if the Project View (not the Log view) is active! While the analysis is running, a progress panel is displayed and you cannot work with the Canoco for Windows during that time. You may disable the progress panel in the Options dialog box (see section 5.14).

The analysis results are stored in the Log View window of your project. Under Windows 9x, the capacity of the log is limited; so if the size of the analysis log exceeds the Log window capacity, part of the output is removed. You get the opportunity to save the whole analysis log in a text file first. If there is a log from a previous analysis, you get the opportunity to save this log to a file, before it is deleted from the Log View to make place for the log from the new analysis. Fortunately, it does not happen often that the log exceeds the Log View capacity, except when you have very many environmental variables which result in a large correlation matrix. These capacity problems are not present in Windows NT, Windows 2000, or Windows XP.

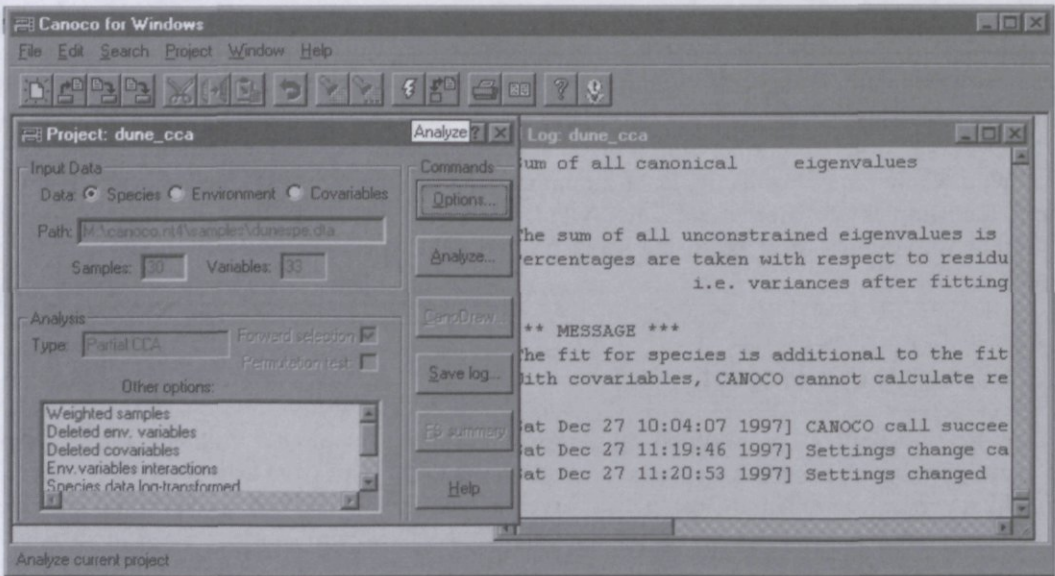


Figure 5-30 Canoco for Windows workspace after analysis.

You can save the log results using either the **Save Log...** button in the project view or by the **Save log...** command in the **File** submenu. You can print the analysis log using either the **Print** button in the program toolbar or using the **Print...** command in the **File** submenu (or using the **Ctrl-P** keyboard short-cut). Before printing the log, you can preview the output using the **Print Preview** facility available either via the **Print Preview** button in the program toolbar or with the **Print Preview...** command in the **File** submenu.

The log produced by the CANOCO has usually quite long lines. To fit these lines on the printed pages, it is advantageous to set the print orientation to the Landscape mode in the dialog box invoked by the **Page Setup...** command in the **File** submenu.

5.13 Creating ordination diagrams

After you finished the analysis of a CANOCO project, you can visually inspect the analysis results using the ordination diagrams created by the program **CanoDraw**. You can start the program using the **CanoDraw...** button in the **Project View** window. This button is enabled if the analysis results are available and the file with analysis results is younger than the file with Canoco project.

After you clicked the **CanoDraw...** button or used an alternative invocation method (either the **Run CanoDraw** command in the **Project** submenu or the **Ctrl-C** keyboard short-cut), **CanoDraw** is started and the current project settings are read together with the analysis solution file and with the input data files. **CanoDraw** then must save all this information into its own project file. Once the **CanoDraw** project was created, **CanoDraw** does not refer to the files used or created by the Canoco for Windows program. If you change Canoco project and do a new analysis, you must recreate the **CanoDraw** project with the new settings and results.

Note that Canoco for Windows merely starts the **CanoDraw** program, so that you can continue work with Canoco for Windows, while **CanoDraw** is running.

5.14 Specifying program options

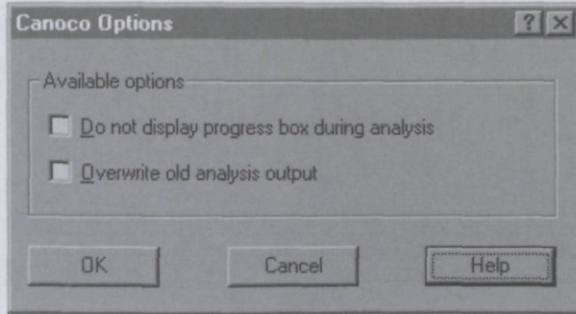


Figure 5-31 Canoco Options dialog.

You can use the **Options...** command in the **Project** submenu to specify options for the Canoco for Windows program (see Figure 5-31).

The first option, if checked, disables the creation of the progress window which is normally displayed while a CANOCO project is being analyzed. This possibility might come handy if you want to run a long analysis and work with other software during that period. Normally, when Canoco for Windows does long permutation tests, it brings to foreground and updates the progress window from time to time to confirm to the user that the Canoco application is still working. This might be a real nuisance, however, if you want to run the analysis on the background.

If the second option is specified, the old analysis output is deleted each time a new analysis log is being appended. This might be useful if you try to find proper ordination model and/or analysis options iteratively, and you are interested only in the final model.

6. Results of the analysis

6.1 Introduction

This chapter describes the numerical results of an ordination analysis by Canoco that are obtained after clicking the Analyze... button in Canoco for Windows. The results consist of three parts, listed in the log-window, the solution file and the species-environment table. The log-window provides summary statistics of the analysis, including the results of significance tests. Look also in the log-window for any possible warning and error messages. The solution file contains the ordination scores, and various other statistics that are linked to individual species, samples and, if available, environmental variables. This is the file that contains the information for drawing ordination diagrams and is thus used by CanoDraw. The species-environment table contains a table of correlation coefficients (or covariances) between species and environmental variables in linear ordination methods or a table of weighted averages in unimodal ordination methods. The subsequent sections describe each part of the output in turn.

This chapter is also relevant to you if you run the console version of Canoco (chapter 7). The results described in section 6.2 (Log window) can be found in the CANOCO output file.

6.2 Log-window

It is recommended to check in the log-window that the analysis has been carried out as planned. The first part gives essential information on the number of active samples, species, environmental variables, and covariables. Later parts give diagnostics on outliers in the environmental and covariable data and summary statistics of the ordination. The results are discussed in the order in which they are listed in the log-window. The example analyses are a canonical correspondence analysis (CCA) and a redundancy analysis (RDA) applied to the extended Dune Meadow data in Table 4.5 and Table 4.3 with the species numbered 31 - 33 made supplementary (see also Appendix A).

6.2.1 Log of reading of the project CON-file and the data files

Table 6.1 shows the first part of the log-window of a canonical correspondence analysis (CCA) applied to the Dune Meadow data in Table 4.5 and Table 4.3. Some of the output is a little bit cryptic because some options are indicated by number rather than by a description. For example, in this output, CCA is indicated by analysis type 5 and this is the chosen number (Answer = 5) and, below the listing of names of the data files, no forward selection has been chosen as indicated by a 0 and the chosen scaling of ordination scores is 2. Canoco for Windows always gives full diagnostics as indicated by a 1 or 3. The precise meaning of the codes can be found in chapter 7 on the console version of CANOCO. In most cases, the number reflects the order of options in the wizard pages. Table 6.2 and Table 6.3 explain the numerical codes of the scaling of ordination scores in terms of the choice you made in the wizard in unimodal and linear methods, respectively.

Table 6.1 First part of the log-window of a canonical correspondence analysis on the Dune Meadow data.

[Mon Jul 28 14:50:14 1997] Log file created

[Mon Jul 28 14:52:35 1997] CON file [C:\cfw\samples\ccaman1.con] saved

[Mon Jul 28 14:52:38 1997] Running CANOCO:

Program CANOCO Version 4.0 September 1997 - written by Cajo J.F. ter Braak
 Copyright (c) 1988-1997 Centre for Biometry Wageningen, CPRO-DLO
 Box 100, 6700 AC Wageningen, the Netherlands.
 CANOCO performs (partial) (detrended) (canonical) correspondence analysis,
 principal components analysis and redundancy analysis.
 CANOCO is an extension of Cornell Ecology program DECORANA (Hill,1979)

For explanation of the input/output see the manual, 'Unimodal models' and
 ter Braak, C.J.F. (1995) Ordination. Chapter 5 in:
 Data Analysis in Community and Landscape Ecology
 (Jongman, R.H.G., Ter Braak, C.J.F. and Van Tongeren, O.F.R., Eds),
 Cambridge University Press, Cambridge, UK, 91-173 pp.

*** Type of analysis ***

Model	Gradient analysis		
	indirect	direct	hybrid
linear	1=PCA	2= RDA	3
unimodal	4= CA	5= CCA	6
,,	7=DCA	8=DCCA	9
	10=non-standard analysis		

Type analysis number

Answer = 5

*** Data files ***

Species data : C:\cfw\samples\dunespe.dta

Covariable data :

Environmental data : C:\cfw\samples\duneenv.dta

Initialization file:

Forward selection of envi. variables = 0
 Scaling of ordination scores = 2
 Diagnostics = 3

File : C:\cfw\samples\dunespe.dta

Title : SPECIES - DUNE MEADOW DATA (M. BATTERINK AND G. WIJFFELS, 1983)

Format : (I5,9(I5,F2.0))

No. of couplets of species number and abundance per line : 9

No samples omitted

Number of samples 22

Number of species 33

Number of occurrences 226

File : C:\cfw\samples\duneenv.dta

Title : ENVIRONMENTAL DATA IN FULL FORMAT - DUNE MEADOW DATA

Format : (I5,F5.0,1X,2F3.0,3X,3F2.0,3X,4F2.0)

No. of environmental variables : 10

No interaction terms defined

No transformation of species data

Weight .00 is given to species 31

Weight .00 is given to species 32

Weight .00 is given to species 33

No sample-weights specified

No downweighting of rare species

No. of active samples: 20
 No. of passive samples: 2
 No. of active species: 30

Total inertia in species data=
 Sum of all eigenvalues of CA = 2.11526

***** Collinearity detected when fitting variable 6 *****
 ***** Collinearity detected when fitting variable 10 *****

Table 6.2 Codes for the options for the scaling of ordination scores in unimodal methods (CA, CCA and DCCA).

Focus scaling on	Scaling type	
	biplot scaling	Hill's scaling
Inter-sample distances	1	-1
Inter-species distances	2	-2
Symmetric	3	-3

Table 6.3 Codes for the options for the scaling of ordination scores in linear methods (PCA and RDA).

Focus scaling on	Species scores	
	Divide by standard deviation (Correlation biplot)	Do not post-transform (Covariance biplot)
Inter-sample distances	1	-1
Inter-species correlations	2	-2
Symmetric	3	-3

Among the list of names of data files is also the initialization file, if any. This file is useful in the console version of CANOCO (see section 7.3) but has little effect in Canoco for Windows as most options are fully specified by the Project Setup Wizard.

Table 6.1 also gives details on the data files, starting with the file specified as species data file and followed by the file specified as environmental data. Among other things, it is reported how many samples and species there are and how many values are non-zero (number of occurrences). After the report on the covariable (if any) and environmental file and any data transformations, there is a list of species and samples (if any) that are given a non-default weight. The species (or samples) that were deleted or were made supplementary are listed here with a weight of .00 as they have no influence on the analysis. In the example, the species numbered 31, 32, and 33 (Hip rha, Poa ann, and Ran acr) were made supplementary and thus receive zero weight. Then, the number of active samples and species is listed that jointly determine the ordination. Unless the user specifies otherwise, a sample is active when it occurs (1) with non-zero values for the active species in the species data and (2) in the environmental and covariable data files specified for the analysis. A sample is made supplementary (passive) when it is made so by the user, or when the sample is not encountered either in the file with environmental data or in the file with the covariables. In the example, the samples with numbers 17 and 20 do not occur in the environmental data and are therefore made supplementary by CANOCO. Active species are species that have non-zero values for active samples in the species

data and which are not deleted or made passive. The number of active species can be lower than the highest species number encountered in the species data, because some species may be absent in the data.

The total inertia in a unimodal method is the total variance in the species data as measured by the chi-square statistic of the sample-species table divided by the table's total (Greenacre 1984). The inertia is equal to the sum of all eigenvalues of CA. The value is repeated in the ordination summary of section 6.2.5. In linear methods (PCA and RDA), the total sum of squares (TSS) and the total standard deviation in the species data (TAU) are given here. The TSS and TAU are calculated after any data transformation and data standardization, using the notation of Table 6.24 on page 135,

$$(6.1) \quad TSS = \sum_i \sum_k y_{ik}^2 \quad \text{and} \quad TAU = \{ \sum_i \sum_k y_{ik}^2 / (nm) \}^{1/2}$$

User-specified weights for samples and/or species (if any) are used in the calculations. In a linear method, all species values are subsequently divided by the TAU. This has the advantage that the eigenvalues issued by PCA and RDA are fractions of the total sum of squares.

6.2.2 Collinear environmental variables and collinear covariables

At the bottom of Table 6.1 there are the messages

```
***** Collinearity detected when fitting variable    6 *****
***** Collinearity detected when fitting variable    10 *****
```

These messages indicates that the environmental variables numbered 6 and 10 are collinear with the environmental variables with lower numbers and, if present, the covariables. The code names of these variables are Pasture and NM, as can be deduced from Table 6.6 or any listing of environmental variables in the solution file and, with some more difficulty, from the data file in Table 4.3.

A variable is collinear if it can be written as a linear combination of the other variables. Collinear environmental variables occur always, when

- the environmental data contain nominal variables, or
- the number of active samples is less than the number of independent explanatory variables (covariables + environmental variables).

In other cases you may need to check your data or the options you chose. The environmental variables are perhaps more correlated than expected, possibly because fewer samples are active than you expected, or perhaps there are coding errors.

When the environmental data contain nominal variables, the variable indicating the last class or category is always collinear with the preceding classes. The variables Pasture and NM are, indeed, the last categories of the nominal variables Use and Management regime, respectively, in Table 4.3. As you can verify in Table 4.3, the value for NM is 1 minus the sum of the values of the other management categories; therefore NM is a collinear variable. Despite the collinearity message, the order of the class variables does not affect the ordination results, except for the regression coefficients, the corresponding t-values, and the t-value biplot scores in the solution file. The regression coefficients of collinear variables are set to 0, as are the corresponding t-values and t-value biplot scores.

Another message that may appear reads like:

```
***** Variable      10 has negligible variance      *****
***** (possibly after adjustment for covariables) *****
```

This message occurs when an environmental variable does not show variation in the data or, if there are covariables in the analysis, when the environmental variable is collinear with the covariables, for example, when a variable occurs in both the covariable and environmental data. Such environmental variables are automatically deleted from the analysis and do not appear in the solution file.

Collinear covariables are indicated by a message that reads like:

```
Covariable NM          is linearly dependent, hence ignored
```

6.2.3 Outliers in the explanatory variables: check on influence

Direct gradient analysis (canonical 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. CANOCO checks for each sample the leverage:

1. for each separate environmental variable.
2. in the space of the covariables, i.e. for all covariables jointly
3. in the joint space of the covariables and environmental variables, i.e. for all the covariables and environmental variables jointly.

Check 1, the check for separate environmental variables, detects univariate outliers. A sample is reported if it has more than five times the average leverage. Such samples have a value that is more than 3 standard deviations from the mean. This check is skipped for indicator variables (0/1-variables). In Table 6.4, an excerpt from the example CCA, the sample with identification number 15, is reported to have 8.2 times the average leverage for variable 1. This indicates that the value of the A1 horizon in this sample is an outlier. This value is 11.5 (Table 4.3). 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 from the mean.

The other two checks detect multivariate outliers. If a sample has more than three times the average leverage, then CANOCO reports the sample number and how many times the average its leverage is. In the example of Table 6.4 there are no multivariate outliers. They would have been reported below the headings Covariable influence (check 2) and +Environmental space influence (check 3).

It is important to remark that the leverage includes the (implicit) weights of samples. In linear methods all implicit weights are 1. In unimodal methods the implicit weight of a sample is the sample total (the row total across species). A sample may therefore be reported as an outlier simply because it has a large implicit weight. Even if there are no covariables in the analysis, a sample can be reported as an outlier in the covariable space, if it has an extreme sample total. The logic of this is that such a sample has much influence on the analysis. To exclude the sample weights from the influence check, carry out an RDA without any additional weighting of samples.

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 conclusions. But, always be hesitant to remove the sample in the analysis you report. More discussion on this topic can be found in any modern book on regression and on outliers.

Table 6.4 Regression diagnostics.

```

***** Check on influence in covariable/environment data *****
The following sample(s) have extreme values
Sample Environmental      Covariable + Environment space
variable Influence      influence      influence

      15      1      8.2x
***** End of check *****

```

6.2.4 Correlation matrix, means, standard deviations, inflation factors

With environmental data in the analysis, the log-window also contains a matrix of correlation coefficients such as shown in Table 6.5 and a list of means, standard deviations, and inflation factors as shown in Table 6.6. This output is TAB-delimited for transfer to word processors and spreadsheets. The variables in these tables are ordination axes, indicated with the labels “SPEC AX1” ... “SPEC AX4” and “ENVI AX1” ... “ENVI AX4”, and environmental variables, indicated with their code names. Each of the four ordination axes calculated by CANOCO is represented by two variables. The reason is that there are two set of samples scores in an analysis with environmental data: one set is derived from the species data and has the prefix “SPEC” and the other set is derived from the environmental data and has the prefix “ENV”. From Table 6.5 we see the ordination axes that are derived from the species (SPEC AX’s) show modest correlations among themselves in the range -0.05 — 0.13, whereas the ordination axes that are derived from the environmental data (ENVI AX’s) are mutually uncorrelated. If we would have applied an indirect gradient analysis instead, the SPEC AX’s would have been uncorrelated. The correlations between the SPEC AX’s and ENVI AX’s of the same axis number are called the species-environment correlations. These correlations can also be found in the summary of the ordination (section 6.2.5). The correlations between the SPEC AX’s and the environmental variables in the top part of Table 6.5 are called inter-set correlations, whereas those between the ENVI AX’s and the environmental variables are called the intra-set correlations. The bottom part of Table 6.5 shows the correlations among all environmental variables. For example, the correlation between A1 and Manure is -0.23.

In linear methods (PCA and RDA), the correlations are the product-moment correlation coefficients. With user-defined weights for samples, these weights are used in calculating the means, standard deviations, and correlation coefficients in the obvious way (Kendall & Stuart, 1973: p 301). In unimodal methods (CA, CCA, DCA, and DCCA), the sample total (y_{i+}) acts as a sample weight, even if default weights are used, giving weighted means, weighted standard deviations, and weighted correlation coefficients, as indicated in Table 6.5 and Table 6.6. With non-default weights $\{w_i^*\}$ and $\{w_k^*\}$, the applied weights are calculated as $w_i = w_i^* \sum_k w_k^* y_{ik}$.

If covariables are present, the correlations are adjusted for the covariables, i.e. they are partial correlations (Kendall & Stuart, 1973, Chapter 27). Partial correlations are calculated by regressing each of the environmental variables on to the covariables and by calculating the correlations among the residuals of these regressions.

In Table 6.6 we see that the standard deviation of each of the ENVI AX’s is 1. This reflects the choice in the scaling of ordination scores. In a direct gradient analysis with scaling 2 (Table 6.2) the variance of the ENVI AX scores is set equal to 1.

Table 6.5 Correlations among environmental variables and ordination axes.

**** Weighted correlation matrix (weight = sample total) ****

SPEC AX1	1.0000								
SPEC AX2	-.0387	1.0000							
SPEC AX3	.0773	-.0400	1.0000						
SPEC AX4	-.0506	.1273	-.1110	1.0000					
ENVI AX1	.9580	.0000	.0000	.0000	1.0000				
ENVI AX2	.0000	.9018	.0000	.0000	.0000	1.0000			
ENVI AX3	.0000	.0000	.8554	.0000	.0000	.0000	1.0000		
ENVI AX4	.0000	.0000	.0000	.8888	.0000	.0000	.0000	1.0000	
A1	.5392	-.1562	.5042	-.0972	.5629	-.1732	.5894	-.1094	
Moisture	.8833	-.1535	-.1199	.1507	.9221	-.1702	-.1402	.1696	
Manure	-.2962	-.6895	-.1687	-.1603	-.3092	-.7646	-.1972	-.1803	
Hayfield	-.0724	.5453	-.2158	.2508	-.0756	.6046	-.2523	.2821	
Haypastu	-.1647	-.4992	-.1129	-.0768	-.1719	-.5535	-.1320	-.0864	
Pasture	.2677	-.0281	.3660	-.1875	.2795	-.0312	.4279	-.2110	
SF	.1421	-.6273	-.3601	-.0768	.1484	-.6956	-.4210	-.0864	
BF	-.3491	.1578	-.0255	-.5195	-.3645	.1750	-.0298	-.5845	
HF	-.3459	-.1047	.3758	.4643	-.3611	-.1161	.4394	.5224	
NM	.5464	.6656	.0007	.0379	.5704	.7381	.0008	.0426	
	SPEC AX1	SPEC AX2	SPEC AX3	SPEC AX4	ENVI AX1	ENVI AX2	ENVI AX3	ENVI AX4	
A1	1.0000								
Moisture	.4154	1.0000							
Manure	-.2283	-.2204	1.0000						
Hayfield	-.1845	.0251	-.6118	1.0000					
Haypastu	.1588	-.1671	.4800	-.6023	1.0000				
Pasture	.0210	.1634	.1231	-.4096	-.4816	1.0000			
SF	.0768	.1595	.6838	-.4661	.5600	-.1283	1.0000		
BF	-.3069	-.3759	-.1809	.0277	-.0512	.0282	-.2956	1.0000	
HF	-.1444	-.1780	.1361	.0857	-.2581	.2008	-.4375	-.3049	1.0000
NM	.3551	.3641	-.7422	.3933	-.2831	-.1083	-.3463	-.2413	-.3049
	A1	Moisture	Manure	Hayfield	Haypastu	Pasture	SF	BF	HF
HF	1.0000								
NM	-.3572	1.0000							
	HF	NM							

Table 6.6 also shows a column head "inflation factor". It is the **Variance Inflation Factor (VIF)** of a variable in a multiple regression equation (Montgomery & Peck 1982: section 8.4.2). The name derives from the fact that the variances of estimated regression coefficients $\{c_j\}$ are proportional to their VIF's, namely

$$(6.2) \quad \text{var}(c_j) = \text{VIF}(\text{residual variance}) / (n - q - 1)$$

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

$$(6.3) \quad \text{VIF} = 1/(1 - R_j^2)$$

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

High VIF's indicate multicollinearity among the environmental variables. If an environmental variable is completely multicollinear, its VIF is set to 0, and its regression coefficient and associated t-value are set to 0. Normal VIF's are always greater than 1.0. For mutually uncorrelated environmental variables all VIF's are equal to 1.0, but this happens only in designed experiments. **If all VIF's are given as 1.0000, then CANOCO probably did not calculate them at all.**

Table 6.6 Means, standard deviations and inflation factors of environmental variables.

N	name	(weighted) mean	stand. dev.	inflation factor
1	SPEC AX1	.0000	1.0439	
2	SPEC AX2	.0000	1.1088	
3	SPEC AX3	.0000	1.1691	
4	SPEC AX4	.0000	1.1251	
5	ENVI AX1	.0000	1.0000	
6	ENVI AX2	.0000	1.0000	
7	ENVI AX3	.0000	1.0000	
8	ENVI AX4	.0000	1.0000	
1	A1	4.6850	1.8613	1.7814
2	Moisture	2.8015	1.7312	1.8500
3	Manure	1.9022	1.3629	8.3034
4	Hayfield	.3387	.4733	2.7057
5	Haypastu	.4146	.4927	2.2125
6	Pasture	.2467	.4311	.0000
7	SF	.2978	.4573	9.2126
8	BF	.1708	.3763	2.4671
9	HF	.3109	.4629	4.5651
10	NM	.2204	.4145	.0000

6.2.5 Summary of the ordination

6.2.5.1 Analyses without covariables.

Table 6.7 shows the summary of the ordination from the example CCA of the extended Dune Meadow data. Results are given for the first four ordination axes. By default, this output is TAB-delimited for optimal display in word processors and spreadsheets.

Table 6.7 Summary of a CCA of the Dune Meadow data.

**** 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 eigenvalues					2.115
Sum of all canonical eigenvalues					1.220

The **eigenvalues** measure the importance of each of the axes (values between 0 and 1). The first eigenvalue is 0.461, the second 0.298, and so on.

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 (see equation (6.38); Greenacre, 1984). The **total inertia** of the species data is 2.115 in Table 6.7. 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, as shown in Table 6.8, because the species data are scaled in this way (see page 124).

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 (see e.g. McCune 1997). The amount explained is given by the eigenvalue in constrained analyses (RDA/CCA) and by $r^2 \times$ eigenvalue in unconstrained analyses (PCA/CA) with r the species-environment correlation. The amounts of explained variance are given in the next row.

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 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 Unimodal Models p. 162). The total variance of the fitted values is precisely the sum of all canonical 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 canonical eigenvalues, e.g., for axis 2, $100 * (\lambda_1 + \lambda_2) /$ (sum of all canonical eigenvalues). In PCA/CA, the formula is a bit more difficult (as 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 environment-derived sample scores.

There exists another interpretation of the percentage variance of the species-environment relation. Linear relationships can be well summarized by correlation coefficients. In linear methods (PCA/RDA), the relationships between the species data and the environmental data can thus be summarized in a table of species by environmental variables with, as entries, the correlation between each particular species and each particular environmental variable (Unimodal Models p. 140: Fig. 1). Each axis explains a part of the variance in 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 be displayed in a two-dimensional biplot of species scores (adjusted for species variance) and the biplot scores of environmental variables. Each adjusted species score is divided by the standard deviation of the species. The biplot displays covariances, instead of correlations, if the species scores are not adjusted for species variance (i.e. if the species scores are not post-transformed; with scaling of ordination scores < 0 , see Table 6.3). With covariables in the analysis, partial covariances are displayed. In unimodal methods (CA/CCA/DCA), the relationships between the species data and the environmental data can be summarized by weighted averages of species with respect to environmental variables (Unimodal Models p. 158: Fig. 1). The entries of the table are thus weighted averages (instead of correlations, as in linear methods), but for the rest the interpretation is the same. The species-environment table itself is also produced by CANOCO (section 6.4; Table 6.60).

Correlation coefficients and weighted averages are good summaries of the species-environment relationship if the environmental variables are quantitative. When the environmental variables are nominal, class means and totals are appropriate summaries in linear methods and unimodal methods, respectively. Ter Braak (1994) and Ter Braak & Verdonschot (1995) showed that the percentages of explained variance of the species-environment relation also apply to tables of means and of totals, and also to tables of mixtures of correlations and class means and of weighted averages and class totals. The total weighted variance in the table is precisely the sum of all canonical eigenvalues. For the mathematical proof of this, see Unimodal Models (p 151) and, to extend the results to unimodal methods, see section 17.2.

Summarizing, we see in Table 6.7 that a CCA-triplot of samples, species and environmental variables based on the first two axes explains 35.9% of the variance (inertia) in the species data, 62.3% of the variance in the fitted species data, and the same percentage (62.3%) of the variance in the weighted averages and the class totals of the species with respect to the environmental variables.

From Table 6.8, a RDA-triplot of samples, species and environmental variables based on the first two axes explains 43.4 % of the variance in the species data, 69.2% of the variance in the fitted species data, and the same percentage (69.2%) of the variance in the correlations and the class means of species with respect to the environmental variables.

Table 6.8 Summary of a RDA of the Dune Meadow data.

**** Summary ****

Axes	1	2	3	4	Total variance
Eigenvalues	: .264	.170	.067	.041	1.000
Species-environment correlations	: .955	.899	.924	.797	
Cumulative percentage variance					
of species data	: 26.4	43.4	50.2	54.3	
of species-environment relation:	42.1	69.2	79.9	86.5	
Sum of all eigenvalues					1.000
Sum of all canonical eigenvalues					.628

Table 6.9 shows the summary of a DCA using detrending-by-segments. The summary has an additional line for the lengths of gradient. These are reported only if detrending-by-segments is requested. The length of gradient is a measure of how unimodal the species responses are along an ordination axis. It is the range of the sample scores divided by the average within-species standard deviation along the axis. The gradient length is expressed in standard deviation units of species turnover (SD). If a gradient length is over 4 SD, there are species in the data that show a clear unimodal response along the gradient. The first axis of the Dune Meadow data has a length of 3.402 SD. The Dune Meadow thus show a modest amount of unimodality. Note that the gradient lengths are not necessarily decreasing in value. For further information see Hill & Gauch (1980) and Jongman et al. (1987: 106).

Table 6.9 Summary of a DCA with detrending-by-segments (with interpretation by the environmental variables).

**** Summary ****

Axes	1	2	3	4	Total inertia
Eigenvalues	: .536	.256	.083	.035	2.115
Lengths of gradient	: 3.402	3.120	1.517	1.438	
Species-environment correlations	: .869	.855	.898	.703	
Cumulative percentage variance					
of species data	: 25.3	37.5	41.4	43.1	
of species-environment relation:	29.4	45.2	.0	.0	
Sum of all eigenvalues					2.115
Sum of all canonical eigenvalues					1.220

Sometimes, the summary of the ordination contains zeroes at places beyond the first axis, indicating that the values were not calculated for these later axes. Table 6.9 is a case in point: the values for the axes 3 and 4 for the cumulative percentage variance of the species-environment relation are zero. With detrending-by-segments, the environmental biplot scores (that are used to make inferences about the table of the weighted averages) depend on the dimension of the biplot³. By default, the dimension is set to two, which is fine for the usual ordination diagram of the second axis against the first axis. The remaining values are not calculated. The values for the first two axes would change if the dimension was set to four. The default value of two can be changed in the CANOCO initialization file (CANOCO.INI).

Two other cases of zeroes in the summary are:

- The first few axes are constrained, whereas the later axes are unconstrained. This happens in hybrid analysis and when the number of environmental variables is small (1–4). The percentages of variance for the species-environment relation will not be calculated for the unconstrained axes.
- With only a few species, all variation may be contained in less than four axes, resulting in zero eigenvalues.

Finally, it should be noted that the eigenvalues are not necessarily decreasing in value if the first few axes are constrained, whereas later axes are unconstrained.

³ The environmental biplot scores are calculated by a separate multivariate regression (Unimodal Models p. 72) of the table of weighted averages on the species scores. The biplot scores depend on the dimension of the biplot because the species scores of different axes are not orthogonal in detrending-by-segments.

6.2.5.2 Analyses with covariables: partial ordination

Table 6.10 shows a summary of a partial ordination, i.e. an ordination with covariables. The example is a CCA of the Dune Meadow data with the A1 horizon, Moisture and Manure used as covariables. We see from the table that the sum of all eigenvalues is no longer equal to the total inertia, because the covariables have already explained some of the inertia in species data, namely $2.115 - 1.346 = 0.769$. In a CCA with the A1 horizon, Moisture, and Manure as the only environmental variables, the sum of all canonical eigenvalues is indeed 0.769. The additional inertia explained by the other environmental variables is 0.450. Note that the sum of $0.769 + 0.450 = 1.219$, which is, apart from rounding errors, equal to the sum of all the canonical 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). Ecological applications of the decomposition are given by Borcard et al. (1992) and Økland & Eilertsen (1994). An example is given in section 8.3.1.2.

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. $100 * 0.166/1.346 = 12.3\%$. This is $100 * 0.166/0.450 = 36.9\%$ of what, in total, can be explained by the current environmental variables. One finds these percentages in the summary table (Table 6.10). A computational formula for the sum of all canonical eigenvalues is given in section 17.3.

Table 6.10 Summary of a partial CCA of the Dune Meadow data. Covariables are A1 horizon, Moisture, and Manure. Environmental variables are Use and Management regime.

**** 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 eigenvalues					1.346
Sum of all canonical eigenvalues					.450

The sum of all eigenvalues is after fitting covariables
 Percentages are taken with respect to residual variances
 i.e. variances after fitting covariables

6.2.6 Global permutation test

Table 6.11 summarizes the results of the global permutation tests to judge the significance of the relation between species and environment in the Dune Meadow data using CCA. The test of significance based on first canonical eigenvalue is reported first. The first canonical eigenvalue (cf. Table 6.7) is 0.461 and the F-ratio (calculated using equation (3.12)) is 3.067. The resulting P-value is 0.010, indicating that the first canonical axis is statistically significant at

the 1% level. Thereafter, the test based on the sum of all canonical eigenvalues (the trace) is reported. The trace is 1.220 (cf. Table 6.7) , leading to an F-ratio of 1.873 (calculated by equation (3.5)). The resulting P-value is 0.0050, demonstrating that the relation between the species and the environmental variables is highly significant ($P < 0.01$).

With the console version of CANOCO it is possible to obtain the results of individual permutations (Table 6.12). The results are displayed on the screen and have a didactic purpose only. With Table 6.12, we can explain the permutation test in some more detail. See also section 3.7.2 on page 43. The 20 active samples in the species data are randomly shuffled 199 times (unrestricted permutation), while keeping samples in the environmental data in place. Two test statistics, labeled F-ratio and F-ratio of axis 1, were calculated for the original unpermuted data and for each of the 199 permutations. The first F-ratio is based on the sum of all canonical eigenvalues and the second F-ratio (F-ratio of axis 1) is the F-ratio based on the first canonical eigenvalue. See section 3.7.5 for the precise definitions. Table 6.12 shows the resulting values for the original data (after "Data"), and for the first 5 and last 5 permutations. The reported P-value is the rank of the statistic for the data divided by the number of calculated values (the number of permutations plus one for the non-permuted data). The F-ratio for the data, 1.873, was the largest value, so that the P-value is $1/200 = 0.005$. The F-ratio of axis 1 for the data, 3.067, was the second largest value, so that the P-value is $2/200 = 0.01$, as reported at the bottom of Table 6.12. If only one of the F-ratios is calculated, the other is set to 0.00 and the corresponding P-value is set to 1.000. The results are summarized in the log-window as shown in Table 6.11.

"Permutation under the reduced model" (Table 6.12 and Table 6.11) means that the residuals from the reduced model have been shuffled. Without covariable data in the analysis, as in the example, the reduced model is the overall null model, i.e. the model without any explanatory variables, so that the residuals are the same as the raw data. With covariables, the reduced model contains the covariables, but not the environmental variables. The alternative is "Permutation under the full model". The full model contains both the covariables (if any) and the environmental variables.

Table 6.11 Summary of the global permutation test of the relation between species and environment in the Dune Meadow data using CCA.

**** Summary of Monte Carlo test ****

```

Test of significance of first canonical axis: eigenvalue =   .461
                                             F-ratio   =   3.067
                                             P-value   =   .0100

Test of significance of all canonical axes : Trace    =   1.220
                                             F-ratio   =   1.873
                                             P-value   =   .0050
    
```

(199 permutations under reduced model)

Table 6.12 Monte Carlo permutations to test the significance of the relation between species and environment in the Dune Meadow data using CCA.

*** Unrestricted permutation ***

Seeds: 23239 945

Number of permutations= 199

*** Permutation under reduced model ***

No	F-ratio	F-ratio of axis 1
Data	1.873	3.067
1	.923	1.853
2	1.134	2.317
3	.917	1.981
4	1.406	2.636
5	.765	1.364

..... Permutations 6 - 194 not shown

195	1.226	2.653
196	.934	1.477
197	.921	1.654
198	.922	1.373
199	.761	1.530

P-value .0050 .0100 (number of permutations= 199)

If you specified blocks to exclude exchanges of samples between blocks, the log-window lists the samples in each block by their identification number. An example is shown in Table 6.13. In this table, the last category of Management type is not listed among the block-defining covariables, because it is redundant for specifying the block structure. If you specified a split-plot design in blocks, the log-window lists, per block, the samples in each whole-plot and how whole-plots and split-plots are being permuted in the test (Table 6.14).

Table 6.13 Blocks defined by Management type in the Dune meadow data.

```
*** Specification of blocks ***  
  
*** The permutations are conditioned on      3 covariable(s), namely: ***  
  
covariable SF  
covariable BF  
covariable HF  
  
*** Sample arrangement in the permutation test ***  
  
Samples in block   1 :  
  1   3   4   12   13   16  
The      6 plots are permuted completely at random  
  
Samples in block   2 :  
  2  10  11  
The      3 plots are permuted completely at random  
  
Samples in block   3 :  
  5   6   7   8   9  
The      5 plots are permuted completely at random  
  
Samples in block   4 :  
 14  15  17  28  29  30  
The      6 plots are permuted completely at random
```

Table 6.14 The second block of a split-plot design containing 6 whole-plots with 4 split-plots each.

```
Samples in block   2 :  
Whole plot   1 :  
 25  26  27  28  
Whole plot   2 :  
 29  30  31  32  
Whole plot   3 :  
 33  34  35  36  
Whole plot   4 :  
 37  38  39  40  
Whole plot   5 :  
 41  42  43  44  
Whole plot   6 :  
 45  46  47  48  
  
These      6 whole plots are permuted completely at random  
The        4 split plots are not permuted
```

6.2.7 Forward selection of environmental variables

In Canoco for Windows, the results of an automatic forward selection are summarized in two tables of marginal and conditional effects, accessible from the Project View as described in section 5.8.2. Examples are given in sections 8.2.9 and 8.3.4. In this section we describe the full results of forward selection as given in the log-window. We do this by presenting an example of a manual forward selection using a CCA of the extended Dune Meadow data with the Monte Carlo permutation box checked. The essential results are also displayed on screen during the manual

selection process in Canoco for Windows. In the console version of CANOCO, the full results are displayed both on screen and in the output file.

After the check of influence (section 6.2.3), the first step of the selection process is reported (Table 6.15).

Table 6.15 Step 1 in manual forward selection of the Dune Meadow data: the marginal effects of the environmental variables.

```

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

*** Unrestricted permutation ***

Seeds: 23239 945

  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
Environmental variable      2 added to model
Variance explained by the variables selected:      .41
"      "      "      all variables      :      1.22

```

In each step, 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 was the only environmental variable. The same list could thus be obtained manually in ten runs of CANOCO, each run with another environmental variable. The effects shown in the first step are called marginal effects. In the first step the variable with the highest extra fit, in our case, Moisture (variable 2) was included (added) in the model. The inertia explained by this variable is 0.41. If all variables would be included, the explained inertia would be 1.22. This value is the sum of all canonical eigenvalues (Table 6.7).

Table 6.16 Step 2 in a manual forward selection of the Dune Meadow data with Moisture already selected.

```

  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
Environmental variable      3 added to model
Variance explained by the variables selected:      .64
"      "      "      all variables      :      1.22

```


With moisture already selected (Table 6.16), 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 variance (the sum of all canonical eigenvalues) of this CCA is 0.64, which is 0.23 more than with Moisture alone.

For illustration Manure was included to the model, instead of the variable with the highest extra fit (NM). The inertia explained by the selected variables, Moisture and Manure, is indeed 0.64. After inclusion of Manure, NM is no longer the best variable to add, as we see in Table 6.17.

Table 6.17 Step 3 of a manual selection of the Dune Meadow data after Moisture and Manure have been selected.

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

This can be explained by noting that the Nature Management meadows do not receive manure, so that the variable NM can largely replace the variable Manure in the model (Jongman et al 1987: 54-55). The best variable to add is now A1. In the example run, the additional effect of the variable A1 on the species is tested at this point for its statistical significance. The test is reported in the log-window as shown in Table 6.18.

Table 6.18 A significance test in forward selection.

```

Environmental variable      1 tested
Number of permutations=    199

*** Permutation under reduced model ***

P-value   .100 (variable   1; F-ratio=  1.55; number of permutations=  199)

```

Because the additional effect of the best variable (A1) is not significant at the conventional 5%-level, it was decided to stop adding more variables. 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.

There may be covariables in the analysis at the start of the forward selection. The extra fit is calculated in precisely the same way as described above.

Warning: Significance tests in forward selection are often too liberal. If none of a large number of variables has a real effect, the reported P-value of the best variable in the forward selection may be well below the conventional 5% level, just because of the selection! Bonferoni-type adjustments may help to solve this problem (Miller 1990, Legendre & Legendre 1998).

6.3 Solution file

6.3.1 Introduction

The solution file contains a series of tables with scores on the four ordination axes for species, samples, environmental variables, and supplementary environmental variables, along with summary information per item. The solution file is used by CanoDraw to produce ordination diagrams. To inspect a solution file and to make simple scatter plots yourself, import the solution file into a spreadsheet program as a tab-delimited (Windows) ASCII text file. The layout of the tables is identical: a heading for the analysis (e.g. Table 6.19) and, as shown in Table 6.20, followed by a heading for the type of item (e.g. Spec: Species scores), labels for the columns, a summary statistic per column (e.g. EIG) and then per row item, an identification number, a code name, and four values (one for each of the first four ordination axes), and, sometimes, two additional columns with statistics per row item. Strict zero columns in the solution file indicate that the corresponding entries are not calculated.

The heading for the analysis differs slightly between methods as shown in Table 6.19 - Table 6.21. In brief,

1. the first line of the heading is the title of analyzed species data file.
2. the second line says which ordination method was applied, how many ordination axes are canonical (i.e. constrained by environmental variables), how many independent covariables there are, and a code for the scaling that is applied to the ordination scores. The codes are explained in Table 6.2 and Table 6.3 on page 117.
3. the third line gives the types of centering and standardization that are applied to the species data in linear methods. In unimodal methods with detrending the detrending options are reported.

The example in Table 6.19 is a redundancy analysis (RDA) on the Dune Meadow data with 4 canonical axes, no covariables, and scaling type 2 (i.e. focus on inter-species correlations in which the species scores are divided by their standard deviation). The species data are neither centered nor standardized by samples. The species data are centered by species but not standardized by species.

Table 6.20 is a heading in a canonical correspondence analysis (CCA). The codes for scaling type are explained in Table 6.2. The third line is blank.

Table 6.21 is a heading in an analysis that used detrending-by-segments. The detrending options in Table 6.21 are the default values of detrending-by-segments: 4 iterations are used for nonlinear rescaling of the axes scores, 26 segments are used in the detrending process, and axes are always non-linearly rescaled, whatever their gradient length. These options are further explained in section 7.6 and in Hill (1979). Detrending by polynomials is abbreviated to "DETR-POLY3", the "3" indicating that third-order polynomials were used in the detrending.

Table 6.19 Heading of each table in linear methods.

SPECIES - DUNE MEADOW 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

Table 6.20 Heading of each table in unimodal methods, followed by a table of species scores.

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

No transformation

Spec: Species scores (Biplot scaling)

N	NAME	AX1	AX2	AX3	AX4	WEIGHT	N2
	EIG	.4612	.2981	.1601	.1337		
1	Ach mil	-.8402	.3816	.0276	-.3341	16.00	6.10
2	Agr sto	.7704	-.5000	-.1143	-.0801	48.00	9.14
3	Air pra	.7395	1.7874	-1.0769	.5318	5.00	1.92
4	Alo gen	.3541	-.9700	-.3470	.1389	36.00	6.61

Table 6.21 Heading of each table with detrending-by-segments.

SPECIES - DUNE MEADOW DATA (M. BATTERINK AND G. WIJFFELS, 1983)
 DCA Canonical axes: 0 Covariables: 0 Scaling: -1
 DETR-SEGME Rescaling: 4 Segments: 26 Threshold: .00
 No transformation

Table 6.22 lists all possible tables in order of their appearance in a solution file. The code of each table uniquely identifies the table. The description of each item may vary slightly depending on the analysis type. The symbols are those used in equations in this manual. The symbol u_k is reserved for species scores in unimodal methods where it has the same dimension as the sample scores. The symbol b_k is used for species scores in linear methods, but may occasionally be used in unimodal methods when these are recast in a linear context. The last two columns indicate the content of the two columns that appear after the four columns of scores for ordination axes. The solution file may contain less tables than listed here. For example, without environmental variables in the analysis the tables numbered 7-14 are missing. From the tables with ordination diagnostics (numbered 3 - 6), the tables with species tolerances and sample heterogeneities are calculated for unimodal methods only, whereas the tables with species fits and residual lengths are not calculated in analyses that use detrending-by-segments. The residual lengths are not available in analyses with covariables.

Table 6.22 The order of tables in the solution file with their codes and symbols.

N	Code	Item	Symbol	Extra	columns
1	Spec:	Species scores	b_k or u_k	weight w_k	1 or N2
2	Samp:	Sample scores that are derived from the species	x_i^*	weight w_i	1 or N2
3	Tol:	Species tolerances	t_k	RMSTOL	N_2
4	Het:	Sample heterogeneities	h_k	RMSTOL	N_2
5	CFit:	Cumulative fit per species	$\text{var}(f_k)$	$\text{var}(y)$	%EXPL
6	SqRL:	Residual lengths per sample	length_i	SQLENG	%FIT
7	Regr:	Regression/Canonical coefficients	c_j		

N	Code	Item	Symbol	Extra	columns
8	tVal:	t-values of regression coefficients			
9	StBi:	Species coordinates for t-value biplot			
10	EtBi:	Environmental coordinates for t-value biplot			
11	CorE:	Inter-set correlations of environmental variables			
12	BipE:	Biplot scores of environmental variables	c_j^*		
13	CenE:	Centroids of environmental classes	c_j^-		
14	SamE:	Sample scores that are derived from the environment	x'_i	weight w_i	%FIT

The results for supplementary species and samples are listed among those for the active ones. The results for supplementary environmental variables, if specified, are given after all the tables of the active variables. The tables given for supplementary environmental variables are like those numbered 7 - 14 in Table 6.22, but have a negative value for the number of canonical axes (Table 6.23). In an indirect analysis, there is no theoretical distinction between normal environmental variables and supplementary ones; in the output there is no distinction either because the number of canonical axes is 0.

Table 6.23 Heading of a table for supplementary environmental variables.

```
SPECIES - DUNE MEADOW 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
```

The tables in the solution file are by default tab-delimited, i.e. the values are separated by tabs, and can therefore easily be copied via the Clipboard into spreadsheets or tables of word processors. In addition there are one or more spaces around most values. The tabs can be replaced by spaces or other delimiters (e.g. comma's) by modifying the CANOCO initialization file.

6.3.2 Relationships between ordination scores

This section lists the relationships between species, sample, and environmental scores per ordination axis, in the case without user-defined weights for samples and species. In later sections, more detail is given and formulae are generalized to include user-defined weights. We use the notation of Table 6.22 and Table 6.24. Biplots display approximate values of data tables (see also section 3.5). The approximation is given in algebraic form, based on scores of the first two ordination axes. For this, the scores are also subscripted with an index for the axis. For example, the species score is indicated by b_k in Table 6.22, so that b_{k2} indicates the k^{th} species' score on the second axis.

Table 6.24 Notation for input data and eigenvalues.

y_{ik}	= value of species k in sample i ($i = 1, \dots, n$; $k = 1, \dots, m$). In CA, DCA, CCA, and DCCA, $y_{ik} > 0$. In PCA and RDA the data $\{y_{ik}\}$, which may also be negative, is divided by the total standard deviation in the species data (TAU, page 118) after optional centering and/or standardization by species and/or by samples.
p_{ik}	= $y_{ik} y_{++} / (y_{i+} y_{+k})$, a transformation used in CA and CCA, with $y_{i+} = \sum_k y_{ik}$, the sample total, $y_{+k} = \sum_i y_{ik}$, the species total, and $y_{++} = \sum_i y_{i+} = \sum_k y_{+k}$, the overall total.
z_{ij}	= value of environmental variable j in sample i after centering and standardization ($i = 1, \dots, n$; $j = 1, \dots, q$). With covariables in the analysis, the environmental data are regressed on the covariables; the values $\{z_{ij}\}$ then denote the residuals of this regression.
z_{ij}	= indicator variable for environmental class j ($z_{ij} = 1$ if sample i belong to the j^{th} class, $z_{ij} = 0$ otherwise).
λ	= eigenvalue of the ordination axis.

6.3.2.1 Principal Components Analysis (PCA)

In PCA a reciprocal regression relation holds true between the species scores and the species-derived sample scores (Table 6.25): the species score b_k is the slope coefficient of the simple regression of the data of the k^{th} species on the sample scores $\{x_i^*\}$ and, the other way round, the sample score x_i^* is the slope coefficient of the simple regression of the data of the i^{th} sample on the species scores $\{b_k\}$. These regression relations are the key to the interpretation of the PCA-biplot: the species and sample scores together form a biplot that displays approximate species value y_{ik} as indicated as two-dimensional approximation in Table 6.25 (see also section 3.5).

Table 6.25 Transition formulae and scaling in PCA.

PCA with focus on inter-sample distances (scaling -1)	
$b_k = \sum_i y_{ik} x_i^* / \sum_i x_i^{*2} = \lambda^{-1} \sum_i y_{ik} x_i^* / n$	with $\sum_k b_k^2 / m = 1$
$x_i^* = \sum_k y_{ik} b_k / \sum_k b_k^2 = \sum_k y_{ik} b_k / m$	with $\sum_i x_i^{*2} / n = \lambda$
\Rightarrow biplot: $y_{ik} \approx b_{k1} x_{i1}^* + b_{k2} x_{i2}^*$	
PCA with focus on inter-species correlations (scaling -2)	
$b_k = \sum_i y_{ik} x_i^* / \sum_i x_i^{*2} = \sum_i y_{ik} x_i^* / n$	with $\sum_k b_k^2 / m = \lambda$
$x_i^* = \sum_k y_{ik} b_k / \sum_k b_k^2 = \lambda^{-1} \sum_k y_{ik} b_k / m$	with $\sum_i x_i^{*2} / n = 1$
\Rightarrow biplot: $y_{ik} \approx b_{k1} x_{i1}^* + b_{k2} x_{i2}^*$	
If the species scores have been divided afterwards by the standard deviation (sd_k) of each species (scaling +1 and +2), the symbol b_k in the above formulae must be replaced by $sd_k b_k$ where the latter b_k is the adjusted score given by CANOCO.	
For the adjusted species scores:	
$b_k = \sum_i \{y_{ik}/sd_k\} x_i^* / \sum_i x_i^{*2}$	
\Rightarrow biplot: $y_{ik} / sd_k \approx b_{k1} x_{i1}^* + b_{k2} x_{i2}^*$	
The adjusted species score b_k is the correlation of the k^{th} species with $\{x_i^*\}$ if the focus is on inter-species correlations (scaling +2)	
$\sum_i x_i^* / n = 0$ (zero mean sample score) if the species data are centered by species ($y_{+k} = 0$)	

The scaling of scores is also given in Table 6.25. With the focus on inter-sample distances (scaling -1), the mean square of the sample scores along an ordination axis is equal to the eigenvalue of the axis, the mean square of the (unadjusted) species scores is equal to 1, and the sample scores are a weighted sum of the species scores (divided by m). With the focus on inter-species correlations but without post-transformation of species scores (scaling -2), the mean square of the species scores along an ordination axis is equal to the eigenvalue of the axis, the mean square of the sample scores is equal to 1 and the species scores are a weighted sum of the sample scores (divided by n). The consequences of these two scaling types for the interpretation of distances between sample points and correlations between species arrows in the ordination diagram can be seen from the eigenvalue equations of PCA (Table 6.26). In scaling ± 1 , the score for a particular sample is the slope coefficient of the simple regression of the inter-sample inner products on all sample scores. Sample points thus approximate inter-sample inner products. Because of the relation between distances and inner products, distances between sample points in the ordination diagram approximate in scaling ± 1 the inter-sample Pythagorean distances. In scaling -2, the score for a particular species is the slope coefficient of the simple regression of the inter-species covariances on all species scores. Therefore the species points together form a biplot that displays approximate inter-species covariances. In scaling +2, the score for a particular species is the slope coefficient of the weighted regression of the inter-species

correlations on all species scores, the weights in this regression being the species variances. Therefore the species points together form a biplot that displays approximate inter-species correlations.

If the species scores are post-transformed (by division by the standard deviation; scaling +1 or +2), the adjusted species score reported by CANOCO is the slope coefficient of the simple regression of the standardized data $\{y_{ik}/sd_k\}$ of the k^{th} species on the sample scores $\{x_i^*\}$. When, at the same time, the focus is on inter-species correlations (scaling +2), the adjusted species score has a special meaning: it is the correlation of the k -th species with the ordination axis (i.e. with the species-derived sample scores). There is no simple expression for the mean square of the adjusted species scores. The sample scores are unaffected by the post-transformation of species scores.

If environmental variables are present in the analysis by PCA (in any scaling), regression coefficients $\mathbf{c} = (c_1, \dots, c_q)^T$ and environment-derived sample scores are calculated after the ordination has been obtained:

$$\mathbf{c} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{x}^*$$

$$x'_i = \sum_j c_j z_{ij}$$

The regression coefficients $\{c_j\}$ are partial regression coefficients from the multiple regression of $\{x_i^*\}$ on the q environmental variables z_1, \dots, z_q . In contrast, the environmental biplot scores $\{c_j^*\}$ are regression coefficients from the simple regression of x_i^* on the j^{th} environmental variables z_j (Table 6.29).

Table 6.26 Eigenvalue equations and scaling in PCA.

In PCA, the eigenvalue equation for the sample scores is

$$\lambda x_i^* = \sum_j c_{ij} x_j^* / n, \text{ where } c_{ij} = \sum_k y_{ik} y_{jk} / m, \text{ the inner product between samples } i \text{ and } j$$

With focus on inter-sample distances (scaling ± 1), $\sum_i x_i^{*2} / n = \lambda$ so that

$$x_i^* = \lambda^{-1} \sum_j c_{ij} x_j^* / n = \sum_j c_{ij} x_j^* / \sum_j x_j^{*2}$$

$$\Rightarrow \text{biplot: } c_{ij} \approx x_{i1}^* x_{j1}^* + x_{i2}^* x_{j2}^*$$

\Rightarrow approximation of inter-sample Pythagorean distances:

$$d_{ij} \approx \{(x_{i1}^* - x_{j1}^*)^2 + (x_{i2}^* - x_{j2}^*)^2\}^{1/2}$$

$$\text{with } d_{ij}^2 = \sum_k (y_{ik} - y_{jk})^2 / m = c_{ii} + c_{jj} - 2c_{ij}$$

In PCA, the eigenvalue equation for the species scores is

$$\lambda b_k = \sum_l c_{kl} b_l / m, \text{ where } c_{kl} = \sum_i y_{ik} y_{il} / n, \text{ the covariance between species } k \text{ and } l$$

With focus on inter-species correlations (scaling -2), $\sum_k b_k^2 / m = \lambda$ so that

$$b_k = \lambda^{-1} \sum_l c_{kl} b_l / m = \sum_l c_{kl} b_l / \sum_l b_l^2$$

$$\Rightarrow \text{biplot of inter-species covariances: } c_{kl} \approx b_{k1} b_{l1} + b_{k2} b_{l2}$$

If, with focus on inter-species correlations, the species scores have been divided afterwards by the standard deviation (sd_k) of each species (scaling +2),

$$b_k = \sum_l w_l r_{kl} b_l / \sum_l w_l b_l^2 \text{ with } w_l = sd_l^2$$

where $r_{kl} = c_{kl} / (sd_k sd_l)$, the correlation between species k and l

$$\Rightarrow \text{biplot of inter-species correlations: } r_{kl} \approx b_{k1} b_{l1} + b_{k2} b_{l2}$$

$\sum_i x_i^* / n = 0$ (zero mean sample score) if the species data are centered by species ($y_{+k} = 0$)

6.3.2.2 Redundancy Analysis (RDA)

Table 6.27 specifies the relations between scores in RDA. The species scores have a simple relation with the environment-derived sample scores: the species score b_k is the slope coefficient of the simple regression of the data of the k^{th} species on the environment-derived sample scores $\{x'_i\}$. The species-derived sample score x_i^* is the slope coefficient of the simple regression of the data of the i^{th} sample on the species scores $\{b_k\}$. The canonical coefficients $\{c_j\}$ are the partial regression coefficients from the multiple regression of $\{x_i^*\}$ on the q environmental variables z_1, \dots, z_q , and, finally, the environment-derived sample scores are a linear combination of the q environmental variables z_1, \dots, z_q . The simple regressions each lead to a biplot approximation as indicated in Table 6.27. In addition, the last row of Table 6.27 says that the environment-derived sample scores have a PCA-like relation with the species scores: the environment-derived sample score x'_i is the slope coefficient of the simple regression of the fitted values $\{\hat{y}_{ik}\}$ of the i^{th} sample on the species scores $\{b_k\}$. The biplot of species and environment-derived sample scores thus displays approximate fitted species values $\{\hat{y}_{ik}\}$.

The scaling of scores is also given in Table 6.27. With the focus on inter-sample distances (scaling -1), the mean square of the environment-derived sample scores along an ordination axis is equal to the eigenvalue of the axis, the mean square of the (unadjusted) species scores is equal to 1, and the species-derived sample scores are a weighted sum of the species scores (divided by m). With the focus on inter-species correlations but without post-transformation of species scores (scaling -2), the mean square of the species scores along an ordination axis is equal to the eigenvalue of the axis, the mean square of the environment-derived sample scores is equal to 1 and the species scores are a weighted sum of the environment-derived sample scores (divided by n). The consequences of these two scaling types for the interpretation of distances between points and correlations between species arrows in the ordination diagram can be seen from the eigenvalue equations of RDA (Table 6.28). In scaling ± 1 , the score for a particular sample is the slope coefficient of the simple regression of the fitted inter-sample inner products on all sample scores. Sample points thus approximate fitted inter-sample inner products. Because of the relation between distances and inner products, distances between sample points in the ordination diagram approximate in scaling ± 1 the fitted inter-sample Pythagorean distances. In scaling -2, the score for a particular species is the slope coefficient of the simple regression of the fitted inter-species covariances on all species scores. Therefore the species points together form a biplot that displays approximate fitted inter-species covariances. In scaling +2, the score for a particular species is the slope coefficient of the weighted regression of the fitted inter-species correlations on all species scores, the weights in this regression being the species variances. Therefore the species points together form a biplot that displays approximate fitted inter-species correlations.

If the species scores are post-transformed (by division by the standard deviation; scaling +1 or +2), the adjusted species score reported by CANOCO is the slope coefficient of the simple regression of the standardized data $\{y_{ik}/sd_k\}$ of the k^{th} species on the sample scores $\{x'_i\}$. When, at the same time, the focus is on inter-species correlations (scaling +2), the adjusted species score has a special meaning: it is the correlation of the k -th species with the ordination axis (i.e. with the environment-derived sample scores).

Table 6.27 Transition formulae and scaling in RDA.

RDA with focus on inter-sample distances (scaling -1)	
$b_k = \sum_i y_{ik} x'_i / \sum_i x_i'^2 = \lambda^{-1} \sum_i y_{ik} x'_i / n$	with $\sum_k b_k^2 / m = 1$
$x_i^* = \sum_k y_{ik} b_k / \sum_k b_k^2 = \sum_k y_{ik} b_k / m$	
$c = (Z^T Z)^{-1} Z^T x^*$	
$x'_i = \sum_j c_j z_{ij}$	with $\sum_i x_i'^2 / n = \lambda$
⇒ biplot: $y_{ik} \approx \hat{y}_{ik} \approx b_{k1} x'_{i1} + b_{k2} x'_{i2}$	
⇒ biplot: $y_{ik} \approx b_{k1} x_{i1}^* + b_{k2} x_{i2}^*$	
RDA with focus on inter-species correlations (scaling -2)	
$b_k = \sum_i y_{ik} x'_i / \sum_i x_i'^2 = \sum_i y_{ik} x'_i / n$	with $\sum_k b_k^2 / m = \lambda$
$x_i^* = \sum_k y_{ik} b_k / \sum_k b_k^2 = \lambda^{-1} \sum_k y_{ik} b_k / m$	
$c = (Z^T Z)^{-1} Z^T x^*$	
$x'_i = \sum_j c_j z_{ij}$	with $\sum_i x_i'^2 / n = 1$
⇒ biplot: $y_{ik} \approx \hat{y}_{ik} \approx b_{k1} x'_{i1} + b_{k2} x'_{i2}$	
⇒ biplot: $y_{ik} \approx b_{k1} x_{i1}^* + b_{k2} x_{i2}^*$	
If the species scores have been divided afterwards by the standard deviation (sd_k) of each species (scaling +1 and +2), the symbol b_k in the above formulae must be replaced by $sd_k b_k$ where the latter b_k is the adjusted score given by CANOCO.	
For the adjusted species scores:	
$b_k = \sum_i \{y_{ik}/sd_k\} x'_i / \sum_i x_i'^2$	
⇒ biplot: $y_{ik} / sd_k \approx \hat{y}_{ik} / sd_k \approx b_{k1} x'_{i1} + b_{k2} x'_{i2}$	
The adjusted species score b_k is the correlation of the k^{th} species with $\{x'_i\}$ if the focus is on inter-species correlations (scaling +2)	
Relation of x'_i to the species scores (simple regression of fitted abundance values on $\{b_k\}$):	
$x'_i = \sum_k \hat{y}_{ik} b_k / \sum_k b_k^2$	
with \hat{y}_{ik} the fitted value for the k^{th} species in the i^{th} sample based on multiple regression on the q environmental variables, i.e.	
$\hat{y}_k = Z (Z^T Z)^{-1} Z^T y_k$	
⇒ biplot: $\hat{y}_{ik} \approx b_{k1} x'_{i1} + b_{k2} x'_{i2}$	
$\sum_i x'_i / n = \sum_i x_i^* / n = 0$ (zero mean sample score)	

Table 6.28 Eigenvalue equations and scaling in RDA.

<p>In RDA, the eigenvalue equation for the sample scores is</p> $\lambda x'_i = \sum_j c_{ij} x'_j / n, \text{ where } c_{ij} = \sum_k \hat{y}_{ik} \hat{y}_{jk} / m, \text{ the fitted inner product between samples } i \text{ and } j$ <p>With focus on inter-sample distances (scaling ± 1), $\sum_i x'^2_i / n = \lambda$ so that</p> $x'_i = \lambda^{-1} \sum_j c_{ij} x'_j / n = \sum_j c_{ij} x'_j / \sum_j x'^2_j$ <p>\Rightarrow biplot: $c_{ij} \approx x'_{i1} x'_{j1} + x'_{i2} x'_{j2}$</p> <p>$\Rightarrow$ approximation of fitted inter-sample Pythagorean distances:</p> $d_{ij} \approx \{(x'_{i1} - x'_{j1})^2 + (x'_{i2} - x'_{j2})^2\}^{1/2}$ <p>with $d_{ij}^2 = \sum_k (\hat{y}_{ik} - \hat{y}_{jk})^2 / m = c_{ii} + c_{jj} - 2c_{ij}$</p>
<p>In RDA, the eigenvalue equation for the species scores is</p> $\lambda b_k = \sum_l c_{kl} b_l / m, \text{ where } c_{kl} = \sum_i \hat{y}_{ik} \hat{y}_{il} / n, \text{ the fitted covariance between species } k \text{ and } l$ <p>With focus on inter-species correlations (scaling -2), $\sum_k b_k^2 / m = \lambda$ so that</p> $b_k = \lambda^{-1} \sum_l c_{kl} b_l / m = \sum_l c_{kl} b_l / \sum_l b_l^2$ <p>\Rightarrow biplot of fitted inter-species covariances: $c_{kl} \approx b_{k1} b_{l1} + b_{k2} b_{l2}$</p> <p>If, with focus on inter-species correlations, the species scores have been divided afterwards by the standard deviation (sd_k) of each species (scaling +2),</p> $b_k = \sum_l w_l r_{kl} b_l / \sum_l w_l b_l^2 \text{ with } w_l = sd_l^2$ <p>where $r_{kl} = c_{kl} / (sd_k sd_l)$, the fitted correlation between species k and l</p> <p>\Rightarrow biplot of fitted inter-species correlations: $r_{kl} \approx b_{k1} b_{l1} + b_{k2} b_{l2}$</p> <p>$\sum_i x'_i / n = 0$ (zero mean sample score) if the species data are centered by species ($y_{+k} = 0$)</p>

The environmental biplot scores $\{c_j^*\}$ are related to both the sample scores and the species scores (Table 6.29). The score c_j^* is the slope coefficient from the simple regression of $\{x'_j\}$ on the j^{th} environmental variable z_j (Table 6.29). When the environmental biplot scores are plotted

as arrows, the arrow for each environmental variable points in the direction that any particular sample point would move if that variable would increase in value (ignoring the other variables). To define the relation with the species scores, let, as in Table 6.29, r_{jk} be the covariance between the k^{th} species and the j^{th} environmental variable. Then, c_j^* is the slope coefficient of the simple regression of the covariances $\{r_{jk}\}$ of all species with the j^{th} environmental variable on the species scores (Table 6.29). In a biplot with the species scores, the environmental biplot score for a particular variable thus approximates the covariances of the species with this environmental variable. If the species scores are adjusted (scaling +1 or +2), it is the correlations that are approximated instead of the covariances.

Table 6.29 The environmental biplot scores $\{c_j^*\}$ in PCA and RDA.

<p>Relation to the sample scores (simple regression of sample scores on z_j):</p> $c_j^* = \sum_i z_{ij} x_i^* / \sum_i z_{ij}^2 = \sum_i z_{ij} x'_i / \sum_i z_{ij}^2 = \sum_i z_{ij} x'_i / n$ <p>\Rightarrow predict change in x_i^* and/or x'_i due to change in z_j from c_j^*</p>
<p>Relation to the unadjusted species scores (simple regression of covariances on the species scores):</p> $c_j^* = \sum_k r_{jk} b_k / \sum_k b_k^2$ <p>where r_{jk} is the covariance between the k^{th} species and the j^{th} environmental variable</p> $r_{jk} = \sum_i z_{ij} y_{ik} / n$ <p>\Rightarrow biplot: $r_{jk} \approx b_{k1} c_{j1}^* + b_{k2} c_{j2}^*$</p>
<p>For the adjusted species scores (scaling +1 and +2):</p> <p>\Rightarrow biplot: $r_{jk} \approx b_{k1} c_{j1}^* + b_{k2} c_{j2}^*$</p> <p>where r_{jk} is the correlation between the k^{th} species and the j^{th} environmental variable</p>
<p>With focus on inter-species correlations (scaling ± 2), c_j^* is equal to the correlation between the j^{th} environmental variable and the ordination axis that has unit mean square (i.e. the species-derived sample scores in PCA and the environment-derived sample scores of RDA), hence</p> <p>In RDA: $c_j^* = \sum_i z_{ij} x'_i / \sum_i x'^2_i \quad \Rightarrow$ biplot: $z_{ij} \approx c_{j1}^* x'_{i1} + c_{j2}^* x'_{i2}$</p> <p>In PCA: $c_j^* = \sum_i z_{ij} x_i^* / \sum_i x_i^{*2} \quad \Rightarrow$ biplot: $z_{ij} \approx c_{j1}^* x_{i1}^* + c_{j2}^* x_{i2}^*$</p>

When the focus is on inter-species correlations (scaling ± 2), c_j^* is a correlation (last block in Table 6.29). It is thus also the slope parameter of the regression of the data of the j^{th} environmental variable on the sample scores that have unit mean square. These are the environment-derived sample scores in RDA and the species-derived scores in PCA. When the focus is on inter-species correlations, the biplot of the environmental biplot scores with the sample scores approximates the environmental data.

As the name suggests, the centroid score c_j^+ of the j^{th} environmental class is the mean of the sample scores of samples that belong to the j^{th} environmental class (Table 6.30). The centroid scores are also related to the species scores. A class of samples acts as a super sample in the

sense that all relations of the sample scores with the species scores carry over to the class centroids by changing the abundance of a species in the sample to the mean abundance of the species in the class of samples. So, if m_{jk} is the mean abundance of the k^{th} species in the j^{th} environmental class, then c_j^+ is the slope coefficient of the simple regression of the means $\{m_{jk}\}$ of all species for the j^{th} class on the species scores (Table 6.30). In a biplot with the species scores, the centroid scores thus approximate the class means of the species.

The relationships of c_j^+ and c_j^+ to the sample scores $\{x'_i\}$ indicated in Table 6.29 and Table 6.30 do not hold true for supplementary variables in RDA (see Table 6.53 on page 173).

Table 6.30 The centroid scores $\{c_j^+\}$ of environmental classes in PCA and RDA.

Relation to the sample scores (centroid of sample scores):

$$c_j^+ = \sum_i z_{ij} x_{i*} / \sum_i z_{ij} = \sum_i z_{ij} x'_i / \sum_i z_{ij}$$

⇒ centroid principle to predict z_{ij} from x_{i*} and/or x'_i and c_j^+

Relation to the unadjusted species scores (simple regression of class means on the species scores):

$$c_j^+ = \sum_k m_{jk} b_k / \sum_k b_k^2$$

where m_{jk} is the mean abundance of the k^{th} species in the j^{th} environmental class

$$m_{jk} = \sum_i z_{ij} y_{ik} / \sum_i z_{ij}$$

⇒ biplot: $m_{jk} \approx b_{k1} c_{j1}^+ + b_{k2} c_{j2}^+$

⇒ biplot: $m_{jk} / sd_k \approx b_{k1} c_{j1}^+ + b_{k2} c_{j2}^+$ for adjusted species scores (scaling +1 or +2)

The canonical coefficients $\{c_j\}$ are the partial regression coefficients from the multiple regression of $\{x_{i*}\}$ and also of $\{x'_i\}$ on the q environmental variables z_1, \dots, z_q (Table 6.31). When the canonical coefficients are plotted as arrows, the arrow for each environmental variable points in the direction that any particular sample point would move if that variable would increase in value conditional on the values of the other variables in the model. The canonical coefficients are also related to the species scores. To define this relation, let, as in Table 6.31, d_{jk} be the partial regression coefficient of the k^{th} species with respect to the j^{th} environmental variable. Then, c_j is the slope coefficient of the simple regression of the partial regression coefficients $\{d_{jk}\}$ for the j^{th} variable on the species scores $\{b_k\}$. When plotted with the species scores, the canonical coefficients thus approximate the partial regression coefficients.

Table 6.31 Canonical coefficients $\{c_j\}$ in RDA.

Relation to the sample scores (multiple regression of sample scores on z_1, \dots, z_q):

$$\mathbf{c} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{x}^* = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{x}'$$

\Rightarrow predict change in x'_i (or x_i^*) due to change in z_{ij} from c_j , conditional on the values of the other environmental variables

Relation to the unadjusted species scores (simple regression of partial regression coefficients d_{jk} on the species scores):

$$c_j = \sum_k d_{jk} b_k / \sum_k b_k^2$$

where d_{jk} is the partial regression coefficient of the k^{th} species with respect to the j^{th} environmental variable

$$\mathbf{d}_k = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{y}_k$$

\Rightarrow biplot: $d_{jk} \approx b_{k1} c_{j1} + b_{k2} c_{j2}$

In scaling +1 and +2, the division of the species scores by the standard deviation gives a biplot of standardized regression coefficients d_{jk} / sd_k

\Rightarrow biplot: $d_{jk} / sd_k \approx b_{k1} c_{j1} + b_{k2} c_{j2}$

6.3.2.3 Correspondence Analysis (CA) and DCA with detrending by polynomials

In CA, reciprocal averaging relations hold true between the species scores and the species-derived sample scores (Table 6.32 and Table 6.33). With the focus on inter-sample distances (scaling ± 1), the sample score x_i^* is the weighted average of the species scores $\{u_k\}$ and that the species score u_k is proportional to the weighted average of the sample scores. The weights in these weighted averages are the data y_{ik} . When changing the focus to inter-species distances (scaling ± 2), the constant of proportionality (λ^{-1} , which is a value greater than 1) moves from the equation for the species scores to that for the sample scores. With the focus on inter-species distances (scaling ± 2), the species score u_k is the weighted average of the sample scores and the sample score x_i^* is proportional to the weighted average of the species scores $\{u_k\}$. These weighted averaging relations are the key to the interpretation of the joint plot in CA by the centroid principle (section 3.5).

In biplot scaling, a PCA-like reciprocal regression relation holds true in CA (Table 6.33). This relation is particularly useful when the eigenvalues are low (short gradients). The reciprocal regression relation does not use the original data $\{y_{ik}\}$, but the transformed data $\{p_{ik}\}$. The transformation has an explicit meaning in contingency tables: p_{ik} is the observed count divided by the expected count under row/column independence. See also section 3.9.4 on page 60. In the reciprocal regression relations, the species score u_k is the slope coefficient of the simple regression of the transformed data $\{p_{ik}\}$ of the k^{th} species on the sample scores $\{x_i^*\}$ and, the other way round, the sample score x_i^* is the slope coefficient of the simple regression of the transformed data $\{p_{ik}\}$ of the i^{th} sample on the species scores $\{u_k\}$. These regression relations are the key to the interpretation of the CA-biplot: the species and sample scores together form

a biplot that displays approximate transformed species value p_{ik} as indicated in Table 6.33. From the biplot it is easy to infer relative abundances, in particular, the share of species k in the total abundance of sample i (y_{ik}/y_{i+}) and the share that sample i has in the total abundance of species k (y_{ik}/y_{+k}). See page 171 of *Unimodal Models*.

The scaling of scores is also given in Table 6.32 and Table 6.33. In biplot scaling with a focus on inter-sample distances (scaling +1), the weighted mean square of the species-derived sample scores along an ordination axis is equal to the eigenvalue of the axis and the weighted mean square of the species scores is equal to 1. In biplot scaling with a focus on inter-species distances (scaling +2), the weighted mean square of the species scores along an ordination axis is equal to the eigenvalue of the axis and the weighted mean square of the species-derived sample scores is equal to 1. Hill's scaling is derived from the biplot scaling by division of all scores by $\sqrt{(1-\lambda)}$, resulting in the weighted mean squares of species scores and of sample scores given in Table 6.32.

The consequences of the two types of biplot scaling for the interpretation of distances between sample points and between species points in the ordination diagram can be seen from the eigenvalue equations of CA (Table 6.34). In scaling +1, the score for a particular sample is the slope coefficient of the weighted regression of the inter-sample inner products on all sample scores. These inner products are based on the transformed data $\{p_{ik}\}$ and weighted by the species totals. Sample points thus approximate these inter-sample inner products. Because of the relation between distances and inner products, distances between sample points in the ordination diagram approximate in scaling +1 weighed inter-sample Pythagorean distances based on the transformed data $\{p_{ik}\}$, which are chi-square distances in terms of the untransformed data $\{y_{ik}\}$. In scaling +2, the score for a particular species is the slope coefficient of the weighted regression of the inter-species covariances on all species scores. The covariances are based on the transformed data $\{p_{ik}\}$ and weighted by the sample totals. Therefore the species points thus approximate these inter-species covariances. Because of the relation between distances and inner products, distances between species points approximate in scaling +2 weighted inter-species Pythagorean distances based on the transformed data $\{p_{ik}\}$, which are chi-square distances in terms of the untransformed data $\{y_{ik}\}$. In summary, with the focus on inter-sample distances, the ordination diagram displays chi-square distances between samples whereas with the focus on inter-species distances, it displays chi-square distances between species.

In Hill's scaling with focus on samples, the sample scores are in Standard Deviation units of species turnover (SD-units). The distances among samples are thus turnover distances (*Unimodal Models*: p 164). In Hill's scaling with the focus on species, distances among species are (generalized) Mahalanobis distances in reduced space (see section 3.11 and the CVA example in section 8.4.3).

If environmental variables are present in the analysis by CA (in any scaling), regression coefficients $\mathbf{c} = (c_1, \dots, c_q)^T$ and environment-derived sample scores are calculated after the ordination has been obtained:

$$\mathbf{c} = (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W} \mathbf{x}^*$$

with $\mathbf{W} = \text{diag}(y_{1+}, \dots, y_{n+})$

$$x'_i = \sum_j c_j z_{ij}$$

The relations in Table 6.32 and Table 6.33 hold true also in DCA with detrending by polynomials (DCA-POL).

Table 6.32 Transition formulae and Hill's scaling in CA and DCA-POL.

<p>CA in Hill's scaling with focus on inter-sample distances (scaling -1)</p>	
$u_k = \lambda^{-1} \sum_i y_{ik} x_i^* / \sum_i y_{ik}$	<p>with $\sum_k y_{+k} u_k^2 / \sum_k y_{+k} = 1/(1-\lambda)$</p>
$x_i^* = \sum_k y_{ik} u_k / \sum_k y_{ik}$	<p>with $\sum_i y_{i+} x_i^{*2} / \sum_i y_{i+} = \lambda / (1-\lambda)$</p>
<p>⇒ centroid principle to predict y_{ik} in a sample from x_i^* and u_k</p>	
<p>CA in Hill's scaling with focus on inter-species distances (scaling -2)</p>	
$u_k = \sum_i y_{ik} x_i^* / \sum_i y_{ik}$	<p>with $\sum_k y_{+k} u_k^2 / \sum_k y_{+k} = \lambda / (1-\lambda)$</p>
$x_i^* = \lambda^{-1} \sum_k y_{ik} u_k / \sum_k y_{ik}$	<p>with $\sum_i y_{i+} x_i^{*2} / \sum_i y_{i+} = 1/(1-\lambda)$</p>
<p>⇒ centroid principle to predict occurrences y_{ik} of a species from x_i^* and u_k</p>	
<p>$\sum_i y_{i+} x_i^* / \sum_i y_{i+} = 0$ and $\sum_k y_{+k} u_k / \sum_k y_{+k} = 0$ (zero mean sample and species score)</p>	

Table 6.33 Transition formulae and the biplot scaling in CA and DCA-POL.

CA in biplot scaling with focus on inter-sample distances (scaling 1)

$$u_k = \sum_i y_{i+} p_{ik} x_i^* / \sum_i y_{i+} x_i^{*2} = \lambda^{-1} \sum_i y_{ik} x_i^* / \sum_i y_{ik} \quad \text{with} \quad \sum_k y_{+k} u_k^2 / \sum_k y_{+k} = 1$$

$$x_i^* = \sum_k y_{+k} p_{ik} u_k / \sum_k y_{+k} u_k^2 = \sum_k y_{ik} u_k / \sum_k y_{ik} \quad \text{with} \quad \sum_i y_{i+} x_i^{*2} / \sum_i y_{i+} = \lambda$$

with $p_{ik} = y_{ik} / (y_{i+} y_{+k} / y_{++})$

⇒ centroid principle to predict y_{ik} in a sample from x_i^* and u_k

⇒ biplot: $p_{ik} \approx 1 + u_{k1} x_{i1}^* + u_{k2} x_{i2}^*$

For given species k , y_{+k}/y_{++} is constant so that

⇒ biplot: $y_{ik}/y_{i+} \propto p_{ik} \approx 1 + u_{k1} x_{i1}^* + u_{k2} x_{i2}^*$ (project samples on species arrow)

For given sample i , y_{i+}/y_{++} is constant so that

⇒ biplot: $y_{ik}/y_{+k} \propto p_{ik} \approx 1 + u_{k1} x_{i1}^* + u_{k2} x_{i2}^*$ (project species on sample arrow)

CA in biplot scaling with focus on inter-species distances (scaling 2)

$$u_k = \sum_i y_{i+} p_{ik} x_i^* / \sum_i y_{i+} x_i^{*2} = \sum_i y_{ik} x_i^* / \sum_i y_{ik} \quad \text{with} \quad \sum_k y_{+k} u_k^2 / \sum_k y_{+k} = \lambda$$

$$x_i^* = \sum_k y_{+k} p_{ik} u_k / \sum_k y_{+k} u_k^2 = \lambda^{-1} \sum_k y_{ik} u_k / \sum_k y_{ik} \quad \text{with} \quad \sum_i y_{i+} x_i^{*2} / \sum_i y_{i+} = 1$$

⇒ centroid principle to predict occurrences y_{ik} of a species from x_i^* and u_k

⇒ biplots as above for scaling 1

$\sum_i y_{i+} x_i^* / \sum_i y_{i+} = 0$ and $\sum_k y_{+k} u_k / \sum_k y_{+k} = 0$ (zero mean sample and species score),

Table 6.34 Eigenvalue equations and the biplot scaling in CA.

In CA, the eigenvalue equation for the sample scores is

$$\lambda x_i^* = \sum_j y_{j+} c_{ij} x_j^* / y_{++}$$

where $c_{ij} = \sum_k y_{+k} p_{ik} p_{jk} / y_{++}$, the inner product between samples i and j

With focus on inter-sample distances (scaling +1), $\sum_i y_{i+} x_i^{*2} / y_{++} = \lambda$ so that

$$x_i^* = \lambda^{-1} \sum_j y_{j+} c_{ij} x_j^* / y_{++} = \sum_j y_{j+} c_{ij} x_j^* / \sum_j y_{j+} x_j^{*2}$$

⇒ biplot of inter-sample inner products: $c_{ij} \approx x_{i1}^* x_{j1}^* + x_{i2}^* x_{j2}^*$

⇒ approximation of inter-sample chi-square distances: $d_{ij} \approx \{(x_{i1}^* - x_{j1}^*)^2 + (x_{i2}^* - x_{j2}^*)^2\}^{1/2}$

with $d_{ij}^2 = \sum_k y_{+k} (p_{ik} - p_{jk})^2 / y_{++} = \sum_k (y_{++}/y_{+k})(y_{ik}/y_{i+} - y_{jk}/y_{j+})^2 = c_{ii} + c_{jj} - 2c_{ij}$

In CA, the eigenvalue equation for the species scores is

$$\lambda u_k = \sum_l y_{+l} c_{kl} u_l / y_{++}$$

where $c_{kl} = \sum_i y_{i+} p_{ik} p_{il} / y_{++}$, the covariance between species k and l

With focus on inter-species distances (scaling +2), $\sum_k y_{+k} u_k^2 / y_{++} = \lambda$ so that

$$u_k = \lambda^{-1} \sum_l y_{+l} c_{kl} u_l / y_{++} = \sum_l y_{+l} c_{kl} u_l / \sum_l y_{+l} u_l^2$$

⇒ biplot of inter-species covariances: $c_{kl} \approx u_{k1} u_{l1} + u_{k2} u_{l2}$

⇒ approximation of inter-species chi-square distances: $d_{kl} \approx \{(u_{k1} - u_{l1})^2 + (u_{k2} - u_{l2})^2\}^{1/2}$

with $d_{kl}^2 = \sum_i y_{i+} (p_{ik} - p_{il})^2 / y_{++} = \sum_i (y_{++}/y_{i+})(y_{ik}/y_{+k} - y_{il}/y_{+l})^2 = c_{kk} + c_{ll} - 2c_{kl}$

6.3.2.4 Detrended Correspondence Analysis with detrending by segments (DCA)

In default DCA (with detrending by segments and nonlinear rescaling of axes) the relationships between scores are less simple. What remains simple is the centroid relation:

$$x_i^* = \sum_k y_{ik} u_k / \sum_k y_{ik} \Rightarrow \text{centroid principle to predict } y_{ik} \text{ in a sample from } x_i^* \text{ and } u_k$$

The sample scores are in Standard Deviation units of species turnover (SD-units). The distances among samples are thus turnover distances (Unimodal Models: p 164).

6.3.2.5 Canonical correspondence analysis and DCCA with detrending by polynomials

Table 6.35 and Table 6.36 specify the relations between scores in CCA and DCCA with detrending by polynomials. The species scores have a simple relation with the environment-derived sample scores. With the focus on inter-sample distances (scaling ± 1), the species score u_k is proportional to the weighted average of the environment-derived sample scores $\{x'_i\}$, whereas the species-derived sample score x_i^* is the weighted average of the species scores $\{u_k\}$. When changing the focus to inter-species distances (scaling ± 2), the constant of proportionality (λ^{-1}) moves, as in CA, from the equation for the species scores to that for the species-derived sample scores. In all types of scaling, the canonical coefficients $\{c_j\}$ are the partial regression coefficients from a weighted multiple regression of $\{x_i^*\}$ on the q environmental variables z_1, \dots, z_q , and, finally, the environment-derived sample scores are a linear combination of the q environmental variables z_1, \dots, z_q . As in RDA, the canonical coefficient can also be derived from the environment-derived scores $\{x'_i\}$ by a multiple regression on the environmental variables.

Table 6.35 Transition formulae and Hill's scaling in CCA and DCCA-POL.

CCA in Hill's scaling with focus on inter-sample distances (scaling -1)	
$u_k = \lambda^{-1} \sum_i y_{ik} x'_i / \sum_i y_{ik}$	with $\sum_k y_{+k} u_k^2 / \sum_k y_{+k} = 1/(1-\lambda)$
$x_i^* = \sum_k y_{ik} u_k / \sum_k y_{ik}$	
$\mathbf{c} = (\mathbf{Z}^T \mathbf{WZ})^{-1} \mathbf{Z}^T \mathbf{Wx}^*$	
$x'_i = \sum_j c_j z_{ij}$	with $\sum_i y_{i+} x_i'^2 / \sum_i y_{i+} = \lambda/(1-\lambda)$
\Rightarrow centroid principle to predict y_{ik} in a sample from x_i^* (or x'_i) and u_k	
CCA in Hill's scaling with focus on inter-species distances (scaling -2)	
$u_k = \sum_i y_{ik} x'_i / \sum_i y_{ik}$	with $\sum_k y_{+k} u_k^2 / \sum_k y_{+k} = \lambda/(1-\lambda)$
$x_i^* = \lambda^{-1} \sum_k y_{ik} u_k / \sum_k y_{ik}$	
$\mathbf{c} = (\mathbf{Z}^T \mathbf{WZ})^{-1} \mathbf{Z}^T \mathbf{Wx}^*$	
$x'_i = \sum_j c_j z_{ij}$	with $\sum_i y_{i+} x_i'^2 / \sum_i y_{i+} = 1/(1-\lambda)$
\Rightarrow centroid principle to predict occurrences y_{ik} of a species from x'_i (or x_i^*) and u_k	
Relations of x'_i to the species scores (\propto weighted average of $\{u_k\}$):	
$x'_i = \sum_k \hat{y}_{ik} u_k / \sum_k y_{ik}$ with focus on inter-sample distances (scaling ± 1)	
\Rightarrow centroid principle to predict fitted abundance \hat{y}_{ik} in a sample from x'_i and u_k	
$x'_i = \lambda^{-1} \sum_k \hat{y}_{ik} u_k / \sum_k y_{ik}$ with focus on inter-species distances (scaling ± 2)	
\Rightarrow centroid principle to predict fitted occurrences \hat{y}_{ik} of a species from x'_i and u_k	
$\sum_i y_{i+} x'_i / \sum_i y_{i+} = \sum_i y_{i+} x_i^* / \sum_i y_{i+} = 0$ and $\sum_k y_{+k} u_k / \sum_k y_{+k} = 0$ (zero mean scores)	

Table 6.36 Transition formulae and the biplot scaling in CCA and DCCA-POL.

CCA in biplot scaling with focus on inter-sample distances (scaling 1)	
$u_k = \sum_i y_{i+} p_{ik} x'_i / \sum_i y_{i+} x'^2_i = \lambda^{-1} \sum_i y_{ik} x'_i / \sum_i y_{ik}$	$\text{with } \sum_k y_{+k} u_k^2 / \sum_k y_{+k} = 1$
$x_i^* = \sum_k y_{+k} p_{ik} u_k / \sum_k y_{+k} u_k^2 = \sum_k y_{ik} u_k / \sum_k y_{ik}$	
$c = (Z^T W Z)^{-1} Z^T W x^*$	
$x'_i = \sum_j c_j z_{ij}$	$\text{with } \sum_i y_{i+} x_i'^2 / \sum_i y_{i+} = \lambda$
with $p_{ik} = y_{ik} / (y_{i+} y_{+k} / y_{++})$	
⇒ centroid principle to predict y_{ik} in a sample from x_i^* (or x'_i) and u_k	
⇒ biplot: $p_{ik} \approx \hat{p}_{ik} \approx 1 + u_{k1} x'_{i1} + u_{k2} x'_{i2}$	
⇒ biplot: $p_{ik} \approx 1 + u_{k1} x_{i1}^* + u_{k2} x_{i2}^*$	
CCA in biplot scaling with focus on inter-species distances (scaling 2)	
$u_k = \sum_i y_{i+} p_{ik} x'_i / \sum_i y_{i+} x'^2_i = \sum_i y_{ik} x'_i / \sum_i y_{ik}$	$\text{with } \sum_k y_{+k} u_k^2 / \sum_k y_{+k} = \lambda$
$x_i^* = \sum_k y_{+k} p_{ik} u_k / \sum_k y_{+k} u_k^2 = \lambda^{-1} \sum_k y_{ik} u_k / \sum_k y_{ik}$	
$c = (Z^T W Z)^{-1} Z^T W x^*$	
$x'_i = \sum_j c_j z_{ij}$	$\text{with } \sum_i y_{i+} x_i'^2 / \sum_i y_{i+} = 1$
⇒ centroid principle to predict occurrences y_{ik} of a species from x'_i (or x_i^*) and u_k	
⇒ biplots as above for scaling 1	
<p>Relation of x'_i to the species scores (simple regression of fitted values $\{\hat{p}_{ik}\}$ on $\{u_k\}$):</p> $x'_i = \sum_k y_{+k} \hat{p}_{ik} u_k / \sum_k \hat{y}_{+k} u_k^2$ <p>with \hat{p}_{ik} and \hat{y}_{ik} the fitted values from the weighted regression on the environmental variables, namely</p> $\hat{p}_{ik} = 1 + Z(Z^T W Z)^{-1} Z^T W p_{ik} \quad \text{and} \quad \hat{y}_{ik} = (y_{i+} y_{+k} / y_{++}) \hat{p}_{ik}$ <p>⇒ biplot: $\hat{p}_{ik} \approx 1 + u_{k1} x'_{i1} + u_{k2} x'_{i2}$</p> <p>For given species k, y_{+k}/y_{++} is constant so that</p> <p>⇒ biplot: $\hat{y}_{ik}/y_{+k} \propto \hat{p}_{ik} \approx 1 + u_{k1} x'_{i1} + u_{k2} x'_{i2}$ (project samples on species arrow)</p> <p>For given sample i, y_{i+}/y_{++} is constant so that</p> <p>⇒ biplot: $\hat{y}_{ik}/y_{+k} \propto \hat{p}_{ik} \approx 1 + u_{k1} x'_{i1} + u_{k2} x'_{i2}$ (project species on sample arrow)</p>	

In biplot scaling, RDA-like regression relations hold true in CCA between species and species-derived sample scores (Table 6.36). These relations are particularly useful when the eigenvalues are low (short gradients). As in CA, the regression relations do not use the original data $\{y_{ik}\}$, but the transformed data $\{p_{ik}\}$ with p_{ik} the observed count divided by the expected count under row/column independence. See also section 3.9.4 on page 60. In the regression relations, the species score u_k is the slope coefficient of the simple regression of the transformed data $\{p_{ik}\}$ of the k^{th} species on the sample scores $\{x'_i\}$, and the sample score x'_i is the slope coefficient of the simple regression of the transformed data $\{p_{ik}\}$ of the i^{th} sample on the species scores $\{u_k\}$. In addition, the last cells of Table 6.35 and Table 6.36 show that the environment-derived sample scores have CA-like relations with the species scores. The environment-derived sample score x'_i is (proportional to) the weighted average of the species scores $\{u_k\}$ with the fitted abundance data \hat{y}_{ik} as weights (last cell of Table 6.35). The environment-derived sample score x'_i is also the slope coefficient of the simple regression of the fitted values $\{\hat{p}_{ik}\}$ of the i^{th} sample on the species scores $\{u_k\}$. The biplot of species and environment-derived sample scores thus displays the approximate fitted species values $\{\hat{p}_{ik}\}$. These regression relations are the key to the interpretation of the CCA-biplot: the species and sample scores together form a biplot that displays approximate fitted value $\{\hat{p}_{ik}\}$ as indicated in Table 6.36. Because the fitted row and column totals are equal to the observed totals, it is easy to infer fitted relative abundances, in particular, the fitted share of species k in the total abundance of sample i (\hat{y}_{ik}/y_{i+}) and the fitted share that sample i has in the total abundance of species k (\hat{y}_{ik}/y_{k+}). See page 171 of *Unimodal Models*.

The scaling of scores is also given in Table 6.35 and Table 6.36. In words, we have the following. In biplot scaling with a focus on inter-sample distances (scaling +1), the weighted mean square of the environment-derived sample scores along an ordination axis is equal to the eigenvalue of the axis and the weighted mean square of the species scores is equal to 1. In biplot scaling with a focus on inter-species distances (scaling +2), the weighted mean square of the species scores along an ordination axis is equal to the eigenvalue of the axis and the weighted mean square of the environment-derived sample scores is equal to 1. Hill's scaling is derived from the biplot scaling by division of all scores by $\sqrt{1-\lambda}$, resulting in the weighted mean squares of species scores and of sample scores given in Table 6.35.

The consequences of the two types of biplot scaling for the interpretation of distances between sample points and between species points in the ordination diagram can be seen from the eigenvalue equations of CCA (Table 6.37). In summary, with the focus on inter-sample distances, the ordination diagram displays fitted chi-square distances between samples whereas with the focus on inter-species distances, it displays fitted chi-square distances between species.

In Hill's scaling with focus on samples, the sample scores are in Standard Deviation units of species turnover (SD-units). The distances among samples are thus fitted turnover distances (*Unimodal Models*: p 164). In Hill's scaling with focus on species, distances among species are (generalized) Mahalanobis distances in environmental space (see the CVA example in section 8.4.3).

Table 6.37 Eigenvalue equations and the biplot scaling in CCA.

In CCA, the eigenvalue equation for the sample scores is

$$\lambda x'_i = \sum_j y_{j+} c_{ij} x'_j / y_{++}$$

where $c_{ij} = \sum_k y_{+k} \hat{p}_{ik} \hat{p}_{jk} / y_{++}$, the fitted inner product between samples i and j

With focus on inter-sample distances (scaling +1), $\sum_i y_{i+} x'^2_i / y_{++} = \lambda$ so that

$$x'_i = \lambda^{-1} \sum_j y_{j+} c_{ij} x'_j / y_{++} = \sum_j y_{j+} c_{ij} x'_j / \sum_j y_{j+} x'^2_j$$

⇒ biplot of inter-sample inner products: $c_{ij} \approx x'_{i1} x'_{j1} + x'_{i2} x'_{j2}$

⇒ approximation of fitted inter-sample chi-square distances:

$$d_{ij} \approx \{(x'_{i1} - x'_{j1})^2 + (x'_{i2} - x'_{j2})^2\}^{1/2}$$

$$\text{with } d_{ij}^2 = \sum_k y_{+k} (\hat{p}_{ik} - \hat{p}_{jk})^2 / y_{++} = \sum_k (y_{++}/y_{+k}) (\hat{y}_{ik}/y_{i+} - \hat{y}_{jk}/y_{j+})^2 = c_{ii} + c_{jj} - 2c_{ij}$$

In CCA, the eigenvalue equation for the species scores is

$\lambda u_k = \sum_l y_{+l} c_{kl} u_l / y_{++}$ where $c_{kl} = \sum_i y_{i+} \hat{p}_{ik} \hat{p}_{il} / y_{++}$, the fitted covariance between species k and l

With focus on inter-species distances (scaling +2), $\sum_k y_{k+} u_k^2 / y_{++} = \lambda$ so that

$$u_k = \lambda^{-1} \sum_l y_{+l} c_{kl} u_l / y_{++} = \sum_l y_{+l} c_{kl} u_l / \sum_l y_{+l} u_l^2$$

⇒ biplot of fitted inter-species covariances: $c_{kl} \approx u_{k1} u_{l1} + u_{k2} u_{l2}$

⇒ approximation of fitted inter-species chi-square distances:

$$d_{kl} \approx \{(u_{k1} - u_{l1})^2 + (u_{k2} - u_{l2})^2\}^{1/2}$$

$$\text{with } d_{kl}^2 = \sum_i y_{i+} (\hat{p}_{ik} - \hat{p}_{il})^2 / y_{++} = \sum_i (y_{++}/y_{i+}) (\hat{y}_{ik}/y_{+k} - \hat{y}_{il}/y_{+l})^2 = c_{kk} + c_{ll} - 2c_{kl}$$

The environmental biplot scores $\{c_j^*\}$ are related to both the sample scores and the species scores (Table 6.38). The score c_j^* is the slope coefficient from the simple regression of $\{x'_i\}$ on the j^{th} environmental variable z_j (Table 6.38). When the environmental biplot scores are plotted as arrows, the arrow for each environmental variable points in the direction that any particular sample point would move to if that variable would increase in value (ignoring the other variables). To define the relation with the species scores, let, as in Table 6.38, m_{jk} be the weighted average of the k^{th} species with respect to the j^{th} environmental variable, then c_j^* is the

slope coefficient of the simple regression of the weighted averages $\{m_{jk}\}$ of all species with respect to the j^{th} environmental variable on the species scores (Table 6.38). In a biplot with the species scores, the environmental biplot score for a particular variable thus approximates the weighted averages of the species with this environmental variable.

In biplot scaling with the focus on inter-species distances (scaling +2), c_j^* is a correlation (last block in Table 6.38). It is thus also the slope parameter of the regression of the data of the j^{th} environmental variable on the sample scores that have unit mean square. These are the environment-derived sample scores in CCA and the species-derived sample scores in CA. When the focus is on inter-species correlations, the biplot of the environmental biplot scores with the sample scores approximates the environmental data.

Table 6.38 The environmental biplot scores $\{c_j^*\}$ in CA, CCA and D(C)CA-POL.

<p>Relation to the sample scores (simple regression of sample scores on z_j):</p> $c_j^* = \sum_i y_{i+} z_{ij} x_i^* / \sum_i y_{i+} z_{ij}^2 = \sum_i y_{i+} z_{ij} x'_i / \sum_i y_{i+} z_{ij}^2 = \sum_i y_{i+} z_{ij} x'_i / y_{++}$ <p>⇒ predict change in x_i^* and/or x'_i due to change in z_{ij} from c_j^*</p>	
<p>Relation to the species scores (simple regression of weighted averages $\{m_{jk}\}$ on the species scores):</p> $c_j^* = \sum_k y_{+k} m_{jk} u_k / \sum_k y_{+k} u_k^2$ <p>where m_{jk} is the weighted average of the k^{th} species with respect to the j^{th} environmental variable</p> $m_{jk} = \sum_i y_{ik} z_{ij} / \sum_i y_{ik}$ <p>⇒ biplot: $m_{jk} \approx u_{k1} c_{j1}^* + u_{k2} c_{j2}^*$</p>	
<p>In biplot scaling with focus on inter-species distances (scaling +2), c_j^* is equal to the correlation between the j^{th} environmental variable and the ordination axis that has unit mean square (i.e. the species-derived sample scores in CA and the environment-derived sample scores of CCA), hence</p>	
<p>In CA: $c_j^* = \sum_i y_{i+} z_{ij} x_i^* / \sum_i y_{i+} x_i^{*2}$</p>	<p>⇒ biplot: $z_{ij} \approx c_{j1}^* x_{i1}^* + c_{j2}^* x_{i2}^*$</p>
<p>In CCA: $c_j^* = \sum_i y_{i+} z_{ij} x'_i / \sum_i y_{i+} x_i'^2$</p>	<p>⇒ biplot: $z_{ij} \approx c_{j1}^* x'_{i1} + c_{j2}^* x'_{i2}$</p>

As the name suggests, the centroid score c_j^+ of the j^{th} environmental class is the weighted mean of the sample scores of samples that belong to the j^{th} environmental class (Table 6.39). The centroid scores are also related to the species scores (Table 6.39). A class of samples acts as a super sample in the sense that all relations of the sample scores with the species scores carry over to the class centroids by changing the abundance of a species in the sample to the total abundance of the species in the class of samples. So, if y_{jk} is the total abundance of the k^{th} species in the j^{th} environmental class, then c_j^+ is the weighted average of the species scores $\{u_k\}$, with the total abundances acting as weights, when the focus is on inter-sample distances (scaling ± 1). c_j^+ is proportional to the weighted average, when the focus is on inter-species distances (scaling ± 2). In biplot scaling, there is an additional regression relation, as there is for the sample scores. When the class totals $\{y_{jk}\}$ are transformed to relative class totals $\{p_{jk}\}$ by division by the expected totals under row/column independence, then, c_j^+ is the slope coefficient of the simple regression of the relative class totals $\{p_{jk}\}$ of all species for the j^{th} class on the species scores. In

a biplot with the species scores, the centroid scores thus approximate the relative class totals of the species (Table 6.39).

Table 6.39 The centroid scores $\{c_j^+\}$ of environmental classes in CA, CCA and D(C)CA-POL.

<p>Relation to the sample scores (centroid of sample scores):</p> $c_j^+ = \sum_i y_{i+} z_{ij} x_i^* / \sum_i y_{i+} z_{ij} = \sum_i y_{i+} z_{ij} x_i' / \sum_i y_{i+} z_{ij}$ <p>Relation to the species scores: (simple regression of relative class totals p_{jk} on the species scores):</p> <p style="text-align: center;">biplot scaling with focus on inter-sample distances (scaling 1)</p> $c_j^+ = \sum_k y_{+k} p_{jk} u_k / \sum_k y_{+k} u_k^2 = \sum_k y_{jk} u_k / \sum_k y_{jk}$ <p style="text-align: center;">biplot scaling with focus on inter-species distances (scaling 2)</p> $c_j^+ = \sum_k y_{+k} p_{jk} u_k / \sum_k y_{+k} u_k^2 = \lambda^{-1} \sum_k y_{jk} u_k / \sum_k y_{jk}$ <p>with</p> <p>$y_{jk} = \sum_i z_{ij} y_{ik} / \sum_i z_{ij}$, the total abundance of the k^{th} species in the j^{th} environmental class and $p_{jk} = y_{jk} / (y_{j+} y_{+k} / y_{++})$, the corresponding transformed total.</p> <p>For given species k, y_{+k}/y_{++} is constant so that \Rightarrow biplot: $y_{jk}/y_{j+} \propto p_{jk} \approx 1 + u_{k1} c_{j1}^+ + u_{k2} c_{j2}^+$ (project classes on species arrow)</p> <p>For given class j, y_{j+}/y_{++} is constant so that \Rightarrow biplot: $y_{jk}/y_{+k} \propto p_{jk} \approx 1 + u_{k1} c_{j1}^+ + u_{k2} c_{j2}^+$ (project species on class arrow)</p>
--

6.3.2.6 Detrended Canonical Correspondence Analysis with detrending by segments

In DCCA with detrending by segments and nonlinear rescaling of axes, the relationships between scores are less simple. What remains simple are the relations given in Table 6.40.

Table 6.40 Some transition formulae in DCCA with detrending by segments.

$$x_i^* = \sum_k y_{ik} u_k / \sum_k y_{ik} \Rightarrow \text{centroid principle to predict } y_{ik} \text{ in a sample from } x_i^* \text{ (or } x'_i) \text{ and } u_k$$

$$c = (Z^T W Z)^{-1} Z^T W x^*$$

$$x'_i = \sum_j c_j z_{ij}$$

When plotted with the species scores, the environmental biplot arrow approximates the weighted averages of all species with respect to this environmental variable (Table 6.41). The centroids of environmental classes are, indeed, centroids of the sample scores (Table 6.42), but have no simple relation with the species scores.

Table 6.41 The environmental biplot scores $\{c_j^*\}$ in DCA and DCCA with detrending by segments.

Relation to the species scores (multiple regression of weighted averages $\{m_{jk}\}$ on the species scores of two axes):

c_{j1}^* = partial regression coefficient on axis 1

c_{j2}^* = partial regression coefficient on axis 2

with m_{jk} the weighted average of the k^{th} species with respect to the j^{th} environmental variable

$$m_{jk} = \sum_i y_{ik} z_{ij} / \sum_i y_{ik}$$

$$\Rightarrow \text{biplot: } m_{jk} \approx u_{k1} c_{j1}^* + u_{k2} c_{j2}^*$$

Table 6.42 The centroid scores $\{c_j^+\}$ of environmental classes in DCA and DCCA with detrending by segments.

Relation to the sample scores (centroid of sample scores):

$$c_j^+ = \sum_i y_{i+} z_{ij} x_i^* / \sum_i y_{i+} z_{ij} = \sum_i y_{i+} z_{ij} x'_i / \sum_i y_{i+} z_{ij}$$

6.3.3 Weights and the eigenvector sample scores

Let w_k^* denote the user-specified weights for species k and w_i^* the user-specified weight for sample i , which are both, by default, equal to 1, and equal to 0 for species and samples that are deleted or that are made supplementary. For unimodal methods, we adopt the notation that y_{+k} and y_{i+} are the weighted species total and weighted sample total, defined by

$$(6.4) \quad y_{+k} = \sum_i w_i^* y_{ik}, \text{ and } y_{i+} = \sum_k w_k^* y_{ik}$$

and that y_{++} is the overall total, defined by

$$(6.5) \quad y_{++} = \sum_{i,k} w_i^* w_k^* y_{ik} = \sum_i w_i^* y_{i+} = \sum_k w_k^* y_{+k}$$

Then, in linear methods, $w_k = w_k^*$ and $w_i = w_i^*$, and, in unimodal methods, $w_k = w_k^* y_{+k}$, the weighted total abundance of a species, and $w_i = w_i^* y_{i+}$, the weighted total abundance of a sample.

With environmental data in the analysis, there are two sets of sample scores in CANOCO: species-derived sample scores $\{x_i^*\}$ and environment-derived sample scores $\{x'_i\}$. In the subsequent sections of this chapter we use the term “eigenvector sample scores”, denoted by $\{x_i\}$, which are the eigenvectors of the analysis. In direct methods, where there must be environmental data in the analysis,

$$(6.6) \quad x_i = x'_i, \text{ the score of sample } i \text{ that is derived from the environmental data}$$

and, in indirect methods, where there may be, but do not need to be, environmental data in the analysis,

$$(6.7) \quad x_i = x_i^*, \text{ the score of sample } i \text{ that is derived from the species data}$$

6.3.4 Species scores

Table 6.43 - Table 6.45 show examples of tables of species scores. The column “WEIGHT” reports the weights w_k of each species. Without user-specified weights, $w_k = 1$ in linear methods, and $w_k = \sum_i y_{ik}$, the species total, in unimodal methods. With user-defined weights, in linear methods, $w_k = w_k^*$ and $w_i = w_i^*$, and, in unimodal methods, $w_k = w_k^* y_{+k}$, the weighted total abundance of a species, and $w_i = w_i^* y_{i+}$, the weighted total abundance of a sample.

Supplementary species are recognizable by weight 0 (species 31, 32 and 33 in Table 6.43 - Table 6.45). In the example tables, they happen to be at the bottom of the table but this is not the general rule. The weights are useful in interpreting ordination diagrams: in unimodal methods species at the edge of the diagram often carry low weights; such peripheral species have little influence on the analysis and it is often convenient not to display them at all.

In linear methods the final column is headed “1” and contains ones only. In unimodal methods, the final column, headed “N2”, is the effective number of occurrences of the species, defined by

$$(6.8) \quad N_2 = 1 / \sum_i (w_i^* y_{ik} / y_{+k})^2$$

It is analogous to the N_2 -diversity measure of Hill (1973b). N_2 can be understood as follows. For presence-absence data, N_2 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 N_2 is close to 1.

The scores of species depend on the scores of the samples (and vice versa). With environmental data in the analysis, there are two sets of sample scores in CANOCO. The closest is the relation of the species scores with the “eigenvector sample scores”, denoted by $\{x_i\}$, as defined in the previous section.

In linear methods, the species score is defined by

$$(6.9) \quad b_k = \sum_i w_i y'_{ik} x_i / \sum_i w_i x_i^2$$

the linear regression coefficient of the data for species k on the eigenvector samples scores $\{x_i\}$ using the weights $\{w_i\}$. The species score is thus a slope parameter. In linear methods, the data y'_{ik} are the species data after any data transformation, centering and standardization that you may have specified for the analysis. The value of the denominator of (6.9) depends only on the scaling of ordination scores (Table 6.3) and is set to $\lambda^\alpha \sum_i w_i$ in the next section, with λ the eigenvalue of the axis and $\alpha = 1, 0, 1/2$ for scaling $\pm 1, 2$, and 3 of Table 6.3. The formula for the species scores can thus be simplified to

$$(6.10) \quad b_k = \lambda^{-\alpha} \sum_i w_i y'_{ik} x_i / \sum_i w_i$$

If $\alpha = 0$ (scaling ± 2), then $\lambda^{-\alpha} = 1$, so that the species score is a **weighted sum** of the sample scores.

If you specified that the species score must be post-transformed by division by the standard deviation (indicated by positive scaling type numbers, Table 6.3) the species scores is divided by sd_k , the standard deviation of species k . The resulting species scores are said to be adjusted for the species variance (Table 6.43). The adjusted species scores can still be interpreted as in (6.9), but now with y'_{ik} defined by y_{ik} / sd_k . The adjusted species score is the regression coefficient of the standardized species data on to the sample scores. The resulting biplot of species and sample scores thus displays standardized species data, even if the ordination was carried out on unstandardized data. See also *Unimodal Models*, page 146. The sign of the scaling type number in linear methods has no influence on other scores than the species scores. In scaling 2, the species score is precisely the correlation of the species with the ordination axis defined by the sample scores $\{x_i\}$. This is an inter-set correlation in direct methods, because the sample scores $\{x_i\}$ are then derived from the other set, namely the environmental data.

In unimodal methods, the species score is defined by

$$(6.11) \quad u_k = \lambda^{-\alpha} \sum_i w_i^* y_{ik} x_i / \sum_i w_i^* y_{ik}$$

with λ and α as defined above for linear methods. If $\alpha = 0$ (scaling ± 2), then $\lambda^{-\alpha} = 1$, so that the species score is the **weighted average** of the sample scores $\{x_i\}$. In the other types of scaling, the species score is proportional to the weighted average. The weighted average is the center of the species distribution along the ordination axis. It is an approximation of the species optimum if the species response curve is unimodal and symmetric. Recall that the species data y_{ik} must be non-negative in unimodal methods, otherwise (6.11) would not make sense as a weighted average. When nonlinear rescaling of axes is in force, which is the default in detrending-by-segments, the species score u_k is not a simple function of the sample scores and (6.11) does not hold.

In unimodal methods, the species scores have a weighted mean of 0 (except when nonlinear rescaling is in force), whereas in linear methods they do not have a mean of 0, except when the species data are centered by samples. With nonlinear rescaling of the ordination axes (default in detrending-by-segments) and environmental data in the analysis, the centroid of the species scores is reported below the table. Environmental biplot scores should take the centroid as origin of the coordinate system. Arrows for environmental variables in a species-environment biplot should start from this centroid-point.

The formulae (6.9) and (6.11) also define how the scores of supplementary species are obtained.

We now discuss the scaling of ordination scores. In linear methods, the (weighted) mean of squares of the (unadjusted) species scores is equal to

$$(6.12) \quad \sum_k w_k b_k^2 / \sum_k w_k = \lambda^{1-\alpha}$$

In unimodal methods with biplot scaling, the (weighted) mean of squares of the species scores (Table 6.44)

$$(6.13) \quad \sum_k w_k u_k^2 / \sum_k w_k = \lambda^{1-\alpha}$$

whereas, with Hill's scaling, the weighted mean square of the species scores is

$$(6.14) \quad \sum_k w_k u_k^2 / \sum_k w_k = \lambda^{1-\alpha} / (1-\lambda)$$

The factor $(1-\lambda)$ in (6.14) together with the definition of $\{x_i\}$ ensures that the species and sample scores are in Standard Deviation units (SD). Species scores in SD-units on average have, by definition, unit within-sample variance:

$$(6.15) \quad \sum_{i,k} w_i^* w_k^* y_{ik} (u_k - x_i)^2 / \sum_{i,k} w_i^* w_k^* y_{ik} = 1$$

With this scaling, the length of the ordination axis is, by definition, the range of the sample scores $\{x_i\}$.

It is of interest to note that the species scores in unimodal methods can also be interpreted as regression coefficients (slopes), at least if the axes are in biplot scaling. For this interpretation, define

$$(6.16) \quad y'_{ik} = (y_{ik} / y_{i+}) / (y_{+k} / y_{++})$$

which is the share of the species k in the total abundance in sample i (y_{ik} / y_{i+}) compared with the overall share of species k in the data (y_{+k} / y_{++}). The overall share is a constant for a given species. On inserting the definitions for y'_{ik} and w_i in (6.9), we obtain

$$(6.17) \quad b_k = \sum_i \{w_i^* y_{ik} x_i / y_{+k}\} / (\sum_i w_i x_i^2 / y_{++}) = u_k (\lambda^\alpha y_{++} / \sum_i w_i x_i^2) = u_k$$

The last equality only holds true if the biplot scaling is used (scaling 1, 2, or 3), as follows from (6.22) and (6.23) in the next section (with $y_{++} = \sum_i w_i$). This interpretation motivated the term biplot scaling, because in biplots species scores are slopes.

Table 6.43 Species scores $\{b_k\}$ in linear methods.

Spec: Species scores (adjusted for species variance)

N	NAME	AX1	AX2	AX3	AX4	WEIGHT	1
	EIG	.2644	.1701	.0671	.0413		
1	Ach mil	-.6878	.1239	-.1430	.0074	1.00	1.00
2	Agr sto	.6368	-.5218	-.0691	.0421	1.00	1.00
3	Air pra	.2203	.3987	-.0665	-.4094	1.00	1.00
4 — 29	not shown						
30	Cal cus	.6031	.1841	-.1098	.2219	1.00	1.00
31	Hip rha	.0372	.4293	.2180	-.3724	.00	1.00

32	Poa ann	-.3992	-.2753	-.4247	-.2815	.00	1.00
33	Ran acr	-.3467	.0615	.7173	.3179	.00	1.00

Table 6.44 Species scores {u_k} in unimodal methods.

Spec: Species scores (Biplot scaling)

N	NAME	AX1	AX2	AX3	AX4	WEIGHT	N2
	EIG	.4612	.2981	.1601	.1337		
1	Ach mil	-.8402	.3816	.0275	-.3342	16.00	6.10
2	Agr sto	.7704	-.5000	-.1143	-.0800	48.00	9.14
3	Air pra	.7395	1.7874	-1.0769	.5319	5.00	1.92
4 — 29	not shown						
30	Cal cus	1.6569	.4507	.3862	-.2538	10.00	2.94
31	Hip rha	.1048	1.4608	-.4087	.8238	.00	2.67
32	Poa ann	-.4230	-.0766	-.4707	-.3737	.00	8.83
33	Ran acr	-.3526	.0198	1.0076	.7046	.00	5.26

Table 6.45 Species scores {u_k} in DCA and DCCA with non-linear rescaling (default in detrending-by-segments).

Spec: Species scores

N	NAME	AX1	AX2	AX3	AX4	WEIGHT	N2
	EIG	.5360	.2565	.0833	.0349		
1	Ach mil	-.2799	-.0827	1.4066	1.6464	16.00	6.10
2	Agr sto	3.5361	1.7391	.2252	-.1242	48.00	9.14
3	Air pra	-.4583	3.8576	-.2429	1.7514	5.00	1.92
4 — 29	not shown						
30	Cal cus	3.8659	2.0476	.7723	4.6385	10.00	2.94
31	Hip rha	1.1185	3.3987	1.6260	2.9200	.00	2.67
32	Poa ann	1.0532	.3620	1.7688	1.7386	.00	8.83
33	Ran acr	1.3757	2.2444	-1.2399	-.4324	.00	5.26
	Centroid	1.6881	1.4897	.7745	.8722		

6.3.5 Sample scores {x_i^{*}} that are derived from the species

The sample scores that are derived from the species scores are labeled “Samp: Sample scores” in the solution file. They make the species axis (SPEC AX) in section 6.2.4. The format is similar to that of the species scores. The column “WEIGHT” reports the weight w_i of each sample. In linear methods, $w_i = w_i^*$; in unimodal methods, $w_i = w_i^*y_{it}$, the weighted total abundance in a sample where w_i^* are the user-defined weights (see previous section, above equation (6.4)). Supplementary samples are always placed at the bottom of the table (see samples “SupplSAM” and “Duplic17” in Table 6.46).

In linear methods the final column is headed "1" and contains ones only. In unimodal methods, the final column, headed "N₂", is the effective number of species in a sample, defined by

$$(6.18) \quad N_2 = 1 / \sum_k (w_k^* y_{ik} / y_{i+})^2$$

N₂ is a member of Hill's (1973b) family of diversity measures. The N₂-diversity measure is the inverse of the Simpson diversity measure. The meaning of N₂ can be understood as follows. For presence-absence data, N₂ is simply the number of species that occur in a sample. With abundance data, the species in a sample may have abundances 1000, 1, 1, say. CA/CCA/DCA are based on weighted averages. The weighted average for this sample is effectively determined by the species that occurs with abundance 1000 and the value of N₂ is close to 1.

In linear methods, the table of sample scores ends with the scores of a notional "empty" sample that has zero abundance value for all species in the data value. This sample, labeled "ORIGIN" (Table 6.46) indicates the point (0,0, ..., 0) in the original species space before centering has been applied. The point for "ORIGIN" in the ordination diagram can be useful in inferring the alpha-diversity of samples (Ter Braak 1983).

In linear methods, the sample score is essentially defined by

$$(6.19) \quad x_i^* = \sum_k w_k y'_{ik} b_k / \sum_k w_k b_k^2$$

the linear regression coefficient of the data for sample i on to the species scores { b_k } using weights { w_k }. The sample score is thus a slope parameter. In linear methods, the data y'_{ik} are the species data after any data transformation, centering and standardization that you may have specified for the analysis. The value of the denominator of (6.19) depends only on the scaling of ordination scores (Table 6.21) and is set to λ^{1-α} ∑_k w_k with λ the eigenvalue of the axis and α = 1, 0, 1/2 for scaling +/- 1, 2, and 3 (Table 6.21). The formula for the sample scores can thus be simplified to

$$(6.20) \quad x_i^* = \lambda^{\alpha-1} \sum_k w_k y_{ik} b_k / \sum_k w_k$$

If α = 1 (scaling +/- 1), then λ^{α-1} = 1, so that the sample score is a **weighted sum** of the species scores.

In weighted averaging methods, the sample score x_i^{*} is defined by

$$(6.21) \quad x_i^* = \lambda^{\alpha-1} \sum_k w_k^* y_{ik} u_k / \sum_k w_k^* y_{ik}$$

If α = 1 (scaling +/- 1), the sample score is therefore a **weighted average** of the species scores { u_k }. In other scalings, the sample score is proportional to the weighted average.

If there are **covariables** in the analysis, the scores { x_i^{*} } are made uncorrelated to the covariables before they are printed in order to avoid distortion by the effects of covariables. The sample scores printed are the residuals of a regression of the scores (6.20) or (6.21) on the covariables. Scores of supplementary samples for which the values of covariables are available are adjusted by use of the equation of the regression just mentioned.

In unimodal methods (except with nonlinear rescaling) the sample scores have a weighted mean of 0, whereas in linear methods they have a mean of 0 in RDA or if the species data are centered by species. With nonlinear rescaling of the ordination axes (default in detrending-by-segments) and environmental data in the analysis, the centroid of the sample scores is reported below the table. Environmental biplot scores should take the centroid as the origin of the

coordinate system. Arrows for environmental variables in a sample-environment biplot should start from this centroid-point.

In indirect gradient analyses, the transition formulae consist of the formulae (6.9) and (6.19) for linear methods (PCA) and the formulae (6.11) and (6.21) in unimodal methods (CA/DCA). In indirect methods, the sample scores $\{x_i^*\}$, which are derived from the species scores by (6.19) and (6.21), are the eigenvector sample scores, i.e. $x_i = x_i^*$ in the notation of the previous section.

We now discuss the scaling of the sample scores. In linear methods, and in unimodal methods in biplot scaling, the (weighted) mean of squares of the eigenvector sample scores $\{x_i\}$ is equal to

$$(6.22) \quad \sum_i w_i x_i^2 / \sum_i w_i = \lambda^\alpha$$

In unimodal methods in Hill's scaling, the weighted mean square of the eigenvector sample scores is equal to

$$(6.23) \quad \sum_i w_i x_i^2 / \sum_i w_i = \lambda^\alpha / (1 - \lambda)$$

The factor $(1-\lambda)$ in (6.23) together with the definition of $\{u_k\}$ ensures that the species and sample scores are in Standard Deviation units (SD) in which the scores have on average unit within-sample variance (6.15). With this scaling, the length of the ordination axis is, by definition, the range of the eigenvector sample scores $\{x_i\}$.

In detrending-by-segments (with non-linear rescaling of axes) the sample scores are always weighted averages of the species scores, and are in SD-units. The type of scaling is thus most like that of scaling -1 in the CA and CCA; in the heading of the table the value -1 is given (Table 6.21). The minimum value of the eigenvector sample scores $\{x_i\}$ is set to 0, as in DECORANA (Hill 1979). With environmental data in the analysis, the sample scores end with the centroid of the sample scores.

Table 6.46 Sample scores $\{x_i^*\}$ that are derived from the species scores in linear methods.

Samp: Sample scores

N	NAME	AX1	AX2	AX3	AX4	WEIGHT	N2
	EIG	.4612	.2981	.1601	.1337		
1	Sample 1	-1.2192	-.4968	-.9350	-1.2524	18.00	3.77
2	Sample 2	-.8644	-.2504	-.5356	-1.7028	42.00	9.09
3	Sample 3	-.3149	-1.0096	-.9001	-.6378	40.00	8.25
3 — 16	not shown						
17	Sample17	-.3882	2.7700	-1.0653	.9045	15.00	6.08
28	Sample18	-.3107	1.4947	-.1467	-.0740	27.00	7.22
29	Sample19	.6647	2.8731	-2.6645	1.7243	31.00	7.94
30	Sample20	2.0014	1.0029	-.2635	.3276	31.00	7.57
20	SupplSAM	-1.4219	.0180	.2932	-.8283	.00	2.88
21	Duplic17	-.3882	2.7700	-1.0653	.9045	.00	6.08

It is of interest to note that the sample scores in unimodal methods can also be interpreted as regression coefficients (slopes), at least if the axes are in biplot scaling. For this interpretation, define

$$(6.24) \quad y'_{ik} = (y_{ik} / y_{+k}) / (y_{i+} / y_{++})$$

which is the share of sample i in the total abundance of species k (y_{ik} / y_{+k}) compared with the overall share of sample i in the data (y_{i+} / y_{++}). On inserting the definitions for y'_{ik} and w_k in (6.19), we obtain

$$(6.25) \quad x_i^* = \sum_k \{w_k^* y_{ik} b_k / y_{i+}\} / (\sum_k w_k b_k^2 / y_{++})$$

The last term between brackets equals $\lambda^{1-\alpha}$ from (6.12), because $y_{++} = \sum_k w_k$ and $u_k = b_k$ in biplot scaling. Therefore the formulae for x_i^* in (6.25) and (6.21) correspond. This interpretation actually motivated the term biplot scaling, because in biplots sample and species scores are slopes. Note that (6.16) and (6.24) are two ways of writing the same data transformation. In the calculation of the score of species k from the sample scores, the last term between brackets in (6.16) does not depend on the samples and is thus a constant, given the species. *Mutatis mutandis*, the same holds true in (6.24). From a biplot with species and sample points turned into arrows by connecting the origin with each of the points, each species arrow thus points in the direction of maximum rate of change of the shares $\{y_{ik} / y_{+i}\}$ and each sample arrow points in the direction of maximum rate of change of the shares $\{y_{ik} / y_{+k}\}$.

6.3.6 Regression coefficients and associated t-values

The regression/canonical coefficients (Table 6.47) are the coefficients of a weighted multiple regression of the sample scores $\{x_i^*\}$ from the previous section on the standardized environmental variables. Again, there are four columns of coefficients because the regression is calculated for each ordination axis separately. Let z_{ij} be the value of environmental variable j ($j = 1, \dots, q$) in sample i and let z_j and s_j be the mean and standard deviation of variable j as given in section 6.2.4 (Table 6.6). The environmental variable is standardized to mean 0 and variance 1:

$$(6.26) \quad z_{ij} = (Z_{ij} - z_j) / s_j$$

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

$$(6.27) \quad x_i^* = c_0 + \sum_j c_j z_{ij} + \varepsilon_i$$

where c_0 is the intercept, c_j the regression coefficient of environmental variable j and ε_i is the error term with mean 0 and variance inversely proportional to w_i (section 6.3.3). Because the environmental variables are centered to mean 0, the intercept c_0 is equal to the mean of the species axis, i.e. $c_0 = 0$ except when nonlinear rescaling is in force. The other coefficients are estimated - using matrix notation - by

$$(6.28) \quad \mathbf{c} = (\mathbf{Z}^T \mathbf{W} \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{W} \mathbf{x}^*$$

where \mathbf{c} and \mathbf{x}^* are column vectors, $\mathbf{c} = (c_1, c_2, \dots, c_q)^T$ and $\mathbf{x}^* = (x_1^*, x_2^*, \dots, x_n^*)^T$, \mathbf{Z} is an $n \times q$ matrix with elements z_{ij} and \mathbf{W} is an $n \times n$ diagonal matrix with as i th element w_i . The fitted values of the regression are

$$(6.29) \quad x'_i = c_0 + \sum_j c_j z_{ij}$$

Note that the $\{c_j\}$ in (6.29) are estimates whereas in (6.27) they represent the true values, but this difference is not made explicit in the notation; the error term in (6.27) says enough. The fitted values $\{x'_i\}$ are termed **sample scores which are linear combinations of environmental variables** (Table 6.49). They constitute the environmental axis (ENVI AX) of section 6.2.4 (Table 6.5). The correlation between the species axis $\{x_i^*\}$ and the environmental axis $\{x'_i\}$ is the multiple correlation between the species axis $\{x_i^*\}$ and the environmental variables (= species-environment correlation).

Table 6.47 Regression coefficients $\{c_j\}$ of standardized environmental variables for each of the ordination axes.

Regr: Regression/canonical coefficients for standardized variables

N	NAME	AX1	AX2	AX3	AX4
	EIG	.4612	.2981	.1601	.1337
1	A1	.1243	-.2646	.7458	-.5663
2	Moisture	.6840	-.3684	-.4711	-.0142
3	Manure	-.0313	-.1257	-.5311	-1.7254
4	Hayfield	-.2932	-.0185	-.6231	-.1678
5	Haypastu	-.2402	-.2334	-.4044	.2610
6	Pasture	.0000	.0000	.0000	.0000
7	SF	-.2247	-.9869	.2333	1.4094
8	BF	-.2788	-.6415	.1876	-.2459
9	HF	-.4373	-.8888	.6438	1.2964
10	NM	.0000	.0000	.0000	.0000

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

In so-called regression biplots (section 3.8), the canonical coefficients are plotted with the species scores. This biplot approximates the $q \times m$ table of regression coefficients $\{d_{jk}\}$ of the data $\{y'_{ik}\}$ [$i = 1, \dots, n$] of each of m species with respect to q standardized environmental variables $\{z_{ij}\}$. See equations (3.20). This type of biplot is explained in the Unimodal models on pages 238-258 and discussed on pages 63, 148 and 235. For an ecological application of the regression biplot see Baar & Ter Braak (1996). CanoDraw for Windows can create this type of biplot.

The regression/canonical coefficients can also be plotted with the sample scores $\{x'_i\}$. When drawn as arrows, the environmental arrow points in the direction that any particular sample point would move if that particular environmental variable would increase in value, whereas the other variables keep their values (conditional effect).

For interpretation purposes it is sometimes of interest to obtain the regression coefficients, $\{c_j^u\}$ say, corresponding to the environmental variables in their original units of measurement, i.e. without the standardization in (6.26). By inserting (6.26) in (6.29) we obtain after some elementary algebraic manipulation

$$(6.30) \quad x_i' = \{c_0 - \sum_j (z_j/s_j)\} + \sum_j (c_j/s_j) z_{ij}$$

The desired coefficients are thus $c_j^u = c_j/s_j$. The value of s_j is found in the table of means and standard deviations (section 6.2.4). De-standardization is particularly useful for classes of nominal variables, because the standard deviation does not have an attractive interpretation for class variables.

The regression/canonical coefficients of class variables can be plotted as points. For this, the coefficients must first be de-standardized as just indicated and, optionally, be centered per nominal variable (Ter Braak & Looman 1994; Unimodal models: chapter 15). When class variables are plotted as points, the vector difference between two points gives the amount and direction that any particular sample point would move if the class membership of the sample would change from the one class to the other conditional on the values of the other environmental variables.

Table 6.48 shows **t-values** of coefficients of the regression in equation (6.29). The t-value of a regression coefficient c_j of variable on an ordination axis is equal to $c_j / \text{se}(c_j)$, where $\text{se}(c_j)$ - the standard error of the estimate c_j - is the square root of $\text{var}(c_j)$ given in (6.2). In indirect gradient analyses, the coefficients $\{c_j^*\}$ are normal regression coefficients, so that the t-values can be used in Student t-tests in the usual way (e.g. Montgomery & Peck 1982; Jongman et al., 1987, sections 3.2.1 and 3.5.2). To test the null hypothesis that the true coefficient of a particular variable on an axis is equal to 0, the t-value of the variable should be compared with the critical value of a Student t-distribution with $n-q-1$ degrees of freedom (n = number of samples, q = number of environmental variables). A variable is shown to contribute significantly to the regression if its t-value in absolute value exceeds the critical value (the critical value for a t-test at the 5% significance level is ca. 2.1, if $n-q-1 > 18$).

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

Note that a table of regression/canonical coefficients of a direct gradient analysis may contain both canonical coefficients and regression coefficients: even if the first columns contain canonical coefficients, the later columns may contain regression coefficients. How many columns contain canonical coefficients is indicated by the number of "Canonical axes" given in the heading of a table on the output file (e.g. Table 6.19). The different columns of the corresponding table of t-values then have different statistical properties!

The fraction of variance that an ordination axis explains in the species-environment biplot is also given in the table of t-values (FR EXPLAINED). It is the same information as given cumulatively in the summary of the ordination (section 6.2.5) In CCA and RDA the fraction

explained by axis k is simply $\lambda_k / (\text{sum of all canonical eigenvalues})$. See section 6.2.5 for further explanation.

Table 6.48 t-Values of the regression coefficients { c_j } of Table 6.47.

tVal: t-values of regression coefficients

N	NAME	AX1	AX2	AX3	AX4
	FR EXPLAINED	.3781	.2444	.1312	.1096
1	A1	1.0315	-1.3724	3.0602	-2.7288
2	Moisture	5.5690	-1.8752	-1.8968	-.0673
3	Manure	-.1203	-.3019	-1.0094	-3.8512
4	Hayfield	-1.9739	-.0777	-2.0745	-.6563
5	Haypastu	-1.7887	-1.0861	-1.4891	1.1284
6	Pasture	.0000	.0000	.0000	.0000
7	SF	-.8198	-2.2509	.4210	2.9865
8	BF	-1.9656	-2.8275	.6543	-1.0070
9	HF	-2.2666	-2.8797	1.6503	3.9025
10	NM	.0000	.0000	.0000	.0000

When CANOCO detects a collinear variable (section 6.2.2), its regression coefficients and t-values are set to 0.000. In Table 6.47 and Table 6.48, for example, the variables Pasture and NM have zero regression coefficients and t-values. These variables were detected to be collinear with the other environmental variables in the analysis (section 6.2.2). They are each the last category of a nominal environmental variable.

If there are **covariables** in the analysis, then the values z_{ij} in (6.27) are replaced by the residuals of a multivariate multiple regression of the standardized environmental variables on the covariables (without any further standardization). The regression in (6.27) is then a partial multiple regression (Kendall & Stuart 1973 sections 27.8 and 27.25; Seber 1977, Theorem 3.7 (ii); note that there is no need to regress the species data on the covariables). In the variance of the partial regression coefficient c_j (6.2) and in the Student t-test, q must be replaced by $q+p$, where p is the number of covariables. When the partial regression is performed after extracting the ordination, we obtain an indirect analysis of a partial principal components analysis, a partial CA or a partial DCA; when it is incorporated within the iterative ordination algorithm, we obtain a direct analysis: partial RDA, partial CCA or partial DCCA.

6.3.7 Sample scores { x'_i } that are derived from the environment

Table 6.49 shows a table of sample scores { x'_i } that are derived from the environmental variables. These scores are the fitted values of the regression (6.27) and are thus a linear combination of the environmental variables as shown in equation (6.29) with { c_j } the regression coefficients of Table 6.47. Whereas these scores are a by-product in indirect gradient analyses, these scores are the eigenvector scores in direct gradient analyses. In direct gradient analyses, the scores are scaled as indicated in equations (6.22) and (6.23) of section 6.3.5, with $x_i = x'_i$.

Samples made supplementary or samples without environmental data are missing from the table.

The weighted mean square of the sample scores { x'_i } is always a factor R^2 smaller than the mean square of the sample scores { x_i^* } where R is the species-environment correlation of the

corresponding axis (this follows from the regression in (6.27)). This rule is reflected in the standard deviations given for ENVI AX's and SPEC AX's in Table 6.6.

The column "WEIGHT" in Table 6.49 is the same as in Table 6.46. The column "%FIT" is an ordination diagnostic and is discussed in section 6.3.11.2 .

Table 6.49 Sample scores { x_i } that are derived from the environmental variables.

SamE: Sample scores which are linear combinations of environmental variables

N	NAME	AX1	AX2	AX3	AX4	WEIGHT	% FIT
	EIG	.4612	.2981	.1601	.1337		
1	Sample 1	-.8862	-.4334	-1.2765	-.2324	18.00	29.38
2	Sample 2	-1.0430	.1049	-.2283	-1.6488	42.00	62.37
3— 16	<i>not shown</i>						
17	Sample17	.0404	2.1449	-.5150	.4919	15.00	35.84
28	Sample18	-.3146	2.2724	-.0025	.3176	27.00	24.73
29	Sample19	1.2056	1.5490	-1.4515	.5585	31.00	53.83
30	Sample20	1.1923	1.5775	-1.5317	.6193	31.00	39.33

6.3.8 Inter-set correlations of environmental variables with axes

The inter-set correlations of environmental variables with the axes (Table 6.50) are the correlation coefficients between the environmental variables and the species-derived sample scores { x_i^* }. The same correlations can also be found in the full correlation matrix in the log-window (section 6.2.4). If there are covariables in the analysis, the correlations given are partial correlations. The correlations are given for all variables, including all categories of nominal variables. Beware that correlation coefficients for class variables convey little information.

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

In contrast to regression/canonical coefficients, the inter-set correlations do not become unstable when the environmental variables are strongly correlated with each other, i.e. when the VIF's of section 6.2.4 are large. See also pages 63-64 of Unimodal models.

Table 6.50 also shows the **fraction** of the total variance in the standardized environmental data that is **extracted** by each species axis (FR EXTRACTED). The fraction extracted is equal to the mean squared inter-set correlation, $\sum_j r_j^2 / q$, where r_j is the inter-set correlation of environmental variable j (cf. Gittins 1985: section 3.2.2)

Table 6.50 Inter-set correlation of environmental variables with the ordination axes.

CorE: Inter set correlations of environmental variables with axes

N	NAME	AX1	AX2	AX3	AX4
	FR EXTRACTED	.1823	.1943	.0762	.0655
1	A1	.5392	-.1562	.5042	-.0972
2	Moisture	.8833	-.1535	-.1199	.1507
3	Manure	-.2962	-.6895	-.1687	-.1603
4	Hayfield	-.0724	.5453	-.2158	.2508
5	Haypastu	-.1647	-.4992	-.1129	-.0768
6	Pasture	.2677	-.0281	.3660	-.1875
7	SF	.1421	-.6273	-.3601	-.0768
8	BF	-.3491	.1578	-.0255	-.5195
9	HF	-.3459	-.1047	.3758	.4643
10	NM	.5464	.6656	.0007	.0379

6.3.9 Biplot scores of environmental variables

The biplot scores of environmental variables (Table 6.51) are primarily meant to be plotted together with the species scores. This so-called species-environment biplot serves to give, in linear methods, a display of approximate values of correlations (or covariances) between species and environmental variables and, in unimodal methods, of weighted averages of species with respect to environmental variables. The biplot scores of environmental variables are optimized to this aim: in principle, they can be obtained by a weighted regression of these covariances/weighted averages on the species scores (see Table 6.29 and Table 6.38 and section 17.1). This regression is actually calculated in CANOCO when detrending-by-segments or nonlinear rescaling of axes or a ranking method (a nonstandard analysis) is in force. But for other methods there is a short-cut (Jongman et al. 1987: section 5.5): the regression gives biplot scores that are a simple function of the intra-set correlations (section 6.2.4), the standard deviations of the ordination axes (the ENVI AX's of section 6.2.4) and the eigenvalues, namely

$$(6.31) \quad c_j^* = (\text{intra-set correlation}) \times (\text{standard deviation of environmental axis})$$

In direct methods using scaling 2, the standard deviation of the environmental axis is equal to 1, so that the biplot score of an environmental variable is precisely its intra-set correlation. CANOCO uses a slightly different shortcut: as shown in section 17.1, the regression gives biplot scores that are a simple function of the inter-set correlations (section 6.3.8), the standard deviations of the species axes (the SPEC AX's of section 6.2.4) and the eigenvalues, namely

$$(6.32) \quad c_j^* = (\text{inter-set correlation}) \times (\text{standard deviation of species axis})$$

In indirect methods using scaling 2, the standard deviation of the species axis is equal to 1, so that the biplot score of an environmental variable is precisely its inter-set correlation in this case. CANOCO uses shortcut (6.32), because it carries through for supplementary environmental variables whereas (6.31) does not.

For unimodal methods in Hill's scaling, (6.31) and (6.32) must be multiplied by $1-\lambda_k$ with λ_k the eigenvalue of axis k.

In the species-environment biplot, quantitative environmental variables are often represented by arrows. The biplot scores of environmental variables (Table 6.51) give the

coordinates of the heads of the arrows, the coordinates of the species being given in section 6.3.4. Classes of nominal environmental variables are often better displayed by centroid points as described in the next section.

The biplot scores can also usefully be plotted together with the sample scores $\{x_i^*\}$ or $\{x'_i\}$. Consider the simple regression of the samples scores $\{x_i^*\}$ on the j th standardized environmental variable,

$$(6.33) \quad x_i^* = c_0^* + c_j^* z_{ij} + \varepsilon_i^*$$

with c_0^* the intercept, c_j^* the simple regression coefficient of environmental variable j and ε_i^* the error term. Then, c_j^* is equal to the biplot score given by (6.32). If the dependent variable in (6.33), x_i^* , is replaced by x'_i , the regression coefficient is equal to (6.31), and is thus also equal to the biplot score (except for supplementary variables in a direct gradient analysis; see Table 6.53 on page 173). This property of the biplot scores was noted in Appendix B of Ter Braak (1994). When drawn as arrows, the environmental arrow points in the direction that any particular sample point would move if that particular environmental variable would increase in value, ignoring the other environmental variables (which is fine under the assumption that the other variables covary with that one environmental variable in the particular way they do in the data set).

Because the sample scores $\{x_i^*\}$ are a low-dimensional representation of the species data, a related interpretation is that the biplot scores of environmental variables display the marginal effect of each variable on the species, as displayed in the ordination diagram, whereas the regression/canonical coefficients give the conditional effects. This interpretation holds per axis in scaling 2 and for all axes simultaneously in scaling 1, as discussed in Appendix B of Ter Braak (1994). These interpretations also hold true for supplementary environmental variables.

In scaling +/- 2 in linear methods and scaling 2 in unimodal methods (focus on species-relations with biplot scaling), the plot of the environmental biplot scores and the sample scores has an extra bonus: the plot displays the approximate values of the environmental data. The bonus comes about because the biplot scores of the j th environmental variable happen to be equal to the regression coefficients of the regression of the standardized data of that one variable on the ordination axes, at least if the sample scores have unit variance. This is the case for the scores $\{x_i^*\}$ in indirect analyzes and the scores $\{x'_i\}$ in direct analyzes; these scores are the ones that CanoDraw would plot by default. In these cases, the arrows based on the environmental biplot scores also display the approximate values of the correlations among environmental variables. The length of each arrow is equal to the multiple correlation of the variable with the displayed ordination axes. See Ter Braak (1994: Table I) and Ter Braak & Verdonschot (1995: Table 2). Strictly speaking, these interpretations do not hold true for supplementary environmental variables in a direct gradient analysis. For these interpretations the optimal score to plot is the intra-set correlation given in the log-window, which is not equal to the biplot score of a supplementary variable in this case.

Table 6.51 also shows the species-environment correlations, $R(\text{SPEC},\text{ENV})$. A value of 0.00, would indicate that the correlation could not be calculated by CANOCO.

If there are **covariables** in the analysis, (6.31) is multiplied by the residual standard deviation of the regression of each standardized environmental variable on the covariables (= the square root of the diagonal element of the partial covariance matrix displayed at the screen in the console version of CANOCO). In this way, the arrow of an environmental variable becomes shorter, the higher the correlation between this environmental variable and the covariables, (i.e., the more the variation in the environmental variable is already explained by the covariables). The arrow is unaffected when the environmental variable is not correlated to

the covariables, i.e. when it contributes entirely new information about the environment. With covariables in the analysis the species-environment biplot approximates in linear methods **partial** covariances and, in weighted averaging methods, weighted averages with respect to residuals of environmental variables (i.e. the environmental variables after eliminating covariable-effects). See section 17.1.

Environmental variables with long arrows are the most important in the analysis. The larger the arrow, the more confident one can be about the inferred covariances (correlations) or weighted averages, and, roughly speaking, the larger the effect of the variable on the species. The interpretation of the length of arrows based on the environmental biplot scores is discussed in more detail on pages 152 and 169-170 of Unimodal Models.

The rules for constructing and interpreting species-environment biplots are the same as those given in Jongman et al. (1987, section 5.3.4) for PCA biplots. Because the scores for species and for environmental variables are often of a different order of magnitude, the biplot is constructed most easily by drawing separate plots of species and of environmental variables on transparent paper, each one with its own scaling. But note that within each plot the scale units of the axes must have equal physical length. The biplot is obtained by superimposing the plots with the axes aligned and the origins of the coordinate systems coinciding. However, there is an exception to the rule that the origins must coincide: when the mean of a species axis is nonzero, then the origin of the “environmental plot” must coincide with the point in the “species plot” whose coordinates are equal to the means of the species axes (SPEC AX_k) given in section 6.2.4. This exception happens in linear methods when the species data are not centered by species, and in unimodal methods when nonlinear rescaling of axes is in force. In both cases there is an extra line below the species and sample scores, starting with the word “CENTROID” which specifies the means of the species axes. Note that in the linear case this centroid is not necessarily equal to the mean of the species scores. If one does not want to draw separate plots, the head of an environmental arrow can be added to the plot of the species at the point whose coordinates are obtained by the formula

$$(\gamma \times \text{biplot-score-of-environmental-variable}) + (\text{mean-of-species-axis})$$

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

Table 6.51 Biplot scores {c_j*} of environmental variables.

BipE: Biplot scores of environmental variables

N	NAME	AX1	AX2	AX3	AX4
	R(SPEC,ENV)	.9580	.9018	.8554	.8888
1	A1	.5629	-.1732	.5894	-.1094
2	Moisture	.9221	-.1702	-.1402	.1696
3	Manure	-.3092	-.7646	-.1972	-.1803
4	Hayfield	-.0756	.6046	-.2523	.2821
5	Haypastu	-.1719	-.5535	-.1320	-.0864
6	Pasture	.2795	-.0312	.4279	-.2110
7	SF	.1484	-.6956	-.4210	-.0864
8	BF	-.3645	.1750	-.0298	-.5845
9	HF	-.3611	-.1161	.4394	.5224
10	NM	.5704	.7381	.0008	.0426

6.3.10 Centroids of environmental variables in the ordination diagram

Nominal environmental variables can naturally be represented by points in the ordination diagram (Ter Braak 1986, Ter Braak 1994, Ter Braak & Verdonschot 1995). Each class of a nominal variable gives one point which is located at the centroid of the sample scores belonging to the class. CANOCO calculates the centroids from the species axes, i.e. from the sample scores $\{x_i^*\}$ using the formula

$$(6.34) \quad c_j^+ = \sum_i w_i z_{ij} x_i^* / \sum_i w_i z_{ij}$$

where z_{ij} is the value of environmental variable j in sample i before any standardization (section 6.3.6, equation (6.26)). Interestingly, the centroids of the sample scores $\{x_i^*\}$ coincide with the centroids of $\{x_i^*\}$ obtained in (6.34), except for supplementary variables.

As the environmental biplot scores, the centroids are primarily meant to be plotted together with the species scores in a species-environment biplot. For nominal variables, the plot serves to give, in linear methods, a display of approximate mean values of species in classes and, in unimodal methods, of relative class totals. The centroid scores are optimized to this aim: in principle they can be obtained by a weighted regression of these means and totals on the species scores (Table 6.30 and Table 6.39). The resulting scores are simply the centroids given by (6.34), except with Hill's scaling and when nonlinear rescaling of axes is in force. See section 17.2 for details.

In contrast to the environmental biplot scores, the centroids must be plotted in the ordination diagram in the same scale as the sample scores (e.g. Figure 8-2). The reason is that a class of nominal variables acts as a new super sample (see also section 13.2). All rules for interpreting plots of species and samples, thus also apply to plots of species and centroids. Representing environmental variables by points is not only useful for classes (dummy variables with $z_{ij} = 0$ or 1) but sometimes also for non-negative quantitative variables that can be absent, i.e. where the value 0 has a special meaning.

As the environmental biplot scores, the centroids can also usefully be plotted together with the sample scores $\{x_i^*\}$ or $\{x'_i\}$. Consider the multiple regression of the samples scores $\{x_i^*\}$ on all dummy variables (z_{ij}) defining a nominal environmental variable with K classes

$$(6.35) \quad x_i^* = \sum_j c_j^+ z_{ij} + \varepsilon_i^*$$

with c_j^+ the regression coefficient of class j ($j = 1, \dots, K$) and ε_i^* the error term. This multiple regression, without intercept, is equivalent to an analysis of variance. Then, c_j^+ is equal to the centroid score given by (6.34). If the dependent variable in (6.35), x_i^* , is replaced by x'_i , the resulting regression coefficient is the same (except for supplementary variables in a direct gradient analysis). These properties of the biplot scores were noted in Appendix B of Ter Braak (1994). When plotted as points, the vector difference between two class points gives the amount and direction that any particular sample point would move if the class membership of the sample would change from the one class to the other, ignoring the other environmental variables (which is fine under the assumption that the other variables covary with that one nominal environmental variable in the particular way they do in the data set). By contrast, the scores for classes based on the canonical coefficients (page 163) follow from the multiple regression equation (6.27). So, when the de-standardized canonical coefficients for class variables are plotted as points, the vector difference between two points gives the amount and direction that any particular sample point would move if the class membership of the sample would change from the one class to the other class for fixed values of the other environmental variables (Ter Braak & Looman 1994; Unimodal Models: chapter 15).

Because the sample scores $\{x_i^*\}$ are a low-dimensional representation of the species data, a related interpretation is that the centroids for environmental classes of a nominal variable display the marginal effect of the variable on the species, as displayed in the ordination diagram, whereas the regression/canonical coefficients give the conditional effect. This interpretation holds per axis in scaling 2 and for all axes simultaneously in scaling 1, as discussed in Appendix B of Ter Braak (1994). These interpretations also hold true for supplementary environmental variables.

Table 6.52 Centroids $\{c_j^+\}$ of environmental variables in the ordination diagram.

ConE: Centroids of environmental variables (mean.gt.0) in ordination diagram

N	NAME	AX1	AX2	AX3	AX4
	R(SPEC,ENV)	.9580	.9018	.8554	.8888
1	A1	.2236	-.0688	.2342	-.0434
2	Moisture	.5698	-.1051	-.0866	.1048
3	Manure	-.2215	-.5478	-.1413	-.1292
4	Hayfield	-.1057	.8448	-.3526	.3942
5	Haypastu	-.2043	-.6577	-.1569	-.1027
6	Pasture	.4884	-.0545	.7477	-.3686
7	SF	.2278	-1.0681	-.6464	-.1327
8	BF	-.8030	.3856	-.0656	-1.2879
9	HF	-.5375	-.1728	.6541	.7777
10	NM	1.0726	1.3880	.0015	.0801

Centroid scores and biplot scores for supplementary environmental variables may look strange in a direct ordination diagram as explained Table 6.53.

Table 6.53 Centroid scores for supplementary environmental variables.

Centroid scores for supplementary environmental variables may look strange in a direct ordination diagram, because centroids based on the scores $\{x_i^*\}$ differ from those based on $\{x'_i\}$. To illustrate this, let variable C_1 indicate a class that consist of sample 1 only, i.e. C_1 is 1 for sample 1 and 0 for the other samples. If this variable is taken as a supplementary environment variable in a direct gradient analysis, then C_1 does not coincide with the point plotted for sample 1. This point is based on x'_1 , whereas C_1 coincides with x_1^* , which is not plotted in a default direct gradient analysis. The reason for this counterintuitive phenomenon is that the primary aim of the centroid scores is to display the means or relative totals of the species in the class. In this extreme case, the means or relative totals are $\{y_{ik}\}$ and $\{y_{ik}/y_{+k}\}$ [$k = 1, \dots, m$], which are best approximated, for given species scores, by the sample scores x_1^* as follows from the regressions resulting in (6.19) and (6.25). The primary aim is thus best served if the centroid point for C_1 is x_1^* (and not x'_1 , as would be more intuitive). The regression in (6.35) with C_1 and C_{j1} as regressors (with $C_{j1} = 1 - C_1$) yields $c_1^* = x_1^*$, as expected by now. To extend the example to environmental biplot scores, let us treat C_1 as a quantitative variable. Its environmental biplot score, which best approximates the correlations of the species with C_1 , in linear methods, and weighted averages of the species with respect to C_1 , in unimodal methods, would not point in the direction of x'_1 but in the direction of x_1^* (Ter Braak 1994: Appendix A). An identical result would be obtained from the regression in (6.33) with C_1 as single regressor.

CANOCO calculates the centroids defined by (6.34) for all environmental variables whose mean is positive. The value assigned to the other environmental variables is the mean of the sample scores (usually 0 unless the species data are not centered or nonlinear rescaling of axes is in force). For variables whose mean is positive but which have some negative values, (6.34) is nonsensical but is still given by CANOCO.

Why nominal variables are naturally represented by points can also be seen from the case that there is a single nominal environmental variable, i.e. when there is a single pre-defined classification of samples. CCA with a series of dummy variables reflecting this classification provides an ordination to show maximum separation among the pre-defined groups of samples. This analysis is mathematically equivalent with Feoli and Orlóci's (1979) "analysis of concentration" and also with a simple CA of a two-way table of species-by-groups, the cells of which contain the total abundance of each of the species in each of the groups of samples (Greenacre 1984, section 7.1). In the CA ordination diagram the groups would be represented by points as they take the place of the samples. Similarly, RDA with a series of dummy variables is a variant of canonical variates analysis/multiple discriminant analysis, in which the groups are always represented by group means, i.e., by centroids (6.34).

The environmental centroids also have an attractive interpretation when there are covariables in the analysis, namely in terms of adjusted means and adjusted totals (means and totals from which the effects of the covariables have been removed by regression). See section 17.2.

6.3.11 Ordination diagnostics

6.3.11.1 Introduction

Usually, an ordination diagram is not an exact representation of the data. Overall measures of the quality of the approximation are given in the “Summary of the ordination” in terms of percentages of variance accounted for. But not all species or 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 heterogeneity for samples. Tolerance and sample heterogeneity are not defined in PCA/RDA. The fit measures and residual distances are not available in DCA (segments).

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

6.3.11.2 Fit for species and residual distances for samples

Table 6.54 shows an example table of species fits in a CCA. The fit is shown cumulatively and expressed as a fraction of the variance of a species. For example, from Table 6.54 we deduce that the fit of “Ach mil” in a CCA ordination diagram of the first two axes is 39%. The variance of species k is defined as

$$(6.36) \quad \text{var}(y_k) = \sum_i w_i (y'_{ik} - \mu_k)^2 / \sum_i w_i$$

whereby w_i and y'_{ik} are defined in section 6.3.4 and μ_k is 0 in linear methods, and 1 in unimodal methods (μ_k being the weighted mean value of species k after data transformation and/or, in linear methods, centering or standardization by species and samples). In unimodal methods, the column headed VAR(y) contains values obtained from (6.36) but in linear methods the values are rescaled so that their mean is 1.

The reported fit statistics for species k and axis s are the regression sums of squares of the weighted regression of the data $\{y'_{ik}\}$ [$i=1, \dots, n$] for species k on the ordination axes numbered 1, \dots , s , when expressed as a fraction of the total sum of squares for the species. The sample weights used in the regression are $\{w_i\}$ and the samples scores used for each axis are the eigenvector sample scores (page 156). The fit is discussed in more detail below.

In unimodal methods, it is not immediately clear what (6.36) means in term of the original abundance values y_{ik} . We first express (6.36) as a chi-square statistic and then relate it to the regression of the data of the k th species on the sample scores. Chi-square statistics are of the form $(\text{observed} - \text{expected})^2 / \text{expected}$. The expected value under independence in a two-way contingency table is $e_{ik} = y_{i+} y_{+k} / y_{++}$ and observed values are y_{ik} . In this notation y'_{ik} in (6.16) is y_{ik} / e_{ik} . On using this and assuming for the moment that there are no user-defined weights (all $w_k = 1$), we obtain from (6.36), with $\sum_i w_i = y_{++}$,

$$\begin{aligned}
(6.37) \quad \text{var}(y_k) &= \sum_i (y_{i+}/y_{++}) \{ (y_{ik}/y_{i+}) / (y_{+k}/y_{++}) - 1 \}^2 \\
&= \sum_i (y_{i+}/y_{++}) (y_{ik}/e_{ik} - 1)^2 \\
&= \sum_i (y_{i+}/y_{++}) \{ (y_{ik} - e_{ik}) / e_{ik} \}^2 \\
&= \sum_i (y_{i+}/y_{++}) (y_{ik} - e_{ik})^2 / e_{ik}^2 \\
&= y_{+k}^{-1} \sum_i (y_{ik} - e_{ik})^2 / e_{ik}
\end{aligned}$$

so that the variance of species k is the chi-square statistic divided by y_{+k} . Note that

$$(6.38) \quad \text{total inertia} = \text{chi-square} / y_{++} = \sum_k (y_{+k}/y_{++}) \text{var}(y_k)$$

As an aside, equation (6.37) can also be written as (Greenacre, 1984, eq 2.4.2)

$$(6.39) \quad \text{var}(y_k) = \sum_i (y_{++}/y_{i+}) (y_{ik}/y_{+k} - y_{i+}/y_{++})^2$$

The curious aspect of this equation is that samples seem to have weights proportional to $1/y_{i+}$ instead of to y_{i+} , as in the remaining of this manual.

In the regression of the data of species k on the sample scores (section 6.3.4) the essential part of the data is y_{ik}/y_{i+} , because y_{+k}/y_{++} is constant in that regression. Therefore we wish to express the variance of species k in terms of $\{ y_{ik}/y_{i+} \}$, the share of the abundance that the species has in each of the samples. On expanding y'_{ik} and w_i in (6.36) we obtain

$$\begin{aligned}
(6.40) \quad \text{var}(y_k) &= \sum_i w_i^* y_{i+} \{ (y_{ik}/y_{i+}) / (y_{+k}/y_{++}) - 1 \}^2 / \sum_i w_i \\
&= (y_{+k}/y_{++})^{-2} \sum_i w_i^* y_{i+} (y_{ik}/y_{i+} - y_{+k}/y_{++})^2 / \sum_i w_i \\
&= (y_{+k}/y_{++})^{-2} \sum_i w_i (y_{ik}/y_{i+} - y_{+k}/y_{++})^2 / \sum_i w_i
\end{aligned}$$

When y_{ik} is compared with y_{i+} (instead of with y_{+k} as in (6.39)), the implied sample weights are the usual ones, $\{ w_i \}$ or, equivalently, $\{ w_i^* y_{i+} \}$.

We now give explicit formulae for the fitted values of the regressions. Let x_{is} denote the eigenvector sample score of sample i on axis s , and b_{ks} (or u_{ks}) the species scores of species k on axis s . Because of (6.17) we do not need to distinguish between u_{ks} and b_{ks} here. In linear methods, the fitted values of the regression of $\{ y_{ik} \}$ on the ordinations axes are

$$(6.41) \quad \hat{f}_{ik} = b_{k1} x_{i1} + b_{k2} x_{i2} + \dots$$

with as many product terms as there are axes in the regression. In unimodal methods, the fitted abundance values, \hat{f}_{ik} , are

$$(6.42) \quad f_{ik} / y_{i+} = (y_{+k} / y_{++}) (1 + b_{k1} x_{i1} + b_{k2} x_{i2} + \dots)$$

The variance of the fitted values is

$$(6.43) \quad \text{var}(f'_k) = \sum_i w_i (f'_{ik} - \mu_k)^2 / \sum_i w_i$$

with $f'_{ik} = f_{ik}$ in linear methods and $f'_{ik} = (f_{ik} / y_{i+}) / (y_{+k} / y_{++})$ in unimodal methods. Equivalently, for unimodal methods, f_{ik} from (6.42) can be inserted for y_{ik} in (6.40).

The fraction of the variance of a species fitted is $\text{var}(f'_k) / \text{var}(y_k)$. The fit by axis 1 is given under the heading AX1. In CA, 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 (i.e. by q canonical axes) is given in the last column, headed % EXPL.

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 under the heading %EXPL. 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 Data Editing Options. In CA/CCA, species with a large value for weight \times variance may have a large influence.

Table 6.54 Cumulative fit per species as fraction of variance of species.

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
4—29	not shown						
30	Cal cus	.3698	.3972	.4173	.4259	7.42	48.64
31	Hip rha	.0014	.2770	.2986	.3862	7.74	43.61
32	Poa ann	.1431	.1477	.3249	.4366	1.25	57.01
33	Ran acr	.0571	.0573	.5241	.7523	2.18	82.77

Table 6.55 shows an example of the diagnostics given for samples. The entries are derived from the squared Pythagorean distance between the data for the i th sample point and its fit by the ordination axes, $\{y'_{ik}\}$, and $\{f'_{ik}\}$ [$k=1, \dots, m$], respectively, i.e. from

$$(6.44) \quad \text{SQDIST} = \sum_k w_k (y'_{ik} - f'_{ik})^2 / \sum_k w_k$$

Before any ordination axes are fitted, the squared distance is calculated using $f'_{ik} = 0$ in linear methods and 1 in unimodal methods and is given under "SQLEN". After fitting s axes, the squared distance between the sample point and its s-dimensional fit in the ordination space is given under the heading AXs ($s=1, \dots, 4$). The percentage fit (% FIT) is (within rounding error) equal to $100 * (1 - \text{entry AX4/SQLEN})$.

In linear methods, the values of SQLEN are proportional to the ones calculated from the raw data, because the total mean square of the species data is set to 1 in CANOCO.

In unimodal methods (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) divided by its total abundance. The formula and its equivalent forms are the same as given for species in (6.37), (6.39) and (6.40), but with the indices k and i interchanged, i.e.

$$\begin{aligned}
 (6.45) \quad d(y_i) &= y_{i+}^{-1} \sum_k (y_{ik} - e_{ik})^2 / e_{ik} \\
 &= \sum_k (y_{++}/y_{+k}) (y_{ik}/y_{i+} - y_{+k}/y_{++})^2 \\
 &= (y_{i+}/y_{++})^{-2} \sum_k w_k (y_{ik}/y_{+k} - y_{i+}/y_{++})^2 / \sum_k w_k
 \end{aligned}$$

After fitting s axes, the squared distance between the sample point and the s-dimensional ordination space is

$$(6.46) \quad \text{SQDIST}(y_i) = (y_{i+}/y_{++})^{-2} \sum_k w_k (y_{ik}/y_{+k} - f_{ik}/y_{+k})^2 / \sum_k w_k$$

with f_{ik} from (6.42).

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 links an almost 'average' species composition to marked extreme environmental values.

Table 6.55 illustrates a subtle point about the calculation of the diagnostics for supplementary samples in direct gradient analyses. The samples Sample 17 and Duplic17 have identical species compositions, but the entries differ, except for SQLEN. The reason is that for the Sample 17 the scores x'_i are used in the calculations, whereas for Duplic17 the species-derived scores x_i^* are used, being the only available scores for this sample, because it does not have environmental data.

CANOCO does not calculate the squared residual lengths if there are covariables in the analysis.

Table 6.55 Squared residual length per sample.

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	Sample 1	2.4239	2.3515	2.2302	2.1596	3.06	29.38
2	Sample 2	.8389	.8579	.8270	.4398	1.17	62.37
3 — 16	not shown						
17	Sample17	6.6419	4.4713	4.3381	4.2515	6.63	35.84
28	Sample18	2.0199	1.5342	1.5341	1.5538	2.06	24.73
29	Sample19	5.7337	3.7957	2.8949	2.6791	5.80	53.83
30	Sample20	2.2773	2.0759	2.3222	2.3192	3.82	39.33
20	SupplSAM	11.1796	11.1795	11.1658	11.0740	12.11	8.57
21	Duplic17	6.5571	4.2701	4.0884	3.9790	6.63	39.95

6.3.11.3 Species tolerance and sample heterogeneity

The unimodal methods, 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, another being the weighted linear regression approach using relative abundances in the previous subsection. The species distribution approach is important in many ecological applications. In the unimodal methods, the species score is (proportional to) the weighted mean of the sample scores, and thereby indicates the center of the distribution of that particular species. 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 & Barendregt, 1986; Ter Braak & Looman, 1986; Ter Braak & 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 as was first noted in Ter Braak & Verdonschot (1995). After Green's paper appeared, a series of papers in *Ecology* discussed niche measures in canonical space (Dueser & Shugart, 1978; Dueser & Shugart, 1979; Carnes & Slade, 1982; Van Horne & Ford, 1982; Dueser & Shugart, 1982). CANOCO follows the round up by Carnes & 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 given in Table 6.56.

Table 6.56 Species tolerances.

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
4—29	not shown						
30	Cal cus	.4499	1.0854	1.5275	.5928	100.81	2.94
31	Hip rha	.6388	1.0312	.6063	.7442	77.34	2.67
32	Poa ann	.5064	1.0954	.6476	.8654	81.00	8.83
33	Ran acr	1.1855	.3282	.7141	1.0569	88.60	5.26

The species tolerance is calculated as

$$(6.47) \quad t_k = \left\{ \sum_i w_i^* y_{ik} (x_i - u_k)^2 / \sum_i w_i^* y_{ik} \right\}^{1/2}$$

with x_i and u_k as used throughout the manual. Thus, u_k is not the weighted average of the sample scores $\{ x_i \}$, if the scaling of the ordination axes focuses on inter-sample distances or if the scaling is symmetric (scaling +- 1 or 3). The tolerance gives a good impression of the range of the x-values over which a species occurs, but underestimates the true tolerance or true niche-breadth, if the scaling is +- 2. An extreme case is that $t_k = 0$, if a species occurs only once. For a fair statistical comparison of niche breadth, the bias must be removed. This can be achieved (as in Hill, 1979: p. 28) by dividing t_k by the $(1 - 1/N_2)^{1/2}$ in scaling +- 2 (Ter Braak & Verdonschot, 1995). N_2 is given in the last column of Table 6.56; it is the effective number of occurrences defined in (6.8) and explained on page 157 of this manual. This adjustment is performed by CanoDraw for Windows, when it imports a new Canoco project.

For samples, one can calculate the same measure of spread as for species. The sample heterogeneity (Table 6.57) is defined analogously to (6.47) by

$$(6.48) \quad h_i = \left\{ \frac{\sum_k w_k^* y_{ik} (x_i - u_k)^2}{\sum_k w_k^* y_{ik}} \right\}^{1/2}$$

The remarks on the bias of t_k also apply to h_i , but now the bias occurs when the species-derived samples scores are weighted averages of species scores (scaling +1: focus on inter-sample distances). For example, if the sample contains just one species, then $h_i = 0$, at least in an indirect method. In a direct method, h_i is not necessarily 0 in this extreme case, because the sample scores used are the eigenvector sample scores, which are derived from the environmental data in direct methods.

The root mean square heterogeneity across the 4 axes is given under the heading “RMSTOL” in Table 6.57. RMSTOL is a summary of the heterogeneity of the sample in the four-dimensional ordination space.

Table 6.57 Sample heterogeneity.

Het : Sample heterogeneity (root mean squared deviation for samples)

N	NAME	AX1	AX2	AX3	AX4	RMSTOL	N2
	FR FITTED	.2180	.1409	.0757	.0632		
1	Sample 1	.3524	.3640	1.1338	.1370	62.47	3.77
2	Sample 2	.7291	.4110	.2346	1.4485	84.46	9.09
3 — 16	not shown						
17	Sample17	.6037	1.4241	.5983	.4498	85.92	6.08
28	Sample18	.5033	1.8997	.4394	.4671	103.36	7.22
29	Sample19	1.0005	1.0026	1.1575	.3963	93.58	7.94
30	Sample20	.5791	1.4583	1.5726	.6505	115.74	7.57
20	SupplSAM	.7799	.1527	.4978	.9781	67.75	2.88
21	Duplic17	.6000	2.0170	1.0198	.8238	123.96	6.08

As in the previous subsection, one should be aware that the sample heterogeneity is calculated from the eigenvector sample scores $\{x_i\}$ for active samples, whereas for supplementary samples, the species-derived scores $\{x_i^*\}$ are used. If an active and supplementary sample have the same species composition (as Sample 17 and Duplic7 in the Dune Meadow data), then their diagnostics may differ in a direct gradient analysis (Table 6.57). With covariables in the analysis, an additional difference may occur: the diagnostic depends on whether or not values for covariables were entered for the supplementary sample, because, if covariable data are available for the supplementary data, then these are used to adjust the species-derived sample score (page 160).

6.3.12 t-Value biplot

The t-value biplot best approximates the t-values of the regression coefficients $\{d_{jk}\} [j = 1, \dots, q; k = 1, \dots, m]$ of the weighted multiple regression of the data $\{y'_{ik}\} [i = 1, \dots, n]$ of each of the species $[k = 1, \dots, m]$ on to the data of all environmental variables $\{z_{ij}\} [i = 1, \dots, n; j = 1, \dots, q]$. The regression model for the k th species is

$$(6.49) \quad y'_{ik} = a_k + \sum_j d_{jk} z_{ij} + \varepsilon_{ik}$$

and is fitted using sample weights $\{ w_i \}$. From a t-value biplot we can infer which species react significantly to any particular environmental variable and, vice versa, which environmental variables contribute significantly to the regression of any particular species (Ter Braak & Looman 1994).

To put the t-ratio biplot into context, recall that direct gradient analysis is a form of (weighted) multivariate multiple regression of the species on to the environmental variables (section 3.8). To the principal results of a regression analysis belong regression coefficients and associated t-values (Jongman et al. 1987: chapter 3). In section 6.3.6 we considered the regression of the species-derived sample scores $\{ x_i \}$ [$i = 1, \dots, n$] on to the q environmental variables. This resulted, for each ordination axis, in q canonical weights and q associated t-values, which were collected for all four ordination axes in $q \times 4$ tables such as Table 6.47 and Table 6.48). In this section we consider the multiple regression of the values of each species separately on the values of environmental variables. In total, there are m such regressions. The regression coefficients and their associated t-values can be collected in two $q \times m$ tables of environment \times species. Both tables can be biplotted, either separately or jointly, in so-called regression biplots (Ter Braak, 1990, Ter Braak & Looman 1994). The table of regression coefficients is represented by a biplot of the canonical weights (Table 6.47) and the species scores (e.g. Table 6.43). The environment \times species table of t-values is represented by a biplot of the coordinates of species and of environmental variables (Table 6.58 and Table 6.59).

For a correct interpretation of the t-value biplot, the coordinates for species and environmental variables must be plotted on the same scale. The points for the species in the t-value biplot indicate the critical t-value 2. The plot can be interpreted as follows. If the environmental points are projected on to a line through a particular species' point and the origin of the plot, the projection points give the approximate t-values for the environmental variables in the regression of this particular species. On this line, the origin marks the t-value of 0 and the species point marks the t-value of 2. All other marks can be found by linear inter- and extrapolation. For example, the mirror point of the species on the line, marks the t-value of -2. Environmental points that project outside the interval indicated by the species point and its mirror image are thus inferred to have t-values greater than ± 2 in the multiple regression for that species. These environmental variables are inferred to be statistically significant in the regression for that particular species.

It is also possible to indicate in which region of the plot the species lie that react significantly positively to a particular environmental variable (Ter Braak & Looman 1994). This region is a circle with as its diameter the line-segment that joins the environmental point and the origin. Species that have their t-value coordinates in the circle react positively to the environmental variable. Similarly, species that react significantly negatively to a particular environmental variable (Ter Braak & Looman 1994) lie in the circle with as its diameter the line-segment that joins the origin and the mirror image of the environmental point. The circles are called Van Dobben-circles after the person who invented them (Ter Braak & Looman 1994).

The coordinates for the t-value biplot are calculated as follows (Ter Braak & Looman 1994). The coordinate of environmental variable j on an ordination axis is equal to its canonical coefficient c_j on that axis (Table 6.47) divided by the square root of the variance inflation factor (VIF) of the variable (Table 6.6). The coordinates of the species depend on the dimension of the plot. In a two-dimension t-value biplot, the coordinate of the species k on axis s is given by

headings VAR(y) and %EXPL shown in Table 6.58). The %(E+C) column gives for each response variable the joint fit by the covariables and the environmental variables as a percentage of the variance of the response variable (VAR(y) in Table 6.54). The %E column provides the additional fit due to environmental variables. The fit is additional to the fit by the covariables and expressed as a percentage of VAR(y). The fit by the covariables alone is simply %(E+C) - %E.

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

It is instructive to explicitly write out the regression equations in CCA. The definition of the species data y'_{ik} is given in (6.16). In the regression for species k, the term y_{+k}/y_{++} is a constant multiplier. The essential part for the regression is thus y_{ik}/y_{+i} . The regression model for species k on to the environmental variables is thus essentially

$$(6.51) \quad y_{ik}/y_{+i} = c_{0k} + \sum_j d_{jk} z_{ij} + \varepsilon_{ik}$$

The regression is fitted to the data using sample weights $w_i = w_i^* y_{i+}$. The t-value biplot represents the t-values of the estimated coefficients $\{ d_{jk} \}$ in (6.51) in two-dimensions.

6.4 Species-by-environment table

Canoco for Windows can give an extra output file beyond the solution file. This is the file called "SPEC_ENV.TAB" which is placed automatically in the working directory of Canoco for Windows. Each time an analysis is carried out the file overwrites any existing file of this name. The file contains a species-by-environment table. This table contains:

- In linear methods, correlations between species and each environmental variable when species are centered and standardized, and similar covariances when species are just centered.
- In unimodal methods, weighted averages of species with respect to standardized environmental variables

The table is formatted as a CANOCO full format data file (section 4.4.1). Table 6.60 show an example from the CCA applied to the Dune Meadow data. The 33 columns refer to the 33 species and the 10 "samples" of the file to the 10 environmental variables. The names of the species and environmental variables are given in the lines after the data. The title indicates that the table is from a CCA (analysis 5; see Table 6.1) with 10 environmental variables, 33 species and no covariables. From the example table, we see that the third species (Air pra) has a weighted average of -0.4647 with respect to the standardized variable, the thickness of the A1 horizon. Because this value is the lowest of all weighted averages for A1, this species tends to occur in samples with lower value of A1 than the other species. The maximum value, 3.0704, applies to Species 21 (Pot pal). This species occurs in Samples 14 and 15 (Table 16.1), which are indeed the ones with the highest A1-value (Table 16.2).

The species-environment table serves two goals:

It is this table that is represented in the species-environment biplot (with species scores and environmental biplot scores). Inferences from the biplot may be in error, because of the biplot contains the main patterns in the table only. Especially when the biplot represents only a small part of the variation in the table (a low percentage variance explained of the species-environment relation in section 6.2.5), one may wish to verify that inferences drawn from the biplot actually hold true in the actual data on weighted averages or correlations in the table.

If there are many environmental variables compared to samples, the constraints in direct gradient methods are weak so that an RDA or CCA give almost the same results as a PCA or

CA. If one is dissatisfied with the result, one may try and analyze the species-environment table directly by a non-centered PCA. This type of analysis is called co-inertia analysis by Doledec & Chessel (1994). To obtain the standard co-inertia analysis in the CA/CCA context, each species must be weighted by the weights w_k (section 6.3.4). Co-inertia is briefly discussed in Ter Braak & Verdonschot (1995).

Table 6.60 Environment-by-species table in unimodal methods.

Environment-by-species table, analysis 5. 10 Vars 33 Specs 0 Cov
 (I5/ 5(8F9.4:/))

	1	2	3	4	5	6	7	8	9	10
1										
	-.4587	.5330	-.4647	.1215	-.2733	-.2068	-.4396	.7065		
	-.2605	1.1599	-.3143	-.5292	-.5172	.3842	.0287	-.0546		
	-.5088	-.2605	-.4038	-.1036	3.0704	.8370	.0111	-.0967		
	-.4461	-.1471	.1475	-.5157	.1375	.9644	-.2874	-.3714		
	.3793									
2										
	-.8240	.7525	.5768	.5319	-.3804	-.7740	-.7710	1.2699		
	-.4629	1.2699	-.4629	1.2699	.3714	1.1416	.5590	-.0886		
	-.6522	-.9295	-.4389	-.1054	1.2699	1.2699	-.6555	.4324		
	.6398	-1.0406	-.0697	-.8962	.0440	1.2699	-.0297	-.4810		
	-.3054									
....										
10										
	-.2302	.1216	1.8805	-.5318	.3872	-.1606	-.5318	-.5318		
	-.5318	.7226	-.5318	1.8805	1.3445	.4064	-.5318	.3170		
	-.4486	-.0679	-.3307	-.5318	1.8805	.8467	-.5318	-.1699		
	1.8805	-.5318	.0328	.0713	.3052	1.1568	1.2775	-.2302		
	-.0932									
0000										
00000000										
00000000										
00000000										
00000000										
00000000										
Ach mil Agr sto Air pra Alo gen Ant odo Bel per Bro hor Che alb Cir arv Ele pal										
Ely rep Emp nig Hyp rad Jun art Jun buf Leo aut Lol per Pla lan Poa pra Poa tri										
Pot pal Ran fla Rum ace Sag pro Sal rep Tri pra Tri rep Vic lat Bra rut Cal cus										
Hip rha Poa ann Ran acr										
A1 MoistureManure HayfieldHaypastuPasture SF BF HF NM										



7. Console version of CANOCO

7.1 Introduction

The console version of CANOCO 4.x has the same user interface as CANOCO version 3.x. You may want to use the console version of CANOCO, because

- you have no access to Canoco for Windows, for which you need Microsoft Windows 98, Windows NT, or later versions; if so, you need to obtain a version of CANOCO.EXE that runs on your system
- you prefer DOS or similar command-line operating systems,
- you are well experienced in automating CANOCO yourself using the CON-project file (section 7.19 and the readme file in the directory \CANOCO\SAMPLES\PROJECTS) and CANOCO.INI file (section 7.3),
- you want to use one of the options that are not available in the Canoco for Windows.

7.1.1 Differences with Canoco for Windows

The most important features in which the console version (i.e. the application CANOCO.EXE) differs from Canoco for Windows (i.e. the application CanoWin.exe) are that

- you cannot point and click using a mouse; instead you can only use the keyboard,
- you cannot correct mistakes or go back to reconsider options you have set; instead you must continue, or must press Control-C to interrupt CANOCO and start from scratch,
- you cannot browse for the input files; instead you must remember and type the file names without error,
- you cannot indicate species, samples, environmental variables and covariables by their code name; instead you must remember and type their identification numbers,
- you cannot consult on-line help; instead, you must have the manual at hand,
- you cannot obtain a summary of the forward selection results; instead you must assemble the results from the print file,
- you cannot run CanoDraw as an integrated part of Canoco for Windows; instead you must start CanoDraw separately and select the CON-project file yourself when defining new CanoDraw project.

Canoco for Windows uses in the background a DLL-version of the console version of CANOCO. Both versions are identical from the user-point of view. There are, however, some options that are not available in Canoco for Windows. In order of likely importance in Canoco for Windows 4.5, they are

- you cannot transform the species by piece-wise linear transformations, e.g. to presence-absence data (Q 34 on page 200)
- you cannot obtain more than 4 ordination axes (Q 46 on page 208)
- when you use restricted permutation types within blocks, you cannot have different layouts in different blocks (Q 63 on page 215),

- you cannot adjust the maximum data sizes to the particular problem at hand (Q 4 on page 188).

7.1.2 Differences with CANOCO 3.x

The console version of CANOCO 4.5 extends the capabilities of CANOCO 3.x. The extensions concern the size of the data that can be analyzed, and the Monte Carlo permutation tests. The new CANOCO version can perform permutation tests for split-plot designs and related balanced multi-level designs. CANOCO 3.x could, with difficulty, analyze repeated measurement designs, in particular Before-After-Control-Impact designs. The new split-plot design options make the analysis much easier to specify. There are also some minor modifications:

- environmental variables and covariables can be indicated both by selection and by omission (Q 26 and Q 30)
- Monte Carlo permutations must be specified at the very beginning of the forward selection process,
- the scaling of the species scores in linear methods has been made more natural. In the new scaling, the mean square of the species scores is 1 or equal to the eigenvalue (Table 6.25). In earlier versions of CANOCO the divisor for the sum of squares of the species scores was not m (the number of species) but n (the number of samples). In the notation of equation (6.12) on page 159, $\sum_k w_k$ is used now instead of $\sum_i w_i$. This change does not affect the way in which ordination diagrams and biplots are interpreted.

The first two modifications imply that CON-project files made for CANOCO 3.x cannot be used with CANOCO 4.5.

7.2 Ways to answer the questions

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

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

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

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

then the default value is 1 and coincides with the minimum value of valid answers. If the range is given as “1 [3]”, the default value is 3 and coincides with the maximum value of valid answers. If an invalid answer is given, the question is posed again. Real values like 2.5 are permitted only when the values in the range have a decimal point, e.g.

Range of valid answers: 1.0 [3.0]

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

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

By pressing RETURN, the implied answers are thus the values -1 and 0. If a comma is used in the answer, missing entries are replaced by the default value. In the example the answer “3” is interpreted as the values -1 and 3. Answers with commas only (e.g. “,” or, if 6 numbers are expected in the answer “,,,,,”) are interpreted as accepting the default.

7.3 Initialization file

Most of the default values used in the console version of CANOCO can be modified by using an initialization file, called CANOCO.INI. The initialization file must be placed in the working folder or, if you prefer, in the folder C:\CANOCO.

Table 7.1 Default CANOCO.INI file in Canoco version 4.0.

```
*CANOCO (values start in position 2)
1 = range [0,1] = (01) decimal output in solution file
  = TAB char   = (02) separator between decimal values in solution file
  = char       = (03) character by which to enclose names ,, ,, ,,
  = char       = (04) character to close the scores of each item ,, ,,
0 = 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 ibi
2 = range [-3,3] = (08) scaling ordination scores ca/dca/cca/dcca ibi
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
1 = 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 SPEC_ENV.TAB
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 B in log(Ay + B) transformation
7 = range [1,9] = (24) default analysis number (1=PCA 2=RDA, etc.)
                        = (25) answer file (input from file)
SPECIES.DTA           = (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 CANOPLLOT or other prog
SPEC_ENV.TAB         = (31) output file for spec-envi table
2 2 2 2 2 2 2 2 = 8 values in range [0,6] = (32) output ordination results
*ENDCANOCO
```

Table 7.1 shows the format of the initialization file. The values in the table are the defaults obtained without the initialization file. The meaning of the values will become clearer in subsequent sections.

To get the full advantage of the file CANOCO.INI for your own analyses, copy CANOCO.INI to the folder where your data are, make this folder the working folder and edit CANOCO.INI with Notepad or another editor or a word processor. (If you use a word processor,

make sure you save the file as a text-only or DOS-text file). For example, you may wish to replace the default file names by the names of your own data files.

Because file names are no longer restricted in length, the text after the file names in Table 7.1 (option 25 - 31) may have been deleted in the installed CANOCO.INI. The text in Table 7.1 serves as a reminder of the order of the file names in the CANOCO.INI file.

7.4 Introductory questions

Which questions are posed depends on the type of analysis. The questions are numbered as Q1, Q2, ..., in the order in which they are posed. CANOCO always starts with the question

Q 1 Type 0 for input from the screen
1 for input from answer file (2) is reserved for automated input from file
Press RETURN for the default indicated by []
Range of valid answers: [0] 2

By pressing RETURN (or typing 0) CANOCO continues the screen dialogue by asking further questions (see the next sections). The screen dialog is logged in the file CANOCO.CON in the working folder. If the answer is 1, then you need to answer only one more question from the screen:

Q 2 Type name of file with answers to the questions

If you type, for example, MYCCA.CON, then the program reads the answers to subsequent questions from the file MYCCA.CON and the analysis proceeds automatically. Commonly, the file specified here is a modification of the file CANOCO.CON from an earlier analysis. It can also be a project file from Canoco for Windows. An example of such a file is given later on in Table 7.4. See also the example project files in the directory \CANOCO\SAMPLES\PROJECTS. If the file contains too few answers, the screen dialogue starts again at the question left unanswered. See section 7.19 for more information.

The next question allows you to choose between a short and a long dialogue.

Q 3 Type 1 for long dialogue
Range of valid answers: 0 [1]

The long dialogue gives access to all the options in CANOCO. The short dialogue asks fewer questions. The values for the questions that are not posed are taken from the initialization file, if present. In the short dialogue, you can delete samples, covariables and environmental variables and choose a data-transformation for the species data, add supplementary environmental variables and carry out Monte Carlo permutation tests. You cannot make species or samples supplementary and you cannot set the centering/standardization of the data or the scaling of the ordination axes.

In the long dialogue the next question is

Q 4 Type 1 for changing maximum data sizes
Range of valid answers: [0] 1

The default data sizes are listed in the startup screen. It is a good strategy to first press RETURN here and see whether the default data sizes are sufficient; if not, CANOCO reports so

and may suggest the values you need. If you need to change the size, CANOCO asks the questions:

- Q 5 Type the number of active samples and the number of passive samples
Ranges of valid answers: 3 [20000] 25000, AND 0 [5000]
- Q 6 Type number of species
Range of valid answers: 1 [5000]
- Q 7 Type number of covariables and the number of environmental variables
Ranges of valid answers: 0 [100] 1000, AND 0 [100]

To get the full output of CANOCO, the number of environmental variables should be 8 higher than the number actually analyzed. Also, the number of covariables should exceed the number in the covariable data file in some analyses, within which CANOCO internally generates covariables:

- in forward selection, the number of covariables should exceed the number of selected variables.
- detrending-by-polynomials requires 12 extra covariables and
- permutation under the full model uses as many extra covariables as there are environmental variables.

7.5 Selecting data sets and analysis type

In this subsection, CANOCO asks for the data set(s) that you wish to analyze and the type of analysis.

- Q 8 Type name of file with species data
dunespe.dta

The suggested file name comes from the CANOCO initialization file CANOCO.INI. Ordination (PCA, CA, DCA, RDA, CCA, etc.) is applied to the data of the file specified here. In community ecology the data file is typically the species data, but if one wants, for example, a PCA of environmental data, the environmental data file should be specified here. If the name is a valid name of an existing file, CANOCO will attempt to read the species data from it. The data can be either in Cornell condensed format, full format or in free format (see the chapter Data input). Note that the data should not contain negative values if a unimodal method (CA, DCA, CCA, DCCA) is chosen; if a negative value is encountered, CANOCO stops with an error message saying so.

- Q 9 If you wish to eliminate effects of external variables from the ordination (e.g. blocks in experiments, seasons, salinity or other background variables),
type name of file with covariables (S to Skip)

If you specify a name of an existing file here, then the ordination needs the prefix "Partial": the effects of covariables are partialled out from the ordination diagram. If a file name is suggested, but you do not want any, type a single S. With covariables, CANOCO will give an ordination of the residual variation in the species data that remains after fitting the effects of the covariables. The ordination axes will be made uncorrelated to the covariables. Further, environmental variables (if present) will be regressed on the covariables and the residuals of these multiple regressions will take the place of the original environmental values. In this way, the effect of the environmental variables on the species is "corrected" for the effect that the

covariables have on the species. Constrained ordination axes will therefore represent the effect that is "uniquely" attributable to the environmental variables- and not to (linear combinations of) covariables. With environmental variables in the analysis, covariables play the role of concomitant regressors in the multiple regression of the ordination axes.

The allowed data formats are the same as for species data. In addition, the user may also specify here the solution file from a previous analysis. The first set of "Sample scores" in the file will then be treated as the values of four covariables named AX1, AX2, AX3 and AX4. By specifying the solution file of a previous analysis CANOCO can extract further axes beyond the first four. If, for example, the covariables are the first four ordination axes, then the four ordination axes to be extracted will be made uncorrelated to these covariables and will thus be equivalent to ordination axes 5 to 8 of the previous analysis. See section 8.2.4.2 for further details.

The file with covariables may contain more variables than can actually be analyzed by CANOCO, provided the extra variables are deleted later on.

Q 10 If you wish to relate the ordination axes to external variables,
type name of file with environmental data (S to Skip)
duneenv.dta

The data of the file specified here are used to interpret or to constrain the ordination of the data of Q 8. In community ecology, the data of Q 8 are typically species data and the data specified here are typically environmental data. In general the file should contain "external" explanatory variables, i.e. variables by which one wants to explain the variation in the data specified in Q 8. Explanation proceeds by way of a multiple regression of each ordination axis on the explanatory variables and by way of correlation coefficients. In a partial RDA and CCA the file should contain the explanatory variables of primary interest.

If you specify name of an existing file here, CANOCO will attempt to open the file and read the environmental data from it. The data can either be in Cornell condensed format, in full format or free format. In addition, the user may also specify here the solution file of a previous analysis. The first set of "Sample scores" in the file will then be treated as the values of four environmental variables named AX1, AX2, AX3 and AX4.

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

Q 11 Type name of print file
CANOCO.OUT

If the name is a valid file name, CANOCO attempts to create a new file with this name, and write output to this file. Any existing file with the same name is overwritten without warning. The lines of this output file have a maximum length of 132 characters.

Q 12 Type name of solution file for CanoDraw or other program
CANOCO.SOL

If the name is a valid file name, CANOCO attempts to create a new file with this name. Any existing file with the same name is overwritten without warning. CANOCO will write tables with ordination scores to this file. See section 6.3.

Q 13 *** Type of analysis ***

Model	Gradient analysis		
	indirect	direct	hybrid
linear	1=PCA	2= RDA	3
unimodal	4= CA	5= CCA	6
..	7=DCA	8=DCCA	9
	10=non-standard analysis		
Type analysis number			
Range of valid answers:	1	[7]	10
Answer =	7		

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

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

PCA assumes a linear model (row 1) for the relationship between the responses of each species and the ordination axes; CA and DCA assume a unimodal model (rows 2 and 3) for the relationship between the responses of each species and the ordination axes. Ordination axes can be thought of as being theoretical environmental variables or underlying gradients. The linear model is fitted by the method of two-way weighted summation which leads to the least-squares solution. The unimodal model is fitted by the method of two-way weighted averaging. Use of DCA is advised if an ordination by CA shows the arch effect, i.e. if the sample scores on the second ordination axis are approximately a quadratic function of the sample scores on the first axis. (The arch effect is also termed the Guttman effect, Gifi 1990) Use of PCA is advised in particular if in ordinations by CA or DCA the range of the sample scores is less than 1.5 SD. This advice is applicable to each choice between techniques of rows 1, 2 and 3.

Other names for CA are reciprocal averaging and - outside ecology, in particular when analyzing nominal response variables - dual scaling, optimal scaling, homogeneity analysis and multiple correspondence analysis (Gifi, 1990; Greenacre, 1984).

The choice among variants of PCA, such as non-centered PCA, species-centered PCA, standardized PCA, double centered PCA or log-contrast PCA, is considered in section 7.12 as these variants are obtainable by transformation of the species data. Principal coordinates analysis can also be obtained as a variant of PCA (see section 3.12).

CANOCO calculates in a simple run at most four ordination axes. See Q 9 and Q 46 for methods to obtain further ordination axes with CANOCO.

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

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

When CCA is applied to nominal response variables it is termed redundancy analysis for qualitative variables (section 3.10; Israels, 1984).

The maximum number of constrained ordination axes (= canonical axes) is in general equal to the number of environmental variables, unless "detrending" is in force (see for this exception Q 15). Because CANOCO calculates in a single run at most four ordination axes, CANOCO will in general determine four constrained ordination axes, unless the number of environmental variables (q) is less than 4. If q is less than 4, CANOCO will extract, after the q constrained ordination axes, one or more unconstrained ordination axes (see below).

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

Analysis number 10 stands for nonstandard analysis in which the user can specify unusual options or unusual combinations of options (see section 7.20). This option is included for completeness of CANOCO as a tool in methodological research of ordination methods. Its use is not recommended in any other context. Usage of this option is at the full risk and responsibility of the user.

The methods of row 1 will be called linear methods while the methods of rows 2 and 3 will be called unimodal methods or weighted averaging methods, in accordance with the terminology in Ter Braak & Prentice (1988).

7.6 Number of canonical axes and detrending options

Q 14 Type number of canonical axes (1,2 or 3)
 Range of valid answers: 1 [2] 3

This question is posed only for hybrid gradient analyses. See Q 13 for explanation.

Q 15 Type 1 for detrending by segments
 2 for detrending by 2nd order polynomials
 3 for ,, ,, 3rd order ,,
 4 for ,, ,, 4th order ,,
 Range of valid answers: [1] 4

This question is asked in DCA, DCCA and analysis numbers 9 and 10. Detrending is a method for removing the arch effect in CA and CCA (see Q 13). Detrending-by-segments is the method of detrending proposed by Hill & Gauch (1980) and used in the computer program DECORANA (Hill, 1979). Minchin (1987) found that this method sometimes flattens out some of the variation associated with one of the underlying gradients. He ascribed this to an instability in the detrending-by-segments method (see also Kenkel and Orloci, 1986). Detrending-by-polynomials is intended to be a more stable method of detrending. In the usual reciprocal algorithm of CA, trial site scores for a particular axis are made uncorrelated to the ordination axes already extracted in each iteration step. With detrending-by-polynomials they are also made uncorrelated to k-th order polynomials of the axes already extracted (k = 2, 3 or 4) and to first-order cross products of these axes.

When the arch effect crops up, the second CA-axis is approximately a quadratic function (= a second-order polynomial) of the first CA-axis. Detrending by second-order polynomials therefore specifically removes the arch effect. But this may not be enough because when there is

a dominant first gradient, the third CA-axis is also a function of the first axis, namely a cubic function; and the fourth axis is a quartic function, etc., which may also obscure a true second underlying gradient. As the eigenvalues of these polynomial axes steadily decrease, detrending by fourth-order polynomials is presumably sufficient in most applications.

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 analysed 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 the default in DCA since CANOCO 3.0.

In DCCA and partial DCA, the method of detrending-by-segments is unattractive on theoretical grounds, but the method of detrending-by-polynomials can be modified into an acceptable method (see the Appendix of Ter Braak & Prentice (1988); *Unimodal Models*: p. 137). Use of detrending-by-segments is therefore not recommended in DCCA and partial DCA. When detrending-by-polynomials is used in a direct gradient analysis, then the number of canonical axes that can be extracted is less than without detrending. Less than four canonical axes can be extracted if there are less than 10, 13 or 16 environmental variables depending on whether the order of polynomials is 2, 3 or 4, respectively. Detrending is, however, almost never needed in CCA if only a few environmental variables are included in the analysis. Moreover if the arch effect does occur in a CCA, it is an indication that some environmental variable is superfluous.

Q 16 - Q 19 are asked only in the long dialogue for DCA and DCCA with detrending-by-segments. In the short dialogue, CANOCO uses the default values. The defaults can be modified by using the CANOCO initialization file .

Q 16 Specify number of segments for use in the detrending process
Range of valid answers: 10 [26] 46

In order to carry out the detrending process, the systematic relation of the trial scores with each of the existing axes must be determined. For this, a weighted running-means smoother is used in which the axes are divided in a number of segments (Hill, 1979; Jongman et al. 1987). In this question you can specify the number of segments. The default value is 26. The maximum permissible value is 46. CANOCO uses the same subroutine as used in DECORANA (Hill, 1979), except for two minor changes in response to the criticisms of Oksanen & Minchin (1997). See also the section 3.13.

Q 17 Is nonlinear rescaling of axes required?
type 0 (no rescaling), or number of times to be done
Range of valid answers: 0 [4] 20

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

Q 18 Specify rescaling threshold
Range of valid answers: [0] 100.0

Hill (1979) writes: "If the rescaling threshold is set to t , then axes with length less than t SD will not be rescaled, while those with length greater than t will be rescaled. The default value is $t = 0$ ". Here SD stands for the Standard Deviation unit, a measure of the length of an ordination axis compared to the average breadth of the species' response curves; (see also section 6.2.5 near Table 6.9).

Q 19 Type number (1-4) of axes for species-environment biplot
Range of valid answers: 1 [2] 4

Answer here the number of axes of a planned ordination diagram. The question is needed when detrending-by-segments is in force, because the ordination axes are then in general slightly correlated. The optimal biplot scores for the environmental variables will therefore depend on the number of axes chosen.

7.7 Forward selection of environmental variables

Q 20 Type 1 for forward selection of environmental variables
Range of valid answers: [0] 1

Ask for forward selection of environmental variables

- to find a minimal set of variables that explain the species data about as well as the full set,
- to rank environmental variables in importance for determining the species data, and/or
- to determine the statistical significance of the effects on the species of a particular environmental variable, either unconditionally or conditionally on the effects of some other environmental variables (without the need to specify these variables as covariables).

Selection of variables is a standard topic in books on multiple regression, e.g. Montgomery & 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. It tests the effect of the variable given the effects of the environmental variables that are already selected. When applied repeatedly and in a stepwise fashion, the test shares the shortcomings of the usual tests, in that the overall size of the test is not controlled. In practice this means that too many variables will be judged significant, or equivalently that the tests are too tolerant overall.

The questions asked during the forward selection process are given in section 7.18 on page 216.

Whereas the forward selection option of CANOCO is excellent for determining the statistical significance of any single environmental variable, it is inappropriate to judge the significance of all environmental variables jointly. For joint tests, choose the Monte Carlo test later on (sections 7.14 and 7.17)

7.8 Scaling of ordination scores

CANOCO has six ways to scale ordination scores. This section explains all six and gives guidelines to choose among them. All scalings yield the same **ordering** of ordination scores and the same Summary of the ordination. Different scalings yield, however, a different amount of scatter along the one ordination axis relative to that along another axis. As a result, the scaling influences some aspects of the interpretation of ordination diagrams. The differences in interpretation are minor if the ratios of the eigenvalues are close to 1.

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

```
Q 21 *** 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
```

This question is asked for linear methods, in the long dialogue only. Specify here whether you predominantly want to interpret relationships among samples (scaling ± 1) or among species (scaling ± 2) from the ordination diagram (or whether you prefer the symmetric scaling ± 3). Your choice is unimportant if the ratio of eigenvalues of the axes is close to 1. The relationships among samples are expressed as Pythagorean distances (scaling ± 1), those among species as correlations (scaling 2), or covariances (scaling -2). The species scores are divided by the standard deviation of the species if you type a positive scaling number and are left untransformed if you type a negative scaling number. Your answer here sets the value of α in the chapter Results, section Solution file, subsection Species scores (see Table 6.25 - Table 6.28). The choice of scaling is discussed in Unimodal Models, on pages 144-152.

Untransformed, a species' score is proportional to the standard deviation of the species. Thus, species with a large variance (often the dominant species) lie far from the centre of the ordination diagram and so unduly dominate the diagram. To counteract this effect and to make the species scores more comparable, you can opt here to divide them (after extraction of the axes) by their standard deviation. Then, the ordination diagram displays standardized species data, and correlations instead of covariances. In scaling 2, a **correlation biplot** is obtained; the length of a species' arrow is then the multiple correlation R of the species with the ordination diagram. If the analysis is centred and standardized by species (PCA/RDA on a correlation matrix), then positive scalings give the same result as the corresponding negative scaling.

Nominal environmental data define groups of samples. Scalings 1 and -1 then allow you to interpret the distances between the groups. With **quantitative environmental data**, scaling 2 results in an ordination diagram that reflects the environmental data and the correlations among the environmental variables. However, environmental effect sizes are best inferred from diagrams in scaling 1 or -1. With both nominal and quantitative environmental data, either scaling may be appropriate. Scaling 3 is intermediate between 1 and 2. It does not have any extra mathematical optimality, but may be convenient as a compromise. Irrespective of your choice of scaling here, the ordination diagram displays the major patterns in the species data table, the table of correlations between species and quantitative environmental variables (see section Table 6.29) and, for nominal environmental data, the tables of class means per species (Table 6.30), all interpreted by the **biplot rule**.

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. Partial correlations are more difficult to interpret, because the scale of the species and the 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 the species and the environmental variables after fitting the covariables). Especially if the residual variances are small, partial correlations are less stable than partial covariances.

Q 22 *** Scaling of ordination scores ***
 1 = sample scores are weighted mean species scores
 2 = species scores are weighted mean sample scores
 3 = symmetric scaling
 Type corresponding negative number for Hill's scaling
 Range of valid answers: -3 [2] 3

This question is asked for unimodal methods, in the long dialogue only. Specify here whether you predominantly want to interpret relationships among samples (scaling ± 1) or among species (scaling ± 2) from the ordination diagram (or whether you prefer the symmetric scaling ± 3). In scaling ± 1 , the (species-derived) sample scores are weighted mean species scores, i.e. species that occur in a sample will lie around that sample's point in the ordination diagram. In scaling ± 2 , the species scores are weighted mean sample scores, i.e. each species' point will be at the centre of its niche in the ordination diagram; samples in which a species occurs are scattered around it. These interpretations of weighted averages form the **centroid principle**. Your choice is unimportant if the ratio of eigenvalues of the axes are close to 1. 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 (see section 6.3.5). The choice of scaling is discussed on pages 163-173 of Unimodal Models.

The relationships among samples and among species are expressed as chi-square distances in the positive scaling numbers 1 and 2, respectively (see Table 6.34 and Table 6.37). Among the negative scaling numbers, scaling -1 gives ecological distances among samples expressed in standard deviation units of species turnover (SD-units; see text around the Table 6.9). With scaling -2, the distances among species are generalised Mahalanobis distances. Scaling ± 3 is intermediate between ± 1 and ± 2 . It does not have any extra mathematical optimality, but may be convenient as a compromise.

The sign of the scaling number determines also how to infer the species data from the species-sample plot, other than by the centroid principle. Positive numbers yield a biplot scaling which gives a more quantitative interpretation by the **biplot rule** and is most suited for short gradients. In the biplot scaling the values that are approximated for a species are proportional to its relative abundance y_{ik}/y_{i+} (see Table 6.33 and Unimodal Models p. 171). Negative numbers yield Hill's scaling, which equalizes the average niche breadth for all axes and thus allows, for long gradients (strong **unimodal** response), the **distance rule**. This rule extends the centroid principle by taking a species' point as the optimum of its unimodal response.

Nominal environmental data define groups of samples. The scaling ± 1 then allows you to interpret the distances between the groups. With **quantitative environmental data**, scaling 2 results in an ordination diagram that reflects the environmental data and the correlations among the environmental variables. However, environmental effect sizes are best inferred from diagrams in scaling -1. With both nominal and quantitative environmental data, either scaling may be appropriate. Irrespective of your choice of scaling here, the ordination diagram displays the major patterns in the species data table, the table of weighted averages of the species with respect to quantitative environmental variables (Table 6.38 and Table 8.3) and the relative abundances of species across environmental classes (Table 6.39).

In DCA and DCCA with detrending-by-segments, this question is not asked, because there is only one scaling available: the original scaling used in DECORANA (Hill, 1979) and also described in Jongman et al. (1987: p.106). This scaling is akin to scaling -1.

7.9 Ordination diagnostics

There are three types of statistics: measures of fit for species, residual distances for samples, tolerances for species ("niche widths") and heterogeneity for samples. Tolerance and sample heterogeneity are not defined in PCA/RDA. The fit measure and residual distance are not available in DCA (segments).

Ordination diagnostics indicate how well or how badly individual species and samples are represented in the ordination diagram.

```
Q 23 *** Species and sample diagnostics ***
0 = no diagnostics
1 = fit and residual distances
Range of valid answers:      0      [3]
```

This question is asked in the long dialogue for linear methods. Measures of fit for species and residual distances for samples are reported by typing 1, 2 or 3, or by pressing RETURN. The default value for the answer is 3 instead of 1 for uniformity with the same question in unimodal methods (Q 24).

```
Q 24 *** 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]
```

This question is asked in the long dialogue for unimodal methods. You can request tolerance measures and/or for measures of fit, that are related to chi-square statistics. The tolerance measures are based on the unimodal model of species response. The chi-square measures are derived from a linear model for relative abundance data (see *Unimodal Models* pp. 167). In consequence, the tolerance measures are most useful for long gradients (> 4 SD), whereas the chi-square measures are most useful for short gradients (< 4 SD).

7.10 Omitting samples and selecting explanatory variables

```
Q 25 Enter numbers (not names) of samples to be omitted
One at a time, ending list with a zero
Range of valid answers:      [0]      n
n = highest identification sample number in the species data
```

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

```
4
7
10
0
```


If no samples are to be omitted, then simply press RETURN. Samples can also be omitted at a later stage (see Q 40). The advantage of doing it here is that omitted samples are skipped when reading the environmental data and covariables.

```
Q 26 *** Select/omit:                covariables ***
Type -1 to omit   particular variables
      1 to select   ' '           ' '
Press RETURN to select all variables
Notice: If you later wish to carry out Monte Carlo permutations tests
        within blocks, select covariables that define blocks first.
        If you wish to include all covariables, still choose select
        if block variables are not the first ones in the data file,
        and then select the block variables first.
```

If you specified a file with covariable data, CANOCO asks whether or not you wish to use all the variables in the file as covariables. You may select a subset of variables, or omit variables you do not want to use as covariables. CANOCO asks this question just before the actual reading of the file with covariables.

The notice under the question is important only if you want to determine the statistical significance of the environmental variables by a test that requires permutations within blocks. In CANOCO, blocks are indicated by covariables. If there are more covariables than the ones indicating blocks, the block covariables must come first. This can always be achieved by selecting the block covariables before the remaining ones. See the example project file E40_NP.CON in section 8.3.3.

```
Q 27 Enter numbers (not names) of                covariables to be omitted
One at a time, ending list with a zero
Range of valid answers:      [0]      p
p = highest identification number of the covariables
```

If you answered -1 to the previous question, indicate here which variables should not be used, one per line and ending the deletions by pressing RETURN (see Q 25).

```
Q 28 Enter numbers (not names) of                covariables to be selected
One at a time, ending list with a zero
Range of valid answers:      [0]      p
p = highest identification number of the covariables
```

If you answered 1 to Q 26, indicate here which variables should be used, one per line and ending the selections by pressing RETURN (see Q 25).

```
Q 29 *** Interactions of                covariables ***
Enter pairs of numbers of                covariables to define product variables
Press RETURN to continue
Ranges of valid answers:      [0]      p      , AND      [0]      p
p = highest identification number of the covariables
```

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

```
2 3
2 2
```

-1 0 (or merely press RETURN)

has the effect that CANOCO creates three new variables with identification numbers 21, 22 and 23. Variable 21 is obtained by calculating for each sample the product of its MOISTURE value and its MANURE value. Variable 22 will contain squared moisture values and variable 23 will contain the cubed MOISTURE values. It is also possible to use numbers of variables that are in the data file but were deleted or not selected in the previous questions.

Squares (and products) of covariables may be useful in partial CA or partial DCA to prevent the ordination axes from being a quadratic function of the covariables. This may happen if the covariables represent a long gradient in the species data and subsequent gradients are much shorter. See the DETRENDING question Q9 and section 8.2.4.3.

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

After reading the data, CANOCO lists on the screen which variables are used as covariables, e.g.

```
*****
                          Names of covariables
*****
SF          BF          HF          NM
*****
```

Q 30 *** Select/omit: environmental variables ***
 Type -1 to omit particular variables
 1 to select ,, ,,
 Press RETURN to select all variables
 Range of valid answers: -1 [0] 1

If you specified a file with environmental data, CANOCO asks whether or not you wish to use all the variables in the file as environmental variables. You may select a subset of variables, or omit variables you do not want to use as environmental variables. CANOCO asks this question just before reading the file.

Q 31 Enter numbers (not names) of environmental variables to be omitted
 One at a time, ending list with a zero
 Range of valid answers: [0] q
 q = highest identification number of the environmental variables

If you answered -1 to the previous question, indicate here which variables should not be used, one per line and ending the deletions with a zero or just by pressing RETURN (see Q 25).

Q 32 Enter numbers (not names) of environmental variables to be selected
 One at a time, ending list with a zero
 Range of valid answers: [0] q
 q = highest identification number of the environmental variables

If you answered 1 to Q 30, indicate here which variables should be used, one per line and ending the selections with a zero or just by pressing RETURN (see Q 25).

Q 33 *** Interactions of environmental variables ***

Enter pairs of numbers of environmental variables to define product variables
Press RETURN to continue

Ranges of valid answers: [0] q , AND [0] q

q = highest identification number of the environmental variables

This question is analogous to Q 29, but is asked now for environmental variables, when present.

By defining product variables, the user can investigate in very much the same way as in multiple regression analysis whether the effect of one variable depends on the value of another variable (see Jongman et al. 1987, section 3.5.4). In other words, this is a way to investigate *interaction* of effects. With, for example, $P * N$ the product of the variables P and N , the effect that P has on the species can be shown to depend on the value of N , if the first eigenvalue of the analysis turns out to be considerably higher than in the analysis without this product variable, or if the t -value associated with this product variable is appreciably larger than 2 in absolute value. You can also determine the statistical significance of the interaction effect by a Monte Carlo permutation test. For this, use the original variables (P and N) as covariables (see section 8.3.3), or choose Forward selection and test the product variable $P*N$ after including P and N in the model.

Inclusion of squared variables may alleviate the restriction that only linear combinations of environmental variables are considered in the analyses provided by CANOCO. The user should, however, be cautious in defining too many product variables, to avoid "data dredging".

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

7.11 Transformations of species data

Q 34 *** Transformation of species data ***

0 = no transformation

1 = $\ln(Ay+B)$ -transformation

2 = squareroot-transformation

3 = piecewise linear transformation

Range of valid answers: [0] 3

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. Taking logarithms turns linear models into ecologically more plausible, multiplicative/exponential models. The values for A and B in the logarithmic transformation are asked next (Q 35). The variance of count data can be stabilised by taking square-roots. You can also specify your own transformation after typing a 3 (piecewise linear transformation). The transformation that is chosen is applied to all species (in general terms: to all response variables).

Q 35 Type values of A and B for use in $\ln(Ay+B)$ -transformation

Ranges of valid answers: -999.9 [1.0] 999.9, AND 0.0 [1.0] 999.9

If the species data are strictly positive ($y > 0$), use $A = 1$ and $B = 0$. However, usually species data contain zero values. Then, a small value ($B > 0$) must be added, because $\log(0)$ is undefined. For technical reasons, B must then be greater than 1 in Canoco, but this limitation can be circumvented by specifying a value for $A > 1$. For example, if you would like to add 0.1 to the original data, specify $A = 10$ and $B = 1$. See example section 8.4.2.

Q 36 Enter couplets of old and new values for
piecewise linear transformation, ending with -1 0
Ranges of valid answers: -1.0 [-1.0] 999.9, AND [.0] 999.9

The piecewise linear transformation works as in DECORANA (Hill, 1979). The following description of this transformation is taken with minor modifications from the DECORANA manual. A typical transformation might be

```
0 0
0.1 1
2 2
5 3
10 4
20 5
-1 0 (or merely press RETURN)
```

The negative number -1 serves to terminate the transformation data, and it must be followed by a dummy value such as 0. The meaning of this transformation is that a quantity 0 in the data is transformed to 0, 0.1 to 1, 2 to 2, 5 to 3, etc. For other numbers, the transformation is interpolated linearly. Thus 6.9 is transformed to

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

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

```
20.3 5.2
```

would be a perfectly acceptable couplet.

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

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

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

"Three restrictions should be noted:

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

```
0 0
0.1 1
1 2
```

then this can be put to rights by typing the couplet

0 0

which is not in ascending order, and which therefore nullifies the transformation that has been fed in so far.

3. Not more than 46 couplets can be entered to define the transformation. If more are entered, the program will proceed to the next stage regardless."

Usually a condensed format file does not contain zero abundance values. However, sometimes it does, so as to indicate that the species occurred in a tiny amount. Such zero values are stored when the file is being read as species data file and such zeroes can thus be transformed for the analysis in a non-zero value, for example with the piecewise linear transformation specified by

0 1
2 2
5 3
10 4
20 5
-1 0

In this transformation the explicit zeroes in the condensed format file are transformed to the value 1.

If the minimum abundance value is ≥ 1 , the transformation

0 0
1 1
-1 0

transforms abundance to presence/absence.

Q 37 Type weight to be given to
* species * that you will be asked to specify next
Type 0.01 to make species passive
Type 0 to delete species
Range of valid answers: .0 [1.0] 100.0

Weights (w^*) can be assigned to species in order to give particular species more ($w^* > 1$) or less ($w^* < 1$) emphasis in the analysis, to delete particular species ($w^* = 0$) or to make particular species supplementary ($w^* = 0.01$). Supplementary species are also called passive. Press RETURN if you do not want to specify (other) non-default weights. A "supplementary" species has no influence on the extraction of the ordination axes, but is added to the ordination afterwards by use of the transition formulae (see section 6.3; see also Jongman et al., 1987; exercises 5.2 and 5.3). For $w^* > 0.01$, the weight of a sample can be interpreted as "the number of times" the species is included in the analysis. For example, if $w^* = 2$ for a species, the same ordination could also have been obtained from an unweighted analysis by including that particular species twice in the data file(s). This interpretation is, of course, strictly valid only for integer weights ($w^* = 1, 2, 3, \dots$), but the mathematics works through equally well for any positive weight.

Q 38 Enter numbers (not names) of species to be weighted
One per line, ending list with 0.
Negative numbers denote sequences. For example
a 4 followed by a -8 weights species 4 through 8.
Range of valid answers: -m [0] m
m = highest species identification number

This question is posed if a non-default weight is specified in the previous question. For example, to give species number 3 and the species numbers 11, 12, 13, ..., 20, 21 double weight,

Q 37 should be answered by typing a 2 and giving a RETURN; then Q 38 appears and should be answered by typing

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

i.e. one number per line. Next, Q 37 appears again in order to allow you to give a different weight to other species. If a species is weighted more than once, the last given weight is decisive.

```
Q 39 Type weight to be given to
* samples * that you will be asked to specify next
Type 0.01 to make samples passive
Type 0 to delete samples
Range of valid answers: .0 [1.0] 100.0
```

This question is similar to Q 37, but now posed for samples. Samples that are made supplementary (passive) are placed after the active samples in the output tables of sample scores.

```
Q 40 Enter numbers (not names) of samples to be weighted
One per line, ending list with 0.
Negative numbers denote sequences. For example
a 4 followed by a -8 weights species 4 through 8.
Range of valid answers: -n [0] n
n = highest sample identification number
```

This question is similar to Q 38, but now applies to samples.

```
Q 41 Weighting of species required?
0 = no
1 = downweighting of rare species
Range of valid answers: [0] 1
```

This question is posed only for unimodal methods and is familiar to users of DECORANA (Hill, 1979). Hill (1979) writes: "In some applications individual samples with rare species may distort the analysis. If it is desired to give rare species less weight, while still retaining them in the analysis, then the downweighting parameter can be set to 1. Let AMAX be the frequency of the commonest species. Then the effect of downweighting is to reduce the abundance of species rarer than (AMAX/5) in proportion to their frequency. Species commoner than (AMAX/5) are not downweighted at all." For further details see Hill (1979). Downweights are similar in interpretation as the weights in Q 37. Their joint effect is multiplicative. The downweight times the weight given in Q 37 - Q 38 is listed for all species identification numbers on the output file in a format of 20 species per line.

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

7.12 Centering and standardization in linear methods

Q 42 *** Centering/standardization by species ***
 0 = none (non-centered PCA)
 1 = centering (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

This question is posed for linear methods and defines, in conjunction with the next question, which variant of PCA/RDA is chosen (Table 7.2). The default is centering by species only. Let y_{ik} be the current value of species k in sample i ($i=1, \dots, n; k=1, \dots, m$) and let w_i^* be the weight of sample i . Unless requested otherwise in Q 39 - Q 40: $w_i^* = 1$. By answering 1, 2 or 3 the value of y_{ik} is replaced by the value of

$$(7.1) \quad y'_{ik} = y_{ik} - \sum_i w_i^* y_{ik} / \sum_i w_i^* \quad (\text{answer} = 1)$$

$$(7.2) \quad y_{ik} / (\sum_i w_i^* y_{ik}^2)^{1/2} \quad (\text{answer} = 2)$$

$$(7.3) \quad y'_{ik} / (\sum_i w_i^* y'_{ik}{}^2)^{1/2} \quad (\text{answer} = 3)$$

where y'_{ik} is the value obtained after centring by species (see this question, answer = 1). Note that in RDA centering by species is implicit because of the intercept in the regression of the sample scores on the environmental variables.

With environmental data it is possible to weight species inversely to the error variance that remains after fitting the species to the environment and covariable data (answer = 4). To put this option into context, let us compare RDA with canonical correlation analysis. 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 (1990), is to weight each species inversely to its error variance. CANOCO incorporates this solution and reports the relative weights given to species on the output file, in the solution file (as weights for species alongside the species scores; see section 6.3.4) and the species-environment file, SPEC_ENV.TAB (see section 6.4). If the R^2 of a species exceeds 0.9, then its weight is truncated as if the R^2 were equal to 0.9. This is done to avoid extreme weights (larger than 10) for species that happen to fit extremely well.

The technical details are as follows. 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 onto 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 re-estimated 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. Advantages of the CANOCO approach are that it does not depend on the reduced rank assumption and its solution is much simpler, admitting direct rather than iterative computation.

Q 43 *** Centering/standardization by samples (in the species data) ***

0 = none (standard)
1 = centering (fine for log-percentage data)
2 = standardization by sample norm
3 = both 1 and 2
Range of valid answers: [0] 3

This question is posed for linear methods and defines, in conjunction with the previous question, which variant of PCA/RDA is chosen (Table 7.2). The default is that neither centering nor standardization by samples is applied. Let y_{ik} be the current value of species k in sample i ($i = 1, \dots, n; k = 1, \dots, m$) and let w_k^* be the weight of species k . Unless requested otherwise in Q 39 - Q 40: $w_k^* = 1$. By answering 1, 2 or 3 the value of y_{ik} is replaced by the value of

$$(7.4) \quad y'_{ik} = y_{ik} - \frac{\sum_k w_k^* y_{ik}}{\sum_k w_k^*} \quad (\text{answer} = 1)$$

$$(7.5) \quad y_{ik} / (\sum_k w_k^* y_{ik}^2)^{1/2} \quad (\text{answer} = 2)$$

$$(7.6) \quad y'_{ik} / (\sum_k w_k^* y'_{ik}{}^2)^{1/2} \quad (\text{answer} = 3)$$

where y'_{ik} is the value obtained after centering by samples (see this question, answer = 1).

The centering / standardizations by species and by samples may interact. It should be noted that CANOCO first centers and standardizes by samples and then by species.

How the centering and standardization questions of linear methods (Q 42 and Q 43) interact with Q 37 - Q 40, can be deduced from the following example. Suppose Q 42 = 3, i.e. the values of each species are standardized to mean 0 and variance 1, so that they have equal weight (in a particular sense). If a species is now given double weight in Q 37- Q 38, the weighted analysis gives the same results as an unweighted analysis in which that species is included twice in the data and the same standardization (Q 42 = 3) is in force.

After centering and standardization by samples and species, CANOCO calculates the Total Sum of Squares of the species data (TSS) and the total standard deviation in the species data by

$$(7.7) \quad \text{TSS} = \sum_i \sum_k w_i^* w_k^* y_{ik}^2 \quad \text{and} \quad \text{TAU} = \{\text{TSS} / \sum_i \sum_k w_i^* w_k^*\}^{1/2}$$

Subsequently, all species values are divided by TAU. After division, the total mean square of the species data is equal to 1. This has the advantage that the eigenvalues reported by PCA and RDA are **fractions** of the total sum of squares and that the sum of all eigenvalues in PCA is equal to 1.⁴ When multiplied by 100, these fractions are usually referred to as percentages of variance accounted for by the ordination axes. Note that in a partial PCA and in a RDA the sum of all eigenvalues is less than or equal to 1.

⁴ Except when centering by samples is used (Q 43 = 1 or 3) in conjunction with standardization by species (Q 42 = 2 or 3). Use of these exceptional cases is discouraged and they are not available in Canoco for Windows; the sample means are equal to 0 after Q 43 but not after Q 42; the iterative ordination algorithm will nevertheless calculate an analysis centered by samples. The eigenvalues are fractions of TSS as defined in (7.7) but do not sum to 1.

Table 7.2 Variants of PCA (also available in RDA if Q 42 = 1 or 3).

Answers to:	Q 43	Q 42	Q 21	
	samples	species	scaling	Interpretation of ordination diagram by distances [points] and arrows [inner products or angles]
	cen/stan	cen/stan		
Ordinary PCA	0	1	1	Pythagorean distance between samples [points] (a,c)
	0	1	2	covariances between species [arrows] (b)
Standardized PCA	0	3	1	standardized Pythagorean distance between samples [points] (c)
	0	3	2	correlations between species [arrows] (b)
Double centered PCA	1	1	3	after ln-transformation: appropriate for percentage data(e; see section 3.9.2) and can fit a unimodal model (d)
PCA standardized by sample norm	2	0	1	cosine theta distance (c) between samples [points] = angular separation (f)
PCA standardized by sample norm and centered by species	2	1	1	cosine theta distance (c) between samples [points]
PCA centered and standardized by samples	3	0	1	"correlation coefficient" between samples [arrows] c,f); controversial!
Noncentered PCA	0	0	1;3	(g,h)
Principal coordinate analysis	1	1	3	dissimilarity between sites when input is -(squared dissimilarity) between samples(section 3.12)

References:

(a) Jongman et al. (1987); (b) Corsten and Gabriel (1976); (c) Prentice (1980); (d) Kooijman (1977);(e) Aitchison (1982); (f) Gordon (1981); (g) Noy-Meir (1973); (h) Ter Braak (1983).

7.13 Output options

```
Q 44 **** Output option for ****
Correlation matrix of eigenvectors and environmental variables
Type 0 for no output
    1 (4) output on file CANOCO.OUT
or the value between brackets for output to the screen as well.
Range of valid answers:    0 [4]
```

This question, posed only if there are environmental variables in the analysis, allows you to see the correlation matrix, means and standard deviations of eigenvectors (ordination axes) and environmental variables at the screen and to write them to the print file specified in Q 11. If there is not enough data space available to calculate the (full) correlation matrix some ordination results cannot be computed and a warning is given.

Q 45 **** Output option for ****

Ordination results

Type 0 for no output

1 (4) output on file CANOCO.OUT

2 (5) output on file CANOCO.SOL

3 (6) output on both files

or the value between brackets for output to the screen as well.

Enter your choice for each item, all on one line, e.g. 2222222 <RETURN>

Press RETURN for the defaults indicated below the items.

Spec-scor	Samp-scor	Regr-coef	t-values	Inter-cor	Envi-bipl	Centroids	Linea-com
2	2	2	2	2	2	2	2

This question allows you to see the ordination results of the analysis at the screen and to write them to the print file (Q 11) and/or to the solution files (Q 12). The default is that all output is written to the solution file (section 6.3). The answers for the items must be entered on a single line, for example,

6 2 1 4 4 6 6 5

ending by pressing the return-key. If more numbers are entered on a line than required, the superfluous numbers are ignored. If a solution file has not been asked for in Q 12, the lines beginning with 2(5) and 3(6) do not appear; if the answer given is nevertheless 2, or 3 (or 5 or 6), then CANOCO acts as if the answer 1 (or 4) was given. If an error is detected in the answer, the question appears again.

The items listed in Q 45 depend on the type of analysis. The abbreviations are (between brackets) the symbols used in section 6.3.

SPEC-SCOR = species scores (u_k)

SAMP-SCOR = sample scores, that are species-derived (x_i^*)

REGR-COEFF = regression coefficients (c_j) of the environmental variables for an unconstrained ordination axis, canonical coefficients (c_j) for a constrained ordination axis, and t-values associated with the regression coefficients c_j in the multiple regression of $\{x_i^*\}$ on $\{z_{ij}\}$

T-VALUES = coordinates for species and environmental variables for the t-value biplot

INTER-COR = inter-set correlations between the environmental variables and the ordination scores $\{x_i^*\}$

ENVI-BIPL = scores of environmental variables for drawing a biplot (suitable for quantitative variables) $\{c_j^*\}$

CENTROIDS = centroids of environmental variables in the ordination diagram (suitable for qualitative (nominal) variables) $\{c_j^+\}$

LINEA-COM = sample scores which are linear combinations of the environmental variables (x'_i)

See section 6.3.2 for further explanation.

Warning: If the number of items (species or samples) is larger than or equal to 2000, their order in the extended output to the print file CANOCO.OUT (Q45) is incorrect. The items numbered 2000 and higher are not sorted alongside the others. Sort the scores in Excel instead.

7.14 Additional analyses

Q 46 Type
0 = stop
1 = more analyses with current data
2 = passive analysis of other environmental variables
3 = as 2, but with regressions
4 = more ordination axes
Range of valid answers: [0] 4

The user can stop the program by pressing RETURN or ask for additional analyses using the current species data and covariables. In additional analyses, the answers concerning data transformation (Q 34 - Q 43) and covariables (Q 26 - Q 29) and the output files remain in force and are not presented again.

After answering 1, the user can determine the overall statistical significance of the environmental variables (Q 51), delete environmental variables (Q 48), designate environmental variables as covariables (Q 47), or modify the type of analysis (Q 13 - Q 24). However, the user cannot switch between linear methods and unimodal methods and cannot delete samples or species. If there are no environmental variables, the program continues immediately with Q 13, or else with Q 47.

After answering 2 or 3, CANOCO asks for a (new) set of environmental variables (with so-called supplementary environmental variables) which are used to interpret the current ordination axes. By answering 3 the current sample scores (x_i) which are linear combinations of the previous set of environmental variables, are replaced by fitted values of the regression of the current sample scores (x_i^*) on the newly entered environmental variables. If the answer is a 2, then the current sample scores x_i and x_i^* remain unchanged. In either case, CANOCO will ask for a file name with environmental data (cf. Q 8), which environmental variables are to be deleted (Q 30), whether interactions are to be included (Q 33), which output is required and how to continue (Q 46).

After answering 4, CANOCO asks how many additional axes you want (Q 50).

Q 47 Enter numbers (not names) of environmental variables
to be turned into covariables, one at a line. Press RETURN to continue
Range of valid answers: [0] q
q = highest identification number of the environmental variables

This question, asked after Q 51 = 0 (no Monte Carlo test), allows you to create one or more covariables from the existing environmental variables without the need to specify a file with covariables. Any environmental variable specified here is removed from the list of environmental variables. If you have already specified covariable data, the variables specified here are placed after the existing covariables.

Q 48 Enter numbers (not names) of environmental variables to be deleted
One at a line. Press RETURN to continue
Range of valid answers: [0] q
q = highest identification number of the environmental variables

This question, asked after Q 47, allows you to delete one or more environmental variables. Thereafter, CANOCO calculates a new ordination according to the current type of analysis using the remaining environmental variables, asks what output is required and how to continue (Q 46). If you do not delete any environmental variables, Q 51 appears in direct or hybrid gradient analyses and Q 13 in indirect gradient analyses.

7.15 Supplementary environmental variables

Q 49 Type name of file with supplementary environmental variables

This question appears after Q 46 = 2 or 3. See Q 10 and Q 46. Supplementary environmental variables are used to interpret the current ordination axes. Later on, they can be turned into active environmental variables so as to extract ordination axes which are linear combinations of them (i.e. to extract new canonical axes).

7.16 More ordination axes

Q 50 How many axes more?

Range of valid answers: 1 [4]

This question appears after Q 46 = 4. If you press RETURN here, CANOCO proceeds to calculate four extra ordination axes (i.e. axes 5 - 8, or after a second time, axes 9-12, and so on). In this process the current eigenvector sample scores are moved to the covariable data.

Asking for one more ordination axis allows you to determine the significance of the second ordination axis of a direct gradient analysis. This works as follows. If you ask for one more ordination axis, CANOCO moves the current first ordination axis to the covariables. The new first axis will be the second axis of the original analysis. The significance of this new first axis can then be determined by Monte Carlo permutation (see Q 51).

7.17 Monte Carlo permutation tests

7.17.1 Introduction

The statistical significance of the relation between the species and the whole set of environmental variables, given the covariables, can be evaluated using Monte Carlo permutation tests. A Monte Carlo permutation test is a test of statistical significance obtained by repeatedly *shuffling (permuting) the samples*.

The validity of a 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), a completely random permutation is appropriate, whereas for a randomized block design the permutation must be within blocks. Data from line transects, time series, rectangular grids, repeated measurement studies (e.g. BACI-designs) require specialized permutation types. CANOCO can automatically generate valid permutation types for such data, when recorded at equal intervals. Since version 4.0, CANOCO can also generate permutations that are appropriate for split-plot designs, nested designs and related balanced multi-level designs. If your data require yet another type, you can provide 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 irregular 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 of overcoming the difficulty of statistical tests in the presence of autocorrelation or spatial correlation (Besag & Clifford, 1989: section 5; Ter Braak, 1980: part II, chapter 3). They thus form a viable alternative for traditional parametric tests based on precise modeling of the autocorrelation structure.

We first describe the questions that are asked in all types of permutation tests (section 7.17.2). This section also covers the permutation test that uses unrestricted permutations. In section 7.17.3 we describe block designs. By defining blocks, exchanges of samples from different blocks can be excluded. Blocks are specified by covariables. With blocks of equal size, the required permutations can sometimes be obtained without block-defining covariables by using the split-plot options (see Q 61). Sections 7.17.4 and 7.17.5 describe more advanced permutation types. The background theory of the permutations tests is described in section 3.7.

7.17.2 Unrestricted permutation; common questions

```
Q 51 *** Monte Carlo permutation test ***
      0 = no significance test
      1 = test of significance of first canonical axis
      2 = test of significance of all canonical axes together
      3 = both 1 and 2
Range of valid answers:      [0]      3
```

This is the first question asked after Q 46= 1 (additional analyses) if there are environmental variables in the analysis. You can 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 trace). The resulting tests determine the significance of the first ordination axis and that of all canonical axes together, respectively.

By answering 1 or 3, the statistical significance of the first ordination axis is determined. The null hypothesis of the test is that, given any covariables, there is no relation between species and environment. The test statistic is an F-ratio of the first eigenvalue and the residual sum of squares. This test statistic has maximum power against the alternative hypothesis that there is a single dominating gradient that determines the relation between species and environment. This test statistic requires more computer time than the overall test.

By answering 2 or 3, the statistical significance of the relation between the species and the set of environmental variables is determined. The null hypothesis of the test is the same as above, namely that, given any covariables, there is no relation between species and environment. The test statistic is an F-ratio of the sum of all canonical eigenvalues (which takes the role of the regression sum of squares) and the residual sum of squares. This test statistic yields an omnibus test, i.e. a test which is sensitive to all kinds of deviations from the null hypothesis.

```
Q 52 *** Type of permutation ***
      0 = permutations read from file
      1 = unrestricted
      2 = restricted for split-plot designs, time series, lines and grids
Range of valid answers:      0      [1]      2
```

Unrestricted permutation (Q 52 = 1) is appropriate for completely randomized designs and for simple random sampling. It is also the default for studies without any additional structure. Restricted permutation types (Q 52 = 2) are appropriate for line transects, time series and rectangular grids, if recorded at equal intervals, and for balanced split-plot designs and related designs such as Before-After-Control-Impact (BACI) designs, repeated measurement designs, and many ANOVA designs with random (nested or crossed) factors. With covariables the question is more extended (Q 55) to allow for block designs.

If your data require yet another permutation type, you can feed permutations from an external file into CANOCO. After answering 0, the only other questions asked are Q 64 and Q 54.

Q 53 Type two integers (1-30000) as seeds for the random number sequence, on a single line or press RETURN for default seeds.
Ranges of valid answers: 1 [23239] 30000, AND 1 [945]

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

Q 54 Type number of random permutations
-number to get permutation under reduced model
Range of valid answers: -9999 [-199] 9999

For a test at the 5%-significance level, minimally 19 permutations are required (the result is then significant if the test statistic for the data is larger than that for any of the 19 permutations, because $1/20 = 0.05$). The power of the test increases with the number of permutations, but only slightly so beyond 199 permutations. As each extra permutation costs computer time, taking a number larger than 199 will not usually be worthwhile (see the discussion in Ter Braak & Wiertz 1994).

By typing a negative number of permutations (as in the default), residuals from the reduced model ("null model") are permuted. By typing a positive number, residuals from the full model are permuted. The reduced model is the current model for the species data with the variables for testing being excluded, whereas in the full model the variables for testing are included. In these definitions, the current model contains the covariables (if any) and, in forward selection, the environmental variables that have already been selected.

The reduced model method better maintains the type I error in small data sets. Without covariables, the method yields the exact Monte Carlo significance level (Hope 1968). The full-model method gives slightly lower type II error. Recent research shows that there is little reason to change the default which permutes the residuals of the reduced model (Anderson & Legendre, 1999).

7.17.3 Specifying blocks

Q 55 *** Type of permutation ***
0 = permutations read from file
1 = unrestricted
2 = restricted for split-plot designs, time series, lines and grids
3 = unrestricted (as 1) within blocks
4 = restricted (as 2) within blocks
Range of valid answers: 0 [1] 4

This question is the extended form of Q 52 and is asked when there are covariables in the analysis. In experimental designs with blocks or sampling designs with strata, exchanges of samples between the blocks or the strata must be excluded. This can be achieved by answering 3 or 4 here and defining blocks by covariables in the next question. If samples are taken in a number of different locations, defining locations as blocks provides a test for common within-location variation.

For blocks of equal size, the required permutations can sometimes be obtained without block-defining covariables. Try option 2 (split-plot design) for this. See the project file `mimicblc.con` in the subdirectory `..\SAMPLES\PERMUTIO\PLOUGH` (section 8.3.2).

```
Q 56 *** Specification of blocks ***
Enter number (not names) of covariables that define blocks,
one at a line. Press RETURN to continue
Range of valid answers:      [0]      10
```

Blocks are groups of samples. Usually each block is indicated by one particular dummy covariable that has the value 1 for samples that belong to the block and the value 0 for other samples. You must indicate here which covariables define the blocks. If the covariable indicating the last block is entered, CANOCO will report that this variable is not in the covariable data, even if explicitly entered as such. The reason is that a covariable that is collinear with the previous ones, is deleted by CANOCO. See section 6.2.2. With three blocks indicated by the variables with identification numbers 11, 12 and 13, for example, CANOCO reports nothing special after entering 11 and 12, but after entering 13, CANOCO reports:

```
Covariable 13 is not in the covariable data.
It may have been deleted by CANOCO as being collinear.
Please try again or press RETURN to continue.
Range of valid answers:      [0]      p
p = highest identification number of the covariables
```

Simply pressing RETURN to continue is all that is needed. From the covariables that you specified, CANOCO determines which samples belong to each block and reports the result in the print file.

If there are more covariables than the ones indicating blocks, the block covariables must come first. This can always be achieved by selecting the block covariables before the remaining ones in Q 28 after answering Q 26 = 1. See the project file `e40_np.con` in the subdirectory `..\SAMPLES\PERMUTIO\E40` (section 8.3.3).

If you specify a quantitative variable to indicate blocks, each different value of the variable will yield a block (which may not be what you intended!). In general, the samples of a block have a unique combination of values on the block-defining covariables.

7.17.4 Restricted permutation types

```
Q 57 *** Type of restricted permutation ***
1 = time series or line transects (cyclic shifts)
2 = rectangular spatial grids (toroidal shifts)
3 = split-plot designs (whole plots with linked split-plots)
Type -1 or -2 to disable random shift of mirror image
Range of valid answers:      -2      [1]      3
```

This question appears after Q 52 = 2 and after Q 55 = 2 or 4. You can specify here whether your data are from a line transect / time series (Q 57 = 1), a rectangular grid (Q 57 = 2), or a split-plot design (Q 57 = 3). If you have multiple transects (or series or grid) of equal dimension, you may also select here the split-plot design option, which encompasses the other options. With the split-plot design you can test for split-plot factors (e.g. within-transect variation) as well as for whole-plot factors (e.g. between-transect variation). With the split-plot design, you can also analyze Before-After-Control-Impact (BACI) designs, repeated measurement designs, and many ANOVA designs with random (nested or crossed) factors.

The permutations for series/transects or grids are cyclic or toroidal shifts. It is rarely needed, but you may disable shifts from the mirror image of the series/transect or grid by typing a negative number.

For the background theory see Section 3.7.

Q 58 Type number of rows of the rectangular grid
Range of valid answers: 1 [2] ng
ng = number of active samples

This question is asked after Q 57 = 2 or -2. Rows and columns of the grid are not arbitrary entities in CANOCO: a row consists of samples that are consecutive in the data file. For example, if you specify that your grid has 4 rows and 14 columns, then CANOCO assumes that the first 14 units (samples or whole-plots) in the data file form the first row, the next 14 the second row, etc. In contrast, if you specify that your grid has 14 rows and 4 columns, then CANOCO assumes that the first 4 units (samples or whole-plots) in the data file form the first row, the next 4 the second row, etc. You can check in the print file under "Sample arrangement in the permutation test" whether CANOCO interpreted your specification as intended.

Q 59 Type number of split-plots per whole-plot
Range of valid answers: 1 [2] ng
ng = number of active samples

This question is asked after Q 57 = 3. A split-plot design is a hierarchical design with two levels of units: whole-plots containing split-plots. Split-plots are the lowest level sampling units, i.e. the samples in your data file. Examples are samples-within-estuaries, plots-within-stands, plots-along-transects, relevés-within-time-series (in a permanent plots study). You must specify here the number of samples per whole-plot. CANOCO will answer, for example,

10 whole plots detected with 2 split-plots each

Q 60 Split-plots per whole-plot are found in the data file by the rule:
take K, skip L.
In the default (1 0) split-plots are contiguous, within blocks if present
Ranges of valid answers: [1] 2 , AND [0] 10

This question is asked after Q 59. You can specify here how the samples forming a whole-plot are arranged in the data file. If the samples of a whole-plot are consecutive in the data file, you can apply the default rule (take 1 sample, skip next 0 samples), because no samples need to be skipped. For example, if in a permanent plot study, the vegetation of 50 locations is monitored at 20 points in time, locations are whole-plots and relevés (samples) split-plots. In the data file, the samples may be arranged by locations or by times. Arrangement by locations means that all the data of a single location are consecutive in the data file, so that the default rule applies. (The rule 'take 20 samples, skip 0' would work as well). Arrangement by times means that all data of a single time point are consecutive in the data file. With a standard order of locations within times, the data of each location are found by the rule 'take 1 sample, skip the next 19 samples'.

Here is an example which would require you to specify a take number other than the default. Let the whole-plots A, B, and C consist of 6 samples each and let the samples happen to be arranged as AAB BCC AAB BCC AAB BCC in the data file. Then, the rule take 2, skip 4 correctly specifies the whole-plots. Such data arrangements occur naturally in ANOVA designs with random crossed factors.

```

Q 61 *** Type of permutation for whole-plots ***
0 = none (mimics blocks without covariables)
1 = time series or line transects (cyclic shifts)
2 = rectangular spatial grids (toroidal shifts)
3 = exchangeable (unrestricted )
Type -1 or -2 to disable random shift of mirror image
Range of valid answers: -2 [3]

```

Having defined what whole-plots are (Q 59 and Q 60), you can specify here how the whole-plots must be permuted. The next question asks how split-plots must be permuted. Blocks can be emulated (without the need of block-defining covariables) by not permuting whole-plots. By this choice (Q 61 = 0), split-plot factors can be tested (a split-plot factor is a variable or set of variables that varies within whole-plots). To test for a whole-plot factor (a whole-plot factor is a variable or set of variables that varies between whole-plots), whole-plots must be permuted. In the standard split-plot design, whole-plots are exchangeable and can be randomly permuted (Q 61 = 3), whereas split-plots are not permuted. The whole-plots may themselves form a time series, a line transect or a grid, so that random permutations may not be appropriate. For such cases, CANOCO can limit the permutations to cyclic or toroidal shifts. It is rarely needed, but you may disable shifts from the mirror image of the series/transect or grid by typing a negative number. After answering ± 2 (grids), CANOCO asks for the number of rows of the grid of whole-plots as in Q 58.

```

Q 62 *** Type of permutation for split-plots ***
0 = none (held together )
-----INDEPENDENT ACROSS WHOLE PLOTS -----
1 = time series or line transects (cyclic shifts)
2 = rectangular spatial grids (toroidal shifts)
3 = exchangeable (mimics plots in blocks)
-----DEPENDENT ACROSS WHOLE PLOTS -----
4 = time series or line transects (cyclic shifts)
5 = rectangular spatial grids (toroidal shifts)
6 = exchangeable
Type -1, -2, -4 or -5 to disable random shift of mirror image
Range of valid answers: -5 [0] 6

```

Having defined how whole-plots must be permuted (Q 61), you can specify here how the split-plots must be permuted. To test for a whole-plot factor in a standard split-plot design, split-plots are kept together (no permutation). To test for a split-plot factor in a standard split-plot design, each whole-plot acts as a block: no permutation of whole-plots, and random permutations within whole-plots. The appropriate answer to Q 62 is then 3. If your environmental variables vary little or not at all within whole-plots, the test will never show significant effects.

If the split-plots form a time series, a line transect, or a spatial grid, the split-plot permutations can be restricted to cyclic or toroidal shifts so as to account for autocorrelation among split-plots. If the split-plots form parallel time series and time is an autocorrelated error component affecting all series, the same shift should be applied to all time series. This is specified by using the “dependent across whole-plots” options (Q 62 = 4, 5 or 6). After answering ± 2 or ± 5 (grids), CANOCO asks for the number of rows of the grid of split-plots per whole-plot as in Q 58. It is rarely needed, but you may disable shifts from the mirror image of the series/transect or grid by typing a negative number.

Many designs can be analyzed in the split-plot framework. Section 8.3 gives examples of specifying tests in repeated measurement designs, ANOVA with random nested and crossed factors and the BACI design.

7.17.5 Restricted permutation types within blocks

This section describes the question after you asked for restricted permutation for line transects, grids or split-plots within blocks (Q 55 = 4). If you asked for permutations for time series or line transects (Q 57 = ±1), CANOCO assumes that each block consists of a time series or line transect, and immediately lists which samples belong to each block. It is therefore essential that the samples are in their natural order in the data file within each block. On the other hand, if you asked for permutations for grids or split-plot designs (Q 57 = ± 2 or 3), CANOCO reports on the screen the samples that belong to the first block only. For example, CANOCO reports

```
Block 1 contains the samples numbered:
  1   2   3   4   5   6   7   8   9  10
 11  12  13  14  15  16  17  18  19  20
 21  22  23  24
```

With grids, these samples form a grid and CANOCO asks for the number of rows of the grid as in Q 58. For the split-plot design, these samples are the split-plots of a smaller number of whole-plots and CANOCO asks for the number of split-plots per whole-plot (Q 59), for the rule to find the samples of each whole-plot (Q 60), and how whole-plots and split-plots must be permuted (Q 61 and Q 62). CANOCO then lists the samples that belong to block 2, for example,

```
Block 2 contains the samples numbered:
 25  26  27  28  29  30  31  32  33  34
 35  36  37  38  39  40  41  42  43  44
 45  46  47  48
```

and asks whether this and later blocks have the same layout:

```
Q 63 Type 1 if this and later blocks have the same layout
      else press RETURN
Range of valid answers:      [0]      1
```

If you answer 1 in response to Q 63, CANOCO assumes, for grids, that the grids in this and subsequent blocks have the same number of rows and, in split-plots designs, that the design of this and later blocks is the same as for the previous one. If you press RETURN instead, CANOCO asks the same question as for block 1 (the number of rows of the grid or the details of the split-plot design in this block), again followed by Q 63.

CANOCO uses the convention that sample 1 is always in block 1; the second block starts with the next higher sample number that is not in block 1, etc.

```
Q 64 Type name of file with permutations
```

After Q 52 = 0 or Q 55 = 0, CANOCO asks for the file with the permutations, which must be a text only file. A very small example file is given in Table 7.3 The example file specifies three permutations of the numbers 1, ... , 20 and can be used in conjunction with the Dune meadow files `dunespe.dta` and `duneenv.dta`. The numbers can be in free format, each one permutation starting at a new line. The numbers are not sample identification numbers, but sequential numbers for the 20 active samples. The numbers 18, 19 and 20 thus refer to Sample 18, Sample 19 and Sample 20 (see Table 16.2). In general, 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 in the solution file (even if these samples have other identifying numbers). To

avoid confusion it is safest to make sample numbers consecutive. It is **not** permitted to specify a bootstrap sample from the numbers 1 to n; the CANOCO algorithm does not allow this. CANOCO will detect an error if the values read do not form a permutation. Another example is given in project file permfile.con in the directory ..\SAMPLES\PERMUTIO\BACII1SPE (section 8.3.7).

Table 7.3 Example permutation file: three permutations of the numbers 1, ..., 20.

```

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

```

After you specified the file with permutations, CANOCO asks:

```

Q 65 Type number of random permutations
      -number to get permutation under null model
Range of valid answers:  -9999      [-199]  9999

```

CANOCO simply uses the permutations specified in file, despite the adjective “random” in the phrasing of the question. As in Q 54, CANOCO can either permute residuals from the reduced model or from the full model. In the example file there are 3 permutations; so, possible answers are 3 or -3.

In forward selection, the file with permutations should contain sufficient permutations for each variable tested. For example, if 5 tests are carried with 200 permutations in each test, the file should contain at least 1000 permutations.

7.18 Forward selection dialogue

```

Q 66 **** Start of forward selection of variables ****

*** Monte Carlo permutation test ***
    0 = no significance test
    1 = significance test
Range of valid answers:      [0]      1

```

If you asked for forward selection (Q 20 = 1), CANOCO first asks whether you wish to use Monte Carlo permutation tests at any stage during the selection process. If so, CANOCO asks for the type of permutation using the same sequence of questions as in the previous section (Q 52 - Q 64), except for the number of permutations (Q 54). The number of permutations (Q 54) can be specified each time you ask to test a particular environmental variable.

At each step of the selection process, CANOCO gives a list of the environmental variables that are available for selection in order of the extra variance each one would explain if added to the model, followed by question Q 67, described below. For example:

```
Variance explained by the variables selected: .22
"      "      "      all variables      : .63

N      Name Extra fit
1 A1      .03
6 Pasture .03
8 BF      .04
9 HF      .07
5 Haypastu .07
4 Hayfield .07
7 SF      .12
10 NM     .14
3 Manure  .14
```

```
Q 67 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 [3] 10
```

The second and third lines of the question are missing if you did not want Monte Carlo permutation tests during forward selection (Q 66 = 0). By pressing RETURN, CANOCO would add the best variable at this step of the selection process, but you may decide otherwise by typing the number of another variable. If you answer, for example, -3, then CANOCO will determine the statistical significance of variable 3 (Manure) by carrying out a Monte Carlo permutation test. For each test you can specify the number of permutations to be carried out. The progress of the test can be followed on the screen and the test result is reported, for example

```
P-value .100 (variable 3; F-ratio= 1.93; number of permutations= -199)
```

After this, question Q 67 is repeated, so that you are free to include the variable you just tested, to test or include any other variable, or to stop the selection process. See section 6.2.7 for an example run and the meaning of the statistics reported by CANOCO.

7.19 Canoco project files

CANOCO logs the dialog at the screen in the file CANOCO.CON in the working folder. Table 7.4 shows the CANOCO.CON file from the example CCA on the Dune meadow data, the output of which forms the example in the Results chapter (the lines with file names are shortened for ease of display). The CON-project file can be used to automate further analyses of the same kind. If you want to use the CON-project file yourself with the console version of CANOCO, make sure that you rename CANOCO.CON first, for example to MYCCA.CON. The reason of this is that CANOCO creates a new CANOCO.CON in each analysis, overwriting any existing file of that name.

Canoco for Windows uses CON-project files to specify and modify the ordination analysis. Most project files can be used with both Canoco for Windows and with the console version CANOCO.EXE. Projects from Canoco for Windows using manual forward selection cannot be run with the console version. Projects made with the console version that use features that are not in Canoco for Windows (see section 7.1.1) cannot be opened in Canoco for Windows.

You can edit the CON-project file with Notepad or another editor, or word processor. (If you use a word processor, make sure you save the file as a text-only or DOS-text file). For

example, you may wish to rerun the analysis with downweighting of species. For this change, the “0” in “0 = weighting of species” to a “1” (Q 41).

The CON-project file, e.g. MYCCA.CON, can be used in two ways

- Enter the file MYCCA.CON at Q 2
- Run the file from the command-line with the piping symbols “<” and “>”. Examples are:
 1. prompt> CANOCO < MYCCA.CON
 2. prompt> CANOCO < MYCCA.CON >MYCCA.SCR
 3. prompt> CANOCO < MYCCA.CON >NUL

In the first example, the usual screen dialog flashes again across the screen, in the second example, it is logged to the file MYCCA.SCR, whereas in the last example the screen dialog is lost. In all cases the results of the analysis are written to the print file, the solution file and the species-environment table (see Chapter 6).

Entering the file at Q 2 is similar to the first piping example, but allows you to continue the analysis interactively when the CON-project file is incomplete. The standard CON-file produced by the console version of CANOCO is always incomplete, because it does not contain a line to end the analysis (Q 46 = 0: Stop, Table 7.4). This feature allows you to continue the analysis interactively. If you wish to use this feature with a CON-project file produced by Canoco for Windows, you must remove the last line from the file, including the new line character (remove the last two lines, to be sure).

Table 7.4 Example CANOCO.CON file with added question numbers.

```
Question      Answer      = annotation
Q 13:         2 = DO NOT CHANGE THIS LINE
Q 1:          1 = long dialogue?
Q 4:          0 = changing maximum sizes?
Q 8:          dunespe.dta                = file with species data
Q 9:          S                          = file with covariables
Q 10:         duneenv.dta                = file with environmental data
Q 11:         CANOCO.OUT                  = print file
Q 12:         CANOCO.SOL                  = solution file for CanoDraw
Q 13:         5 = analysis number
Q 20:         0 = forward selection?
Q 22:         2 = scaling of sample and species scores?
Q 23:         3 = spec and sample diagnostics
Q 25:         0 = sample number to be omitted
Q 30:         0 = select/delete of environmental variables
Q 33:         0 0 = product of environmental variables
Q 34:         0 = transformation of species data
Q 37:         .01000 = weight for species ( noweight=1)
Q 38:         31 = species given nonstandard weight
Q 38:        -33 = species given nonstandard weight
Q 38:         0 = species given nonstandard weight
Q 37:         1.00000 = weight for species ( noweight=1)
Q 39:         1.00000 = weight for samples ( noweight=1)
Q 41:         0 = weighting of species?
Q 44:         4 = output of correlations?
Q 45:         2 2 2 2 2 2 2 = ordination output
Q 46:         1 = stop, more analyses, other env. data?
Q 51:         3 = Monte Carlo permutation test?
Q 52:         1 = type of permutation
Q 53:        23239 945 = seeds for random numbers
Q 54:        -199 = number of permutations
```

7.20 Nonstandard analyses

A nonstandard analysis may be obtained by typing the number 10 in response to Q 13. The user is warned that the program has not been tested with as much regard for nonstandard analyses compared to standard analyses, and that the nonstandard analyses do not have a secure theoretical basis. There is no user-support for this option. In a nonstandard analysis the user is allowed to combine options in a nonstandard way. For example, the question about nonlinear rescaling (Q 17) is posed normally only when detrending-by-segments is in force, but in a nonstandard analysis this question is posed for all unimodal methods. In this way the user may specify an analysis in which nonlinear rescaling of axes is used in combination with detrending-by-polynomials or without detrending. (In DECORANA the rescaling question was also posed in basic correspondence analysis, but had no effect.) Nonlinear rescaling of axes is also possible in CCA and DCCA, but its use is somewhat illogical: the optimal linear combinations of environmental variables are searched for, but after these combinations have been determined, they are modified by the nonlinear rescaling of axes, so destroying their optimality property.

The only additional question in a nonstandard analysis, which is posed immediately after Q 13 is:

ITY: Allowed values are -100, -4, -3, -2, -1, 0, 1, 2, 3, 4. The sign of ITY discriminates between linear methods (ITY < 0) and weighted averaging methods (ITY > 0). The absolute value of ITY determines the order of the polynomial used for detrending. The usual orthogonalization procedures in PCA and CA are thus in force if ITY = -1 and 1, respectively. The values ITY = 0 and -100 have a special meaning: if ITY = 0, then detrending-by-segments is in force; ITY = -100 corresponds to the option in Q 42 requiring no centring by species.

NEIGZ: Allowed values are 0, 1, 2, 3, 4. Type 0 for an indirect gradient analysis, 4 for a direct gradient analysis, and 1, 2 or 3 for a hybrid analysis (cf. Q 14).

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

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

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

I do not know whether the reciprocal ranking algorithm gives unique species and samples scores, irrespective of the initial scores. To lessen this possible dependency on initial scores, ranking of scores is performed after two iterations of the iterative ordination algorithm starting from the usual initial scores. (Note that one iteration of this algorithm as implemented in CANOCO involves 4 passes of the data.) For technical reasons, the reciprocal ranking algorithm usually does not converge in 15 iterations; nevertheless the final scores are precise enough for most practical purposes. The final iteration is performed without ranking. If IORD or JORD is equal to 1, then it is implied that the sample scores are derived from the species scores (Q 21/Q 22= 1). As a consequence, the final scores satisfy the equations (6.20), (6.21), (6.28) and (6.29) with $\alpha = 1$. However, the species scores are not a simple function of the samples scores; the equations (6.10) and (6.11) do not hold. In linear methods the mean square of the sample scores is set to 1 (cf. (6.21)). In weighted averaging methods the sample and species scores are scaled in SD-units, either by nonlinear rescaling of axes or by using equation (6.15).

Q 68 also allows detrending-by-polynomials to be used in linear methods. This use is, however, not supported by theory and therefore not recommended: linear methods applied to data arising from unimodal models produce a "horseshoe" which scrambles the order of samples along the first axis. In contrast, the "arch" produced by weighted averaging methods does not scramble the order of samples along the first axis.

8. Canoco examples

8.1 Introduction

This chapter gives a number of examples of the use of CANOCO 4.5 and Canoco for Windows. The data files and project files of examples are placed in four subdirectories of the directory C:\CANOCO\SAMPLES where C:\CANOCO is the name of the directory where you installed CANOCO. The four directories are

- UNIMODAL with examples from the “Unimodal Models to Relate Species to Environment” booklet
- PERMUTIO with examples of permutation tests and the decomposition of variance
- METHODS with examples of other methods that are also available in CANOCO
- PROJECTS with default project files for most methods that are available in CANOCO

Each example has a readme.txt file with references to the original authors of the data. We are grateful to these authors for making their data available for users of CANOCO. The copyrights remain with the original authors.

The directory C:\CANOCO\PROJECTS is mainly for users of the console version of CANOCO. It contains default project files for most methods that are available in CANOCO. The names of the different projects are explained in the readme file in the directory.

For maximum compatibility across different computer platforms, the names of files all use the MS-DOS 8.3 convention. This leads to somewhat cryptic names, but we tried to be as informative as possible (Table 8.1). All data files that can be analyzed with Canoco for Windows have the extension DTA. The project files all have the extension CON. If analyzed, the project gives a solution file with the same name and extension SOL. The content of the log-window (the output file) is in the same directory using the project name with the extension LOG.

Each example data set starts with a short description of the problem being addressed, the data source, what is illustrated, followed by a list of files names with a description of their content or purpose. The example data sets are numbered within the first three main groups mentioned above.

Note that the CanoDraw for Windows documentation uses some of the examples discussed in this chapter for illustration of the various aspects of visualizing ordination results (Chapter 14 of this manual).

Table 8.1 Naming convention of files in the examples.

(*ANA* = type of analysis, e.g. CCA, prefixed or postfixed with extra information indicating a serial number, organism or experimental design; ORGANISM: type of organism, e.g. ALGAE).

File type	Typical names	Extension
Project file	*ANA*	con
Species data	species	dta
	ORGANISM	dta
Environmental data	environm	dta
	explanat	dta
	design	dta
Covariable data	as Environment	dta
Solution file	*ANA*	sol
Log file	*ANA*	log

8.1.1 How to analyze the examples

Search for the project file of the example with the Windows Explorer (available either in the *Start / Programs* submenu, or in *Start / Programs / Accessories* submenu, depending on operating system version). Double click the project file to launch Canoco for Windows with the project file, then click, in the Project View, the **Options** button, to view and/or modify the options of the project, or immediately the **Analyze** button. Alternatively, launch Canoco for Windows, click the **Open** button and **Browse** to the directory with the example project. In order not to overwrite existing project files, also click the **Save as** button and enter a new project name, e.g. test.

On all platforms, open a Prompt-Box (DOS-, Console or Command Box), go to the subdirectory where the example file is and invoke either Canoco for Windows (e.g. type C:\CANOCO\CANOWIN) or the console version of CANOCO by typing C:\CANOCO\CANOCO. With the console version you may also try to use the redirection symbols "<" and ">":

```
C:\CANOCO\CANOCO.EXE < project.con > project.scr
```

After the analysis is completed, the file CANOCO.OUT gives what is in the log-window in CANOCO for Windows.

Most project files can be used with both Canoco for Windows and with the console version CANOCO.EXE. It is clearly indicated if a project runs with the console version only. Projects from Canoco for Windows using manual forward selection cannot be run with the console version.

☞ To see in Canoco for Windows which species, samples, environmental variables or covariables are being analyzed in the examples, open an example project, click Options and, in the Setup Wizard sequence, check the Delete boxes of the Data Editing Choices. Canoco for Windows then lists the available species, samples, covariables and environmental variables. You do not need to actually delete any item, nor do you need to uncheck the Delete boxes.

☞ For the best looking automatic plots in CanoDraw, choose in Canoco for Windows "Focus scaling on inter-species correlations" with species scores "divided by their standard deviation" or, in unimodal methods, on "inter-species distances" with "biplot scaling". If you are satisfied with the analysis, rethink about the scaling and, if needed, adapt the scaling in Canoco for Windows and manually optimize the plots in CanoDraw using the *View / Diagram Settings / Properties 1 / Show rescaling coefficients...* option.

☞ To inspect the values in a solution file and to make simple scatter plots, import the solution file into a spreadsheet program as a tab-delimited (Windows) ASCII text file. Scatter plots are not automatically correct ordination diagrams, however; the aspect-ratio must be changed so that axes are isotropic (equal scales).

8.2 Examples from "Unimodal models to relate species to environment" re-analyzed

The examples in this section show how to reproduce some of the analyses that are published in the booklet "Unimodal models to relate species to environment" (Ter Braak, 1996). The pages where the example is discussed are indicated as "Booklet: page numbers". Many examples lend themselves naturally to further analysis and these are done, where appropriate, in subsequent subsections.

Some of the examples include permutation tests to determine the statistical significance of the environmental variables (e.g. section 8.2.3) and/or of the ordination axes of a canonical analysis (sections 8.2.4.2 and 8.2.4.3). Permutation testing is further illustrated in a series of examples, from simple to advanced, in section 8.3.

Table 8.2 List of examples in Unimodal models (n.a. = not available).

Pages	Directory	Description
pp 54-55:	SPIDER1	DCA on Spider data (n = 100 samples, m = 12 species, environmental data for 28 samples) of Van der Aart & Smeenk-Enserink (1975). Indirect environmental gradient analysis and niche study.
pp 63-67:	SPIDER2	CCA on Spider data (n = 28 samples, m = 12 species, q = 6 - 26 environmental variables) of Van der Aart & Smeenk-Enserink (1975). Environmental gradient and niche study.
pp 67-68:	DYKE	CCA on Dyke vegetation (n = 125 samples, m = 133 plant species; q = 6 environmental variables) of De Lange (1972). Environmental gradients study.
pp 68-69:	ALGAE	CCA and DCCA on algae data (n = 25 samples, m = 34 algae taxa, q = 7 environmental variables) of Fricke & Steubing (1984). Pollution gradient.
pp 78-79:	DUNEBOOK	CCA on Dune Meadow data (n = 20 samples, m = 30 species, q = 8 environmental variables) of Batterink & Wijffels (report). Observational effects of management regimes on vegetation.
pp 79:	WEEDS	CCA on Arable weeds (n = 96 samples, m = 13 species, q = 2 environmental variables) of Post (unpublished). Spatial gradient in weed composition.
pp 79:	SEASHORE	CCA on sea-shore data (n = 63 samples, m = 68 species, q = 2 environmental variables) of Cramer & Hytteborn (1987). See also Jongman et al. (1987: pp 167-168). The samples are along four transects. Succession study. Inference of environmental change.
pp 88-89:	n.a.	Partial CCA on diatoms (n = 402 samples, m = 330 species, q = 24 environmental variables, p = 2 covariables) of Smit (1988). Pollution gradient adjusted for sampling and other background variation.
pp 116-117:	n.a.	PCA on diatoms (n = 57 samples in 16 pools, m = 24 species) of Van Dam, Suurmond & Ter Braak (1981). Acidification study. Change in diversity of diatom assemblages with time. Also data exploration.
pp 118-120:	n.a.	DCA on bird species (n = 526 samples, m = 51 species) of Opdam, Kalkhoven & Philippona (1984). Gradients in landscape ecological context.
pp 120:	DUNEBOOK	RDA on Dune Meadow data (n = 20 samples, m = 30 species, q = 3 environmental variables) of Batterink & Wijffels (report). Environmental gradient study.
pp 122-123:	n.a.	Hybrid CCA on tropical forest data (n = 40 samples of which 16 supplementary samples, m = 285 species, q = 3 environmental variables) of Purata (1986). Succession study. Passive samples.
pp 140, 149:	EPIALGAE	CCA on algae data (n = 198 samples, m = 181 species, q = 29 environmental dummy variables) of Snoeijis & Prentice (1989). Effect of temperature increase on diatom assemblages. CCA on two nominal variables coding for 18 sites and 11 months, with 2 supplementary environmental variables.
pp 145-148:	DUNEBOOK	RDA on Dune Meadow data (n = 20 samples, m = 30 species, q = 8 environmental variables) of Batterink & Wijffels (report). Observational effects of management regimes on vegetation.
pp 156-175:	STREAMS	Partial CCA on macro-invertebrate data (n = 40 samples, m = 197 species, q < 21 environmental variables, p = 5 covariables) of Higler & Repko (1981). Land-use effects on macro-invertebrates, adjusted for sampling season. Forward selection of variables.
pp 189-198:	VEGCHANG	RDA on vegetation data (n = 40 samples, m = 106 species, q = 2 environmental variables) of Ter Braak & Wiertz (1994). What causes the vegetation change? Sample on a 4 x 5 grid, sampled twice. Environmental inference from species data (change model). Monte Carlo permutation tests.
pp 201-216:	n.a.	partial RDA of a designed BACI experiment (n = 84 samples, m = 17 species, q = 1 environmental variable, p = 19 covariables) of Verdonschot & Ter Braak (1994). Before-After-Control-Impact design in 12 ditches sampled 3 times Before and 4 times After the Impact (an insecticide applied in 4 different doses).
pp 201-216:	n.a.	RDA of a completely randomized experiment (n = ??)
pp 229-234:	n.a.	Canonical Correlation analysis on political data. Correlation and Regression biplots.

pp 239-256: DISEASES

RDA and partial RDA on regional mortality ratios of 11 diseases ($n = 39$ samples, $m = 11$ species, $p > 15$ environmental variables) of Kunst et al. (1990). Effect of socio-economic status on causes of death. Regression biplots.

8.2.1 Example SPIDER1 - A niche study by CA and DCA

Problem:	Determination of the niches of hunting spiders in a dune area
Data:	Van der Aart & Smeek-Enserink (1975)
Booklet:	pp 54-55
Directory:	\CANOCO\SAMPLES\UNIMODAL\SPIDER1
Illustration of:	<ul style="list-style-type: none"> • DCA interpretation. • How to plot a species response curve against an ordination axis using CanoDraw. • How to interpret an indirect gradient analysis with CANOCO if there are many samples without environmental data. • How CANOCO links samples in different files. • How to relate the ordination axes of different analyses.

Files	Name	Description
Species	spide100.dta	counts of 12 hunting spiders in 100 pitfalls in a dune area
Environmental	soilveg.dta	data for 26 variables on soil and vegetation around 28 of the pitfalls
	soilvgl.n.dta	ln-transformed soil and vegetation variables for 28 pitfalls
Derived	dca100.dta	sample scores on the DCA axes, derived in project spidedca.con from the spider counts
Project	spid_ca.con	CA of spider counts (100 active samples)
	spid_dca.con	DCA of spider counts (100 active samples)
	indirect.con	regression of the first axis of the DCA on the soil and vegetation variables (28 active samples)
	indiforw.con	as indirect.con but using forward selection
	dca28.con	DCA of spider counts with environmental data, yielding a DCA of 28 pitfalls only (28 active samples)
	dca28100.con	DCA of spider counts with environmental data, relating the two DCA analyses for the 28 pitfalls
	spid_cca.con	CCA of spider counts on 6 soil and vegetation variables (28 active pitfalls)

8.2.1.1 SPIDER1: CA with arch effect and DCA with species response curves

In Unimodal Models; page 54, the spider counts are first analyzed by correspondence analysis (CA). The analysis can be carried out with the project file spid_ca.con. After opening this project file and clicking **Analyze...**, the eigenvalues of the analysis appear in the log-window of the project. The first two eigenvalues are .65 and .42. The ordination diagram with the sample scores can be obtained by invoking CanoDraw. For this, switch back to the Project View using the **F3** shortcut key or the Switch button on the toolbar and then click CanoDraw. When CanoDraw has started, save the new CanoDraw project under suggested name and select the *Create / Scatter Plots / Samples* menu command. The diagram shows an approximate quadratic relation between the sample scores of the second, vertical axis and those on the first, horizontal axis (arch effect). This arch is removed in detrended correspondence analysis (DCA). Switch back to your CANOCO project. To obtain the DCA, either click Options to change the

project options from CA to DCA (with option “detrrending by segments”), or open the existing project file `spid_dca.con`. In the former case, save the modified project using “File | Save as” so as not to overwrite the original project file `spid_ca.con`. After clicking **Analyze...** the eigenvalues and lengths of gradient of the DCA appear in the log-window. The first eigenvalue is always the same as that of CA, but the second eigenvalue decreases from .42 to .086, suggesting that the second axis is unimportant. The length of gradient of the first axis is 4.4 SD, suggesting that some spiders species show a unimodal response to the first DCA axis. This is indeed the case as illustrated in Figure 2 on page 56 of *Unimodal Models*. The fitted curves shown in Fig. 2 can be obtained with CanoDraw. Invoke CanoDraw from the project-view. In CanoDraw (1) save the new CanoDraw project, (2) select *Create / Attribute Plots / Species Response Curves*, and there select the *Generalized linear model* option, select all species except *Pardlugu*, click *Axis 1* and *OK*, (3) in the *GLM Options* dialog select *Quadratic* and *Poisson* and then confirm with the *OK* button all the reports about fitted response models, before the diagram appears. Consult the section 14.2 of this manual for more detailed instructions.

8.2.1.2 SPIDER1: Environmental interpretation of first DCA-axis by multiple regression in CANOCO

On page 55 of *Unimodal Models*, the first DCA-axis of the spider data is interpreted in terms of environmental variables. For this, one typically enters these variables as Environmental data to CANOCO. In this data set, environmental data are available for only 28 of the 100 pitfalls. This presents a problem: if these data are entered as Environmental data to CANOCO and if we continue to ask for an indirect gradient analysis by DCA with detrrending by segments (project file `dca28.con`), the resulting analysis is not that of the section 8.2.1.1; for example, the first eigenvalue changes from .65 to .70. The reason for the difference can be found by inspecting the log-window more carefully. After the report on reading the environmental data, CANOCO reports:

```
No. of active samples:      28
No. of passive samples:    72
No. of active species:     12
```

The DCA is carried out on the spider data from 28 samples only. These are the only samples for which there are environmental data. The remaining 72 samples are treated as passive, i.e. as supplementary samples that do not influence the DCA. This is clearly not the analysis performed on page 55 of *Unimodal Models*. There, the DCA is performed with 100 active samples. From the resulting DCA-scores, the first axis scores are selected for the 28 samples for which there are environmental data. These scores are then taken as the response variable which is regressed on three of the environmental variables. To obtain the correct analysis with CANOCO, you have to carry out the following steps which require the utility `WCanoImp` and a Windows spreadsheet application. The result of these steps is a new file which contains the scores of the 100 samples on the first four DCA axes and which can be read by CANOCO. In the Windows environment the steps are:

- Open the solution file of the original DCA (`spid_dca.sol`) with an editor, word processor or spreadsheet. The solution file is, by default, a tab-delimited ASCII text-file.
- Search for the text “Samp: Sample scores”.
- Copy the body of 100 scores to a new spreadsheet, together with the row of headings for the columns (N, Name, AX1, ..., AX4, Weight, N2).
- Delete the first column (headed N).

- Delete the three rows between the column heading and the scores of the first sample.
- Copy the columns “Name ... N2” and the row with the heading and the 100 rows with the sample scores to the Clipboard.
- Invoke WCanoImp (Start / Programs / Canoco for Windows / Canoimp menu item)
- Save the Clipboard content with WCanoImp.

The result is a file such as dca100.dta with the DCA-scores of all 100 samples. The subsequent interpretation of these scores can be carried out with the project indirect.con. In this project, the file dca100.dta with DCA-scores of 100 samples is opened as Species data, and the file with environmental variables as Environmental data.

The environmental data consist of 26 variables that characterize the soil and vegetation around 28 pitfalls. Many of the variables have a very skew distribution and are therefore log-transformed as in the original paper by Van der Aart & Smeek-Enserink (1975). Because CANOCO cannot transform data entered as Environmental data, the transformation must be carried out outside CANOCO, e.g. in a spreadsheet. The file with log-transformed environmental data is soilvln.dta.

To regress the first DCA axis on three of the environmental variables (project file indirect.con), select direct gradient analysis, enter the file dca100.dta as the Species data file and the file soilvln.dta as the Environmental data file, select RDA as the analysis, and in the wizard page on Data Editing Choices, click the boxes corresponding to Delete Species and Delete Environmental Variables. Recall that the DCA-axes are “Species”. Delete all species, except AX1 and delete all environmental variables except WaterCon, BareSand, CoveMoss. The latter is done most easily by first moving all variables to the right-hand list box and then by moving the three required variables to the left again. After analyzing the project so defined, the summary says that 90.8 percent of the variance in the species data (i.e. the first DCA-axis) is explained by the environmental data. This corresponds to the 90% mentioned in Unimodal Models. In the solution file of this analysis you can find the t-ratios of the regression coefficients under “tVal: t-values of the regression coefficients”. These t-values are the usual ratios of the estimate and the standard error of the estimate as given by any computer program for multiple linear regression analysis. See section 8.4.4 for more information on the use of CANOCO for linear regression.

To understand how CANOCO links the DCA scores of the 100 samples to the environmental data of the 20 samples, you need to inspect the file with environmental data. The sample numbers in the environmental file soilvln.dta are 2, 8, 9, 11, ... , 89. These are the pitfalls for which there are environmental data. Check also the sample names at the bottom of the file. In the solution file indirect.sol, these 28 samples are listed first, followed by the remaining samples.

You may wish to experiment with other choices of environmental variables to explain the first DCA axis, for example by using forward selection. For this, click Options and go to the wizard pages where you can ask for forward selection. After ending the wizard sequence it is prudent to save the modified project in a new file (click Save as on the tool bar). An example of automatic forward selection is in the project file indiforw.con.

In this example we have obtained two DCA analyses, one using 100 active samples and one using 28 active samples. It may also be of interest to know how similar or dissimilar the axes are from these two ordinations. One way to investigate this is to correlate the axes pairwise based on the 28 samples in common. This can be achieved by specifying the file dca100.dta as supplementary environmental data as in the project file dca28100.con. The correlation between the first axes scores of the two analyses is .9963. This number is given in the second correlation matrix in the log-window in the entry SPEC AX 1 and AX1. (The first correlation matrix applies to the analysis of species with respect to the environmental data, the second to the

supplementary environmental data). To find the entry .9963 in the log-window, open the project dca28100.con, click Analyze... , switch to the log-window, move to the beginning of the window, position the cursor by clicking the mouse pointer (so that the insertion point is placed there), select Find from the Search menu, enter .9963 and click the Find next button. It is of interest to note that instead of the file dca100.dta, we could equally have specified the original solution file spid_dca.sol as supplementary environmental file. CANOCO automatically picks the first block headed "Sample scores" from this file. In conclusion, the first DCA axis of the 28 pitfalls is very similar to the first DCA axis of all 100 pitfalls.

Finally, the spider counts are related to 6 selected soil and vegetation variables by canonical correspondence analysis (CCA) in the project file spid_cca.con. The results are the same as those in the next example (Example 8.2.2), except that the 72 pitfalls without environmental data are added as supplementary samples in the solution file. In this analysis, the spider counts are transformed by taking square-roots.

Remark: The theory in the paper from which the DCA-example is drawn (Ter Braak 1985) shows that CA and DCA give approximate solutions to quadratic latent variable models for Poissonian count data. From this perspective it is not needed to transform the counts by taking square-roots or logarithms. However, from the data-analytic point of view, the counts are so over-dispersed and the total counts of the species are so unequal that it seems wise to take their square-root. This is done in the CCA-example in Ter Braak (1986).

☛ To relate the ordination axes of an analysis with those of an earlier analysis, specify the solution file of the earlier one as Supplementary Environmental Data file.

8.2.2 Example SPIDER2 - A niche study by CCA

Problem:	Determination of the niches of hunting spiders in a dune area by CCA
Data:	Van der Aart & Smeek-Enserink (1975)
Booklet:	pp 63-67
Directory:	\CANOCO\SAMPLES\UNIMODAL\SPIDER2
Illustration of:	<ul style="list-style-type: none"> • CCA interpretation. • How to interpret numbers in the file SPEC_ENV.TAB. • How to define products of explanatory variables using the Define Interaction Terms option. • How to choose names for environmental variables when products are to be defined. • Potential problems when there are many environmental variables and how to cope with them.

Files	Name	Description
Species	spid_spe.dta	counts of 12 hunting spiders in 28 pitfalls in a dune area
Environmental	spid_env.dta	6 soil and vegetation variables for 28 pitfalls (ln-transformed)
	spiden26.dta	26 soil and vegetation variables for 28 pitfalls (ln-transformed)
Derived	wa_6.dta	Weighted averages of spiders with respect to the 6 environmental variables (copy of spec_env.tab from spider.con)
	wa_all.dta	Weighted averages of spiders with respect to all 26 environmental variables (copy of spec_env.tab from ca_all.con)
Project	spider.con	CCA of square-root transformed spider counts on 6 environmental variables (ln-transformed)
	cca_all.con	CCA of square-root transformed spider counts on all 26 environmental variables (ln-transformed)
	ca_all.con	CA of square-root transformed spider counts, interpreted using all 26 environmental variables (ln-transformed)
	cca_all2.con	CCA of square-root transformed spider counts on 28 samples and 27 environmental variables
	forward.con	CCA with forward selection using all 26 environmental variables

8.2.2.1 SPIDER2: CCA of spider counts on six environmental variables

In Unimodal Models on page 66, the spider counts of the 28 pitfalls with environmental data are analyzed with CCA. The data are shown in a coded form in Table 3 on page 66; the data files of this example contain the original data on which the analyses were performed. Largely on *a priori* grounds the number of environmental variables was reduced from the original 26 to six. The reported CCA can be generated with the project file spider.con. The eigenvalues and species-environmental correlations of this CCA, reported in Table 1 on page 65 of Unimodal Models, can be found in the ordination summary at the end of the log-window. The intraset correlations in Table 2 on page 65 of Unimodal Models are also in the log-window. They are part of the weighted correlation matrix. Search in this matrix for the rows starting with WaterCon. The reported numbers are in the fifth and sixth column and are the correlations with the environmental axes labeled ENVI AX1 and ENVI AX2. The numbers in the first and second column of the matrix are the corresponding interset correlations. These correlations are also given in the solution file under the heading "CorE:" All the other output of the CCA listed in Unimodal Models depends on the chosen scaling of the ordination scores, which is Hill's scaling

with focus on inter-sample distances. This scaling is discussed in Table 6.35 and on pages 171-172 of Unimodal Models.

The standardized canonical coefficients (Table 2 on page 65 of Unimodal Models) can be found in the solution file spider.sol in the table headed "Regr: Regression/canonical coefficients for standardized variables". These coefficients define the linear combinations of environmental variables and the resulting sample scores are listed in the last block of scores in the solution file. The ordination diagram of Fig. 1 on page 63 consists of the blocks of scores from the solution file headed:

- Spec: Species scores (Hill's scaling).
- Samp: Sample scores (Hill's scaling).
- BipE: Biplot scores of environmental variables.

Following the later advice on page 143 and 171 of Unimodal Models, CanoDraw uses by default in CCA and RDA the second set of sample scores in the solution file which is headed

- SamE: Sample scores which are linear combinations of environmental variables.

The first set of sample scores better represents the community structure and the second set better represents the community response to the environmental variables in the analysis (McCune 1997; see also Table 6.35 and Table 6.36). Because the species-environmental correlation is .96, there is not much difference here between the two sets of scores. *If you wish to use the first set of sample scores in CanoDraw, click Project / Settings and check the Plot SAMP scores even for constrained axes option.*

Note that the biplot scores of environmental variables are, for each axis, proportional to both the intra- and the inter-set correlations. The scaling of the environmental biplot scores is optimized for the interpretation of the weighted averages of species with respect to environmental variables by the rule indicated in Fig. 1 on page 63 of Unimodal Models. From Fig.1 the approximate order of the weighted averages can be inferred. The exact order of the weighted averages can be obtained from the additional output file of CANOCO, called spec_env.tab. This text file has been copied to wa_6.dta. The first 12 values are

-1.1413	.2930	-1.7284	.6562	-2.2010	.3471	.2920	-.3403
.4253	.3871	.2979	.4864				

These are the weighted averages of the 12 hunting spiders with respect to Water Content, expressed as deviations from the mean Water Content (2.6694, given in the log-window). The values are standardized by division by the standard deviation of Water Content (.6842). The values are thus weighted averages with respect to the standardized environmental variable. For example, the lowest value is -2.2010. It is the 5th value and therefore applies to the 5th species Arct peri (see spider.sol or the order of names in wa_6.dta). The highest value is .4864 for the last species (Zora spin), but there are several species with similar values. The order of the weighted mean Water Content of species is well reflected in the ordination diagram of Fig. 1. Also, the species with positive values lie to the left of the origin and the species with negative values to the right. The file spec_env.tab can be inspected by using for example the Notepad program. Table 8.3 shows the same information in transposed format. Using the biplot rule, the spider points and environmental arrows jointly explain 88.5 % of the variance in this table, as given in the summary of the ordination in the log-window: 88.5 is the second entry in the row named "Cumulative percentage variance of the species-environment relation". Unimodal Models (p. 66) reports 87% rather than 88%.

Table 8.3 Weighted averages of hunting spiders with respect to the six standardized environmental variables.

(1 = WaterCon, 2 = BareSand, 3 = FallTwig, 4 = CoveMoss, 5 = CoveHerb, 6 = Refl Lux).

	Environmental variable number					
	1	2	3	4	5	6
Alop acce	-1.14	.70	-.62	.94	-.20	.88
Alop cune	.29	-.27	.19	-.20	.06	-.08
Alop fabr	-1.73	1.59	-.62	1.07	-.75	.96
Arct lute	.66	-.10	-.13	-.39	.50	-.18
Arct peri	-2.20	1.98	-.65	1.09	-1.26	1.21
Aulo albi	.35	-.23	-.05	-.29	.51	-.07
Pard lugu	.29	-.42	1.30	-.61	-.92	-1.11
Pard mont	-.34	.09	-.41	.68	.08	.51
Pard nigr	.43	-.21	-.14	-.42	.42	-.10
Pard pull	.39	-.36	-.26	-.18	.54	.04
Troc terr	.30	-.23	.35	-.30	-.12	-.36
Zora spin	.49	-.14	.36	-.41	.00	-.55

Table 3 on page 66 of *Unimodal Models* can be prepared by importing the CANOCO solution file into Microsoft Excel® and using the values on the first axis of the Species Scores (Spec:) to arrange the species in species data (spid_spe.dta) and the Environment-derived scores (SamE:) to arrange order of samples both in species and environmental data. Note that the datasets (spid_spe.dta and spid_env.dta) can be imported into Excel by first transforming them into TAB-separated format with the CanoMerge program (see section 4.2).

After having inspected the CCA-results, you may wish to modify the project file spider.con to generate the eigenvalues and correlations of DCA and DCCA in Table 4 on page 67 of *Unimodal Models*. To obtain the DCA, you need to select indirect gradient analysis in the first wizard page on Available Data. To obtain the DCCA, select direct gradient analysis and detrending by segments.

8.2.2.2 SPIDER2: CCA of spider counts with many environmental variables

The authors of the spider data (Van der Aart & Enserink, 1975) measured a total of 26 environmental variables, nearly as many as there are samples. All these variables are in the file spiden26.dta. A CCA using all 26 variables (project file cca_all.con) is theoretically a constrained and direct analysis, but because of the number of environmental variables, there are, in practice, few constraints. This CCA generates a solution that is very close to that of an unconstrained analysis by CA of these data (project file ca_all.con), as is apparent by comparing the eigenvalues of the two analyses. The solutions can be made precisely equivalent by adding yet another extra environmental variable by asking for an interaction term. In the project file cca_all2.con, a product between two existing variables is added to the environmental variables. The eigenvalues of this analysis are equal to those of the CA (at least, in theory; in practice there may be small differences because of numerical problems). Indeed, CANOCO reports in the log-window that it has numerical problems in deriving the CCA by saying

*** BEWARE ***

Residual for axis 1 bigger than tolerance, which is .000001

The maximum of the variance inflation factors of the variables is over 50000. Therefore, the canonical coefficients are intrinsically unstable. Variance inflation factors should normally be

less than 20, say, to warrant usage of the canonical coefficients. The lesson is to use either far fewer variables or to stick to the unconstrained, indirect analysis, as the DCA in Example 8.2.1.

One statistical way to solve the problem of the number of environmental variables is to use forward selection. The project file `forward.con` gives an example.

Another way is to represent the environmental data by their first few principal components. For this, (1) perform a PCA with the environmental data in the file `spider26.dta` as Species Data file and (2) specify the solution file of this PCA as Environmental Data file in a CCA in which the spider counts are the Species Data.

The project `cca_all2.con` also serves to illustrate a tip on the naming of environmental variables when product variables are to be defined. The product of the two variables `CoveCory` and `CoveHerb` (variable number 13 and number 7) is given the name `Cove*Cov`. The square of `CoveHerb` would be given the same name. To avoid these ambiguities in name giving, make sure that the start of the name is as informative as possible. Better names in this example are `CoryCove` and `HerbCove`.

8.2.3 Example DYKE - CCA of presence / absence data

Problem:	How plant species in fresh-water dykes are related to water chemistry and soil type.
Data:	De Lange (1972)
Booklet:	pp 76-68
Directory:	\CANOCO\SAMPLES\UNIMODAL\DYKE
Illustration of:	<ul style="list-style-type: none"> • How to code and display a nominal variable (soil type). • How to select species for plotting. • How to inspect axes 3 and 4. • How to test the statistical significance of the relation between species and environment. • How to evaluate the importance of (sets of) environmental variables.

Files	Name	Description
Species	dyke_spe.dta	occurrence (1/0) of 133 water plant species in 125 fresh water dykes
Environmental	dyke_env.dta	3 water chemistry variables and 3 soil type classes
Derived	chlorsor.txt	Weighted Averages of species sorted on chloride
Project	dyke_dca.con	DCA of plant species interpreted by environmental variables
	dyke_cca.con	CCA of plant species on all environmental variables
	cca_dca.con	project relating the CCA and DCA axes
	testall.con	significance test of species-environment relation
	water.con	CCA on water chemistry variables only
	soil.con	CCA on soil type only
	forward.con	forward selection

8.2.3.1 DYKE: Occurrence of water plant species explained by CCA

As mentioned in Unimodal Models on page 67, the DCA axes extracted from the water plant occurrences are poorly related to the available environmental data. The first species-environment correlation (in the summary in the log-window) is .524. When the axes are extracted by CCA, this species environment-correlation increases to 0.824. At the same time the first eigenvalue drops from 0.344 in DCA to 0.197 in CCA. As expected, CCA explains less of the species data but more of the species-environment relation. The latter is best expressed by the percentages variance of the weighted averages of the species with respect to the environmental variables (file spec_env.tab produced by CANOCO; cf Table 8.3) that is explained in two dimensions: 55.3 % in CCA against 19.2% in DCA (see the row "species-environment relation" in the log-window). The first (species-derived) axes of CCA and DCA have a correlation of -0.71, as can be obtained by running the project file cca_dca.con. This project differs from the CCA-project file dyke_cca.con only in that the solution file of the DCA (dyke_dca.sol) is specified as a Supplementary Environment file. The final species-environment correlations in the log-window of this analysis range from 0.915 for the first CCA-axis to 0.391 for the fourth axis. These are the multiple correlations of the regressions of each species-derived CCA-axis onto the four DCA axes.

The entries in Table 5 on page 67 of Unimodal models can be found in the log-window and in the solution file of the CCA (dyke_cca.sol), except for the canonical coefficients for Peat and Sand. The footnote to the table says that these two 0/1 variables were not standardized to unit variance. The table thus gives the coefficient for the original 0/1 variables, which implies that the canonical coefficients in the solution file must be divided by their standard deviations (in the log-window: .4026 and .2626, respectively). For example for the 2nd axis: $.1755/.4026 = .44$ and $-.0792/.2626 = -0.30$ (cf Equation (6.30)). Note that the canonical coefficients for Clay are all zero. CANOCO takes Clay (variable 6) as the reference of the soil type classes because it is the last of the three soil type variables in the data file. This is also visible in the log-window from the message:

```
***** Collinearity detected when fitting variable      6 *****
```

and from the fact that the variance inflation factor of this variable is zero (last column in the table of means and standard deviations; Search for Clay in the log-window). You may wish to check that the eigenvalues and species and sample scores of the CCA do not change when another soil type is taken as the reference class. For this, modify the project (click Options.), check the Delete box in the Data Editing choices after environmental variables, explicitly delete, for example, Sand and analyze the modified project. You may also wish to open the environmental data file dyke_env.dta with Notepad to see how the soil type is coded in the data. Soil type is represented by the last three columns in the file that consist of zeroes and ones only with one "1" per triplet. The columns indicate whether the dyke is on Peat, Sand or Clay, respectively; for example the first sample is on Clay.

In the ordination diagram on page 68, the soil types are displayed by arrows based on the environmental biplot scores (BipE: in the solution file). As discussed on page 65 and later on page 170, it is more natural to display the soil types as points. To achieve this in CanoDraw, select Files, Select nominal env. vars, and select Peat, Sand and Clay. Qualitatively the ordination diagram changes little.

There are 133 species in the data file whereas the published ordination diagram (Fig. 2 on page 68) displays, for clarity, only 31 species. It is difficult to present general rules about which species to select for display. For Fig. 2 all species were plotted except those that would appear close to the center. Because the data set also contains species that occur only once or twice, such rare species are prominently displayed if they happen to occur at relatively extreme environmental positions. Examples in Fig. 2 are the species Pota *dec and Call hamu, both at the bottom of the diagram, which occur only once. From the diagram we know that the two dykes in which these species were found, happen to have very low chloride and relatively low EC, as can be seen by projecting their position on the extended arrows for Chloride and EC. This can be verified in the data by inspecting the file spec_env.tab or the sorted file Chlorsor.txt with the Notepad program. The data alone are, however, insufficient to conclude reliably that this is the true niche of these species! A better, data-driven selection of species, which is available in CanoDraw, is to plot only those species that occur, for example, 10 or more times and whose fit to the diagram is 5% or more. This results in 23 selected species (see section 14.3).

Finally, it should be noted that the second and third CCA eigenvalues are quite close (both ca. 0.12). In addition to the ordination diagram on page 68, one should therefore also inspect the third ordination axis (see section 3.13). Ordination diagrams of the second and third axes can be made with CanoDraw by selecting *Project, Settings, Contents*, the *Second and third axis* value in the *Axes to plot*, and then creating the diagram from the *Create* menu.. The first two axes, shown in Fig. 2 in the paper, are largely defined by EC, Phosphate and Chloride Ratio whereas the 3rd and 4th axis are defined by soil type (Table 5). The water chemistry thus appears to explain more of the species occurrences than soil type.

8.2.3.2 DYKE: Water chemistry and soil type as determinants for plant distribution

The summary of the CCA ordination (in the log-window of the project dyke_cca) also shows that individual plant species occurrences cannot be explained well by the environmental variables. All four axes together just explain 6.3 % of the total inertia (8.257). These summary figures have little meaning for presence/absence data and for abundance data with many zeroes. For one thing, the relation between the species and the environmental variables is statistically highly significant ($P < 0.01$). The test can be obtained by changing the CCA project by selecting a Monte Carlo permutation test. In the project testall.con, both available test statistics are chosen and, further, all defaults are followed as there is no clear reason to deviate from them. The resulting P-values are both 0.005 and thus the above phrase “ $P < 0.01$ ”.

The suggestion from section 8.2.3.1 that water chemistry explains more of the species occurrences than soil type, can be underpinned by running separate analyses on the water chemistry variables and on the soil variables. The analyses can be obtained by modifying the project dyke_cca by checking the Delete Environmental Variables box and by deleting either the soil variables or the water chemistry variables. The resulting sums of the canonical eigenvalues are 0.393 and 0.222, respectively.

Another way to underpin the suggestion is to use forward selection. A forward selection, starting from all variables, selects Phosphate first, followed by Peat, Chloride Ratio, EC and sand (Table 8.4). This table has been copied from the Forward Selection Summary dialog box in Canoco for Windows. To obtain Table 8.4, open the project dyke_cca.con, click Save as.. and type a file name, e.g. my-forward. Then click Options and select forward selection with permutation tests. After clicking **Analyze...**, click **FS Summary** and **Copy**, switch to your word processor and paste the Clipboard content.

Table 8.4 Forward selection of water chemistry and soil type variables to determine their importance in explaining the occurrence of water plants in fresh-water dykes.

Variable	Conditional Effects	Var.Num.	Lambda-A	P-value	F-value
Phosphat		2	0.17	0.005	2.57
Peat		4	0.12	0.005	1.87
Chloride		3	0.12	0.005	1.85
EC		1	0.09	0.010	1.41
Sand		5	0.09	0.070	1.30

8.2.4 Example ALGAE - A study of a pollution gradient

Problem:	Effect of pollution on algae distribution in rivulets
Data:	Fricke & Steubing (1984)
Booklet:	pp 68-69
Directory:	\CANOCO\SAMPLES\UNIMODAL\ALGAE
Illustration of:	<ul style="list-style-type: none"> • How to obtain a DCCA. • How to interpret canonical coefficients. • How to test the significance of ordination axes. • How, given the first axis, a second ordination axis is extracted and tested. • How to define powers of explanatory variables using the Define Interaction Terms option. • Detrending-by-polynomials algorithm in DCCA.

Files	Name	Description
Species	algae.dta	34 algae species (scale 0-5) in 25 sites within rivulets
Environmental	pollutio.dta	7 pollution variables (of which 6 log-transformed)
Derived	alg_cca.xxx	solution file of alg_cca.con with text Sample scores modified to Xample scores
	alg_dcca.xxx	solution file of alg_dcca.con with text Sample scores modified to Xample scores
	ax1pol2.dta	AX1 and AX1-squared of SamE: scores of solution file from pol_dcc2.con, made for the test of the second axis of a DCCA with detrending by second order polynomials
	ax1pol2z.xxx	made by the recipe in section 8.2.4.3 to test for the second axis of a DCCA with detrending by second order polynomials
	weight.txt	default sample weights of (DC)CA used to modify the unweighted RDA project rda_ini.con (See section 8.2.4.3)
Project	alg_dca.con	DCA of algae, interpreted as pollution gradient
	alg_cca.con	CCA of algae on pollution
	alg_dcca.con	DCCA of algae on pollution (detrending by segments)
	pol_dcc2.con	DCCA of algae on pollution (detrending by second order polynomials)
	ax1_cca.con	significance test of the first CCA-axis
	ax2_cca0.con	wrong test of second CCA-axis
	ax2_cca1.con	correct test of second CCA-axis
	ax2_comm.con	significance test of second CCA-axis in the console version of CANOCO
	ax2_dcca.con	approximate test of second DCCA-axis (detrending by segments)
	rda_ini.con	RDA of ax1pol2.dta on pollutio.dta
	rda_samw.con	weighted version of rda_ini.con (sample weights from weight.txt)
	ax2pol2z.con	correct test of second DCCA-axis (detrending by second order polynomials). See recipe in section 8.2.4.3

8.2.4.1 ALGAE: effect of pollution on algae distribution

The first axis of DCA, CCA and DCCA nearly coincide, being a clear pollution gradient (projects `alg_dca.con`, `alg_cca.con` and `alg_dcca.con`). The example is used in *Unimodal Models* (page 69) to show the possible virtue of detrending in CCA, i.e. of DCCA. The reason for this is that the ordination diagram of CCA (Fig. 3) shows the arch effect. Detrending removes the arch (Fig. 4). The second DCCA-axis is of minor importance, as its eigenvalue (0.076) is ten times smaller than that of the first axis. The axis is interpreted on page 69 as being related to the ratio of ammonium to phosphate, the idea being that these variables have on the second axis about equal canonical coefficients of opposite sign (-0.60 and 0.50 in Table 8 on page 70). However, the canonical coefficients are for standardized variables. In terms of log-concentrations, the canonical coefficients must be divided by the standard deviations (2.1153 and 1.2459), yielding -0.28 and 0.40, respectively. The second axis is still a contrast between ammonium and phosphate concentrations, but it is further away from being a log-ratio than the numbers in Table 8 suggest. The interpretation is also open to criticism because the second axis is non-significant as is shown in the next two sections.

8.2.4.2 ALGAE: testing the significance of CCA axes

The significance of the first ordination axis of the CCA can be determined by selecting this option in the wizard page Global Permutation Test (`ax1_cca.con`). The first axis is significant ($P < 0.01$).

To determine the significance of the second ordination axis, we can use the same test after we have modified the project in such a way that the second ordination axis becomes the first axis of the modified project. This can be done by specifying the first CCA axis as a covariable in the new analysis. Let's try and check whether we succeeded. As a first attempt, we specify the solution file of the original CCA as the covariable data file, check the Delete option for covariables and delete all but the variable AX1. Do not ask for a test yet (`ax2_cca0.con`). The first eigenvalue is 0.171 now, which is different from 0.136 (the second eigenvalue in the original analysis). The problem is that CANOCO used the species-derived sample scores from `alg_cca.sol`, whereas it should use the environment-derived scores.

The correct procedure to test for the second ordination axis of CCA or RDA in Canoco for Windows is to

- copy the solution file to a new file, say `SOLUTION.XXX`.
- search for the first occurrence of text "Sample scores" in the file and modify this text for example to "Xample scores".
- specify the modified file to be a covariable file of which you retain the variable AX1.

CANOCO now searches in the file for the first occurrence of the text "Sample scores", which, as you may wish to check, is now the heading of the environment-derived scores ("SampE:").

In the example the modified solution file is called `alg_cca.xxx` and the modified project is `ax2_cca1.con`. After invoking this project, the first eigenvalue is .136 as required and also the 2nd and 3rd eigenvalues are the 3rd and 4th eigenvalues of the original analysis. You may also wish to check that the species scores, sample scores and correlations of the two analyses are similarly related. The only difference is in the canonical coefficients and scores that are derived from them, but this difference is immaterial for our purpose of testing the second ordination axis. If we now select the option to determine the significance of the first axis, the resulting P-value is ca. 0.80, indicating that the second CCA axis of the original analysis is not statistically significant.

To determine the significance of the 3rd axis, simply retain AX1 and AX2 as covariables.

The same procedure can be followed to test the axes of an RDA.

In the console version of CANOCO, the testing can be carried out as described above, but also more directly by asking for “More ordination axes”, and asking for one more axis. See also Q50 in Chapter 7. CANOCO then copies the eigenvector sample scores of the first axis to the covariables so that the first new axis is the second in the new analysis (ax2_comm.con). You can run this analysis from the command prompt in the directory

`\CANOCO\SAMPLES\UNIMODAL\ALGAE` by typing the command

```
c:\CANOCO\CANOCO <ax2_comm.con >nul
```

8.2.4.3 ALGAE: testing the significance of DCCA axes

Testing the first axis of a DCCA presents no difficulty. Testing the second axis is, however, more difficult and, strictly speaking, impossible when detrending by segments is used. The reason for this is that the second axis is detrended with respect to the first axis, but, after the first axis is moved to the covariables as in the previous section, the new first axis is not detrended with respect to this covariable (AX1). To obtain an approximate test, you can define powers of AX1, so mimicking detrending by segments by detrending by polynomials during the test. This can be achieved by entering the modified solution file of the DCCA as covariable data (as SOLUTION.XXX in the previous section), checking the Delete and Define Interactions boxes for Covariables in the project. You can then delete AX2, AX3 and AX4 of the covariable file and then define, for 4th order polynomials, the three extra product variables AX1*AX1, AX1*V5 (=AX1³) and AX1*V6 (=AX1⁴), as V5 = AX1², and V6 = AX1³. An example is the project ax2_dcca.con in which the second axis of a DCCA with detrending by segment is tested using this procedure.

With some extra handwork an exact test can be obtained, if detrending by polynomials is used. The procedure also provides insight into the details of the detrending-by-polynomials algorithm. We illustrate the procedure using detrending by second order polynomials. The initial analysis is done by project pol_dcc2.con. The recipe of the test consists of 14 steps. First the recipe is given and then some explanatory remarks are provided.

1. Open the solution file of the original DCCA (pol_dcc2.sol) with an editor, word processor or spreadsheet. The solution file is, by default, a tab-delimited ASCII text-file.
2. Search for the text “Same: Sample scores” and copy the body of sample scores below this text to a new spreadsheet, together with the row of headings for the columns (N, Name, AX1, ..., AX4, Weight, N2).
3. Delete all column except those headed Name and AX1, and delete the three rows between the column heading and the scores of the first sample.
4. Define an extra column as AX1*AX1 (and further powers if the original DCCA used a higher order polynomial) and copy all columns to the Clipboard.
5. Invoke WCanImp and save the Clipboard content to a file, ax1pol2.dta, say.
6. Run an initial RDA using ax1pol2.dta as Species Data file and the original environmental data, here pollutio.dta, as Environmental data.
7. Save the project as rda_ini.con.
8. Extract, with an editor or spreadsheet, the sample numbers and their weights from the DCCA solution file to a new file.

9. Put the weights as the first column, the sample numbers as the second. Add a column with zeroes and reshape the format so that all numbers are on a single line. Finally save the file as a text only file, say weight.txt.
10. Open the project file rda_ini.con with an editor and include the file weight.txt before the line "1.00000 = weight for sample". Save the file as rda_samw.con.
11. Open the project rda_samw.con in Canoco for Windows, click Options to modify the project to one that uses (a) Inter-species correlations as scaling and (b) Centre and standardize Species.
12. Analyze the modified project rda_samw.con giving the solution file ax1pol2z.sol.
13. Modify the text "Samp: Sample Scores" in the solution file to "Samp: Xample Scores" and save the file as text only file, ax1pol2z.xxx.
14. Open the original DCCA project, save it as ax2pol2z.con, enter the file ax1pol2z.xxx as the Covariable data file, check Delete Covariables, retain the covariables AX1, AX2 (or more if the order of the polynomial is 3 or 4), and ask to test the Significance of the first ordination axis.

Remarks:

- In steps 1-5 a CANOCO-readable file is made that consists of the environment-derived samples scores (SamE:) of the first axis of the original DCCA and as many powers thereof as are needed in the detrending. In the example, the file contains AX1 and AX1-squared.
- In steps 6-12, the variables AX1 and AX1-squared are regressed onto the environmental data of the DCCA, using the sample weights of the DCCA. The RDA-options of step 11 are to improve the numerical precision of the procedure. These are the steps that CANOCO uses to derive a second axis in a DCCA with detrending by polynomials. For this, see page 136 and 137 of Unimodal Models (Ter Braak & Prentice, 1988), in particular, Step A10 and Step A5, which is referred to in Step A10.
- In steps 13-14, the fitted values of the weighted regression (equation (A.10) on page 136) are made available to Canoco for Windows. In the terminology of page 137, Step A10, these steps "add the resulting variables as new variables to the matrix A".
- If the initial DCCA is a partial DCCA, i.e. had a Covariable Data file, these covariables should also be used in Step 6 as additional environmental variables. Also Steps 13-14 need to be modified to the effect that AX1 and AX2 (the environment-derived sample scores of step 12) are added to the Covariable Data File.

On comparing the eigenvalues and eigenvectors of the DCCA for the test of the second axis (in ax2pol2z.sol) with those of the original DCCA (in pol_dcc2.sol), we see that, as intended, the first eigenvalue (.1099) and first eigenvector in the former is identical to the second eigenvalue and eigenvector in the latter (allowing for rounding errors). Note that the second eigenvalue in the former (.0806) differs, however, from the third eigenvalue in the latter (.0655). This is due to the fact that the third axis in the original DCCA is also detrended with respect to AX1*AX2 (see (A.14) on page 137 of Unimodal Models); this variable is not included in the analysis of the test of the second axis. This difference has no consequence for the test of the significance of the second axis.

In the ALGAE example the second DCCA axis is not significant ($P = 0.94$ in ax2pol2z.log). Interpreting of the second axis as a contrast of ammonium to phosphate is therefore not necessarily meaningful here. For the sake of completeness we add that the canonical coefficients of the second axis no longer define a clear contrast of ammonium to phosphate when detrending is by polynomials of order 4.

8.2.5 Example DUNEBOOK - CCA and RDA on observational data

Problem:	Explore the relation between vegetation and the environmental conditions and management of dune meadows
Data:	Batterink & Wijffels (1983)
Booklet:	pp 78-79 (CCA), pp 120 and pp 145-148 (RDA) and text book: Jongman et al (1987)
Directory:	\CANOCO\SAMPLES\UNIMODAL\DUNEBOOK
Illustration of:	<ul style="list-style-type: none"> • How to display species as points in an RDA in CanoDraw. • How to display environmental variables by centroids in CanoDraw. • How to change the scale of the sample scores in CanoDraw. • How to interpret the numbers in the file SPEC_ENV.TAB after an RDA. • How to obtain correlation coefficients between each species and each environmental variable. • The effect of the scaling options of linear methods on the ordination diagram.

Files	Name	Description
Species	table01.dta	30 plant species in 20 dune meadows (Table 0.1 in Jongman et al, 1987)
Environmental	table02.dta	8 environmental variables (of which four define the nominal variable management type) in 20 dune meadows (Table 0.2 in Jongman et al, 1987)
Derived	cov_S_E.dta	covariances between species and 3 environmental variables (file spec_env.tab after running rda_spe.dta)
	cor_S_E.dta	correlations between species and 3 environmental variables (file spec_env.tab after running rda_cor.dta)
Project	cca_hill.con	CCA of plant species to 8 environmental variables (Hill's scaling with focus on inter-sample distances) (pp 78-79, p 164)
	cca_bipl.con	CCA of plant species to 8 environmental variables (biplot scaling with focus on inter-species distances) (p 164)
	rda_sam.con	RDA of plant species to 3 quantitative variables, focus on inter-sample distances (pp 120)
	rda_spe.con	RDA of plant species to 3 quantitative variables, focus on inter-species correlations
	rda_cor.con	RDA of plant species to 3 quantitative variables, focus on inter-species correlations, with species centred and standardized
	rda_eco.con	RDA of plant species to 8 environmental variables, with focus on inter-species correlations, from the Ecoscience paper (pp 145-148)

8.2.5.1 DUNEBOOK: CCA of dune vegetation on 8 environmental variables

The ordination diagram on page 76 of Unimodal Models can be made by Opening and Analyzing the project cca_hill.con. Then click CanoDraw and in CanoDraw ask (after the project is saved) for a triplot (*Create / Triplots / with Environmental variables* command). After you return from CanoDraw, you may wish to check that the first two eigenvalues are .46 and .29, as reported on page 78, and that the ordination diagram accounts for 63.8 % of the variance

in the weighted averages of species with respect to each of the environmental variables (in file `spec_env.tab`). In the project, Hill's scaling is used, focusing on inter-sample distances (scaling -1).

The same CCA analysis is also reported on pages 139-143 of Jongman et al. (1987). For unknown reasons, the canonical coefficients of the second axis in Table 5.10 (page 140 l.c.) are not precisely those found under `Regr`: in the solution file of the analysis, even if the variable `SF` is deleted in the project to ensure that `Standard Farming` is taken as the reference class (cf. the `Peat, Sand, Clay` example in Section 8.2.3). The intraset correlations of both axes in the log-window agree to the two digits reported in Table 5.10.

The file `spec_env.tab` (copied to `wa_8.dta`) can be used to check whether the inferences made from the ordination diagram in the paper hold true for the data.

In a later paper in *Unimodal Models* (pp. 153-187), another type of scaling of ordination axes is introduced, namely the biplot scaling with focus on inter-species distances (scaling 2). This scaling has three attractive features:

- The environmental biplot scores are correlations with the axes (footnote g on page 164 l.c.).
- Species points are weighted averages along the displayed environmental arrows (middle page 169 l.c.).
- The plot of species and sample points can be interpreted both by the centroid principle (page 167 l.c.) and by the biplot rule (page 171 l.c.).

Table 2 on page 164 of *Unimodal Models* summarizes the similarities and dissimilarities among the scalings. See also section 6.3.2.5.

You may wish to modify the project `cca_hill.con` (Open the project, click `Options..` and change in `Scaling: Unimodal Methods to Inter-species distances and biplot scaling`). In the example files the modified project is saved as `cca_bipl.con`. After clicking **Analyze...**, check in the log-window that the summary of the ordination is unchanged. Also the inter- and intra-set correlations of the environmental variables with the axes do not change. To evaluate the effect of the scaling on the ordination diagram click `CanoDraw`, and create in the newly defined `CanoDraw` project a triplot using *Create / Triplots / with Environmental variables*. The extremes of the ordination axes produced by `CanoDraw` do not exceed the scale marks -1 and +1; in scaling 2, this is the maximum range of the environmental biplot scores, because these scores are correlations with the ordination axes. The species and sample scores and centroids are plotted in this diagram by multiplying the scores of the solution file by a certain number. The multiplier can be found and also modified before a graph is created in `CanoDraw` if you select the *View / Diagram Settings* command and in the *Properties 1* page check the *Show rescaling coefficients for composite ordination diagrams* option. If it is checked, `CanoDraw` displays for diagrams, where rescaling might be applied, the *Rescaling of ordination scores* dialog, before the diagrams are created. Figure 8-1 shows the ordination diagram in scaling 2. It is described later in section 8.2.5.3 how to plot nominal environmental variables as symbols.

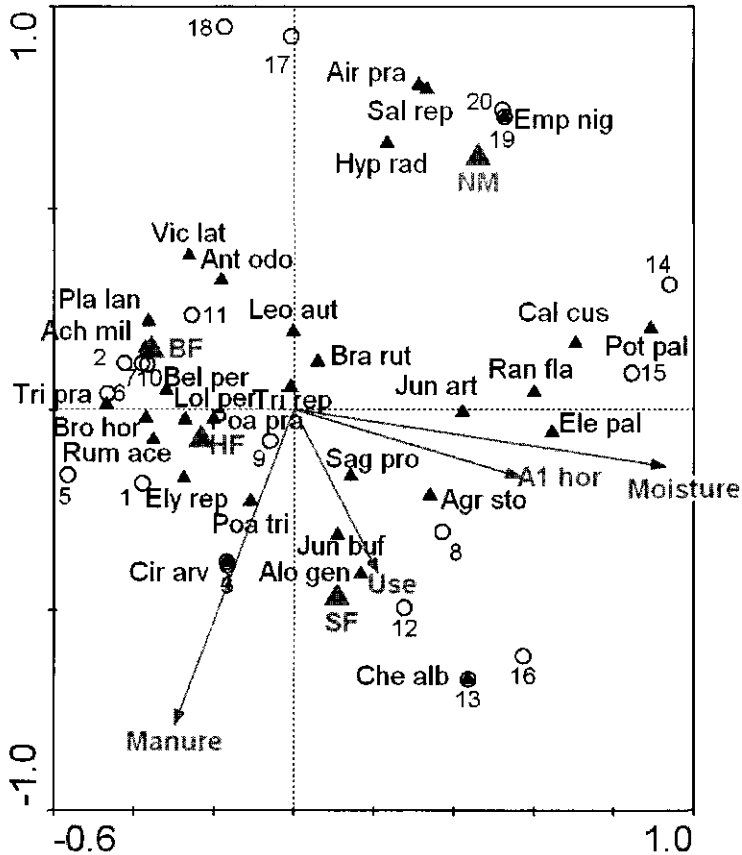


Figure 8-1 Dune meadow data: CCA triplot in biplot scaling with focus on species (scaling 2).

8.2.5.2 DUNEBOOK: RDA of dune vegetation on 3 quantitative environmental variables

The ordination diagram on page 121 of *Unimodal Models* can be made by Opening and Analyzing the project `rda_sam.con`. The scaling focuses on inter-sample distances and species scores are not post-transformed. Only three of the eight environmental variables are retained in the analysis (A1, Moisture and Manure). Then click Canodraw and in Canodraw ask for a triplot.

On page 148 of *Unimodal Models* it is argued that an RDA with quantitative variables only is most easily interpreted quantitatively when scaling 2 is used. In this scaling, the focus is on inter-species correlations and species scores are divided by their standard deviations. You can evaluate the changes in the diagram yourself by Opening and Analyzing the project `rda_spe.con`, and clicking on Canodraw, Triplot. See also the next section. You may wish to check that the summary of the ordination does not depend on the scaling chosen in the project. Also the inter- and intra-set correlations of the environmental variables with the axes do not change.

This example can also illustrate that a project (e.g. `rda_spe.con`) that uses centering by species and scaling 2 ("species scores divided by their standard error") does not yield the same eigenvalues as a project (e.g. `rda_cor.con`) that uses centering *and standardization* by species and scaling 1 or 2. The eigenvalues of the former analysis are .230 and .146 and those of the latter analysis are .186 and .105. In consequence, the ordination axes also differ. Although both resulting triplots display standardized species data, the species data are standardized *after* the extraction of the eigenvectors in the first analysis (`rda_spe.con`) and *prior* to the extraction of the eigenvectors in the second analysis (`rda_cor.con`) The former analysis is on a covariance matrix and the latter on a correlation matrix.

The resulting file `spec_env.tab` contains covariances after an analysis by `rda_spe.con` (file `Cov_S_E.dta`) and correlations after analysis by `rda_cor.con` (file `Cor_S_E.dta`). For example, the first set of 30 values in the file `Cor_S_E.dta` are

-.3176	.3573	-.1582	.0202	-.2010	-.1660	-.2527	.1242
-.0702	.5103	-.2264	-.1242	-.1986	.1477	-.0252	-.2230
-.5511	-.2094	-.5754	-.1935	.8710	.4011	-.0348	-.1071
-.1896	-.0773	.0655	-.2077	.0564	.3186		

These are the correlations between the 30 species and Moisture (the first environmental variable in the Environmental Data File).

☞ To obtain correlation coefficients between each of the species and each of the environmental variables, open a new project, specify your files with species and environmental variables and select an RDA method with centering and standardization by species). After clicking Analyze..., the resulting file `spec_env.tab` contains the desired correlations.

8.2.5.3 DUNEBOOK: RDA of dune vegetation on 8 quantitative environmental variables

The ordination diagram on page 145 of *Unimodal Models* can be made by Opening and Analyzing the project `rda_eco.con`. The scaling is the same as that of `rda_spe.con`; it focuses on inter-species correlations and the species scores are divided by their standard deviations. All eight environmental variables are used in the analysis. Then click the *CanoDraw* button and in *CanoDraw* ask for a triplot. If you do not like the arrows for the species, *CanoDraw* can display them as points if you select the *Project / Settings* command and check the *Display species as symbols even ...* option in the *Appearance* property page. Even when displayed as points, be sure to interpret the species points as arrows in PCA/RDA. The arrows for SF, BF, HF, NM can be changed into points by selecting the *Project / Nominal variables / Environmental variables* command and moving appropriate environmental variables into the right-hand list. The heads of the environmental arrows are based on the environmental biplot scores ("BipE:", Table 6.29), whereas the environmental points are based on the centroid scores for environmental class ("CenE:", Table 6.30).

The extremes of the ordination axes produced by *CanoDraw* carry the scale marks -1 and +1; in scaling 2, this is the maximum range of the species scores and the environmental biplot scores, because both these scores are correlations with the ordination axes. The sample scores and centroids of environmental variables are plotted on this diagram by multiplying the scores of the solution file by a certain number. The multiplier can be found (and changed) using the method described in the section 8.2.5.1. For the diagram on page 145 the multiplier was set to 0.46. Table 1 on page 143 summarizes what can be interpreted in the resulting ordination diagram.

8.2.6 Example WEEDS - A multi-species trend surface

Problem:	Detection of a spatial gradient in vegetation data
Data:	B. Post (1983: unpublished)
Booklet:	pp 79
Directory:	\CANOCO\SAMPLES\UNIMODAL\WEEDS
Illustration of:	<ul style="list-style-type: none"> • How to de-standardize canonical coefficients. • The arbitrariness of the sign of an ordination axis. • The meaning of 100% variance explained in the species-environment relation. • How to interpret large “Covariable influence” without covariable data. • Means, standard deviations and correlations in CCA versus RDA. • How to obtain means, standard deviations and correlations of variables.

Files	Name	Description
Species	weeds.dta	counts of 13 arable weeds across a barley field (96 plots)
Environmental	xy_coord.dta	spatial coordinates of the plot centers
Project	trend.con	CCA of weeds on spatial coordinates
	cca_sqrt.con	CCA of weeds on spatial coordinates (square-roots of counts)
	rda_sqrt.con	RDA of weeds on spatial coordinates (square-roots of counts)

8.2.6.1 WEEDS: A multi-species trend surface across an arable field

On page 79 of Unimodal Models, the results are given of a CCA of counts of weed species in 96 plots from a field of summer barley with the spatial coordinates of the plots used as explanatory variables. The aim of this analysis was to detect a spatial trend across the field, which could then perhaps be interpreted in terms of a known environmental gradient across field. The summary of the analysis from project trend.con shows that the eigenvalues of the CCA are rather small, the first two being 0.089 and 0.015. Because the first eigenvalue is about six times the second eigenvalue, the first ordination axis defines nevertheless a clear gradient. The canonical coefficients of the first axis in the solution file are -.2414 and -.1826 for the x- and y-coordinate. Recall that the coefficients in the solution file are for standardized variables. The field was 50 m x 100 m, reflected in the standard deviations of 9.2402 and 15.5980 for the x- and y-coordinates of the plots, respectively. To obtain the coefficients that apply to the x- and y-coordinates in meters, the standardized coefficients must be divided by the standard deviations of the variables, resulting in $b_1 = -.2414/9.2402 = -0.0261$ and $b_2 = -.1826/15.598 = -0.0117$. The first axis thus makes an angle of $\arctan(b_1/b_2) = \arctan(0.448) = 24$ degrees with the x-coordinate. Note that b_1 and b_2 are reported on page 79 as positive numbers but recall that the sign of an ordination axis is arbitrary. CANOCO could equally well have reported the axis scores with all the signs interchanged.

Two further points are worth noting about the output in the log-window of the analysis. First, the summary of the ordination reports that 85% of the variance of the species-environment relation is explained by the first axis and 100% by the first two axes. This 100% is a simple reminder that there are just two explanatory variables: the variance in the species data that two environmental variables explain (19.6%) is equally well explained by the two CCA axes. Second, the check on the influence of individual samples on the ordination results reports that the covariable influence of sample 87 is 3.0x the influence of an average sample. This would normally mean that sample 87 is an outlier in the space of the covariables. However, there are no covariables specified in the analysis! The reason is that sample 87 has an extremely high weight in the analysis. According to the column WEIGHT in solution file trend.sol, the weight of sample 87 is 2369.00, which is 3.0x the average weight of a sample. Despite the high total abundance in the sample (= WEIGHT), sample 87 contains effectively only $N_2 = 6.00$ species. In total, sample 87 contains 11 species (see data file weeds.dta) but half of them occur in low numbers. Looking at the other values in the column N2 shows that the effective number of species is in the range 4-7. All these statistics suggest that it might be wise to transform the counts by taking square-roots or logarithms.

In the CCA on square-root transformed counts (cca_sqrt.con), there is no report on the check for influence, because CANOCO found no outlying samples. The eigenvalues are even lower, the first being 3 times the second. On repeating the calculations to determine the direction of the first axis across the field, note that the standard deviations of the x- and y coordinates differ somewhat from those in the previous analysis. The means differ also. The reason for this is that these are weighted means, the weights being the sample totals reported as WEIGHT in the solution file. By taking square-roots, the sample totals change, hence the weighted means and standard deviations change. The calculations are $b_1 = -.1344 / 9.4953 = -0.01415$ and $b_2 = -.1295 / 16.1489 = 0.008019$ which results in an angle of 29 degrees with the x-coordinate.

The low first eigenvalue suggests that the gradient is short. By running a DCCA analysis with detrending by segments the gradient length is reported as 1.1 SD when counts are analyzed and 0.73 SD when square-roots are analyzed. This suggests the application of a linear method such as RDA.

A RDA on square-root transformed counts results in a first axis that explains 31% of the variance in the species data and that is 13 times as important as the second axis. The first axis makes an angle of 34 degrees with the x-coordinate, as shown from the calculations of $b_1 = .3746 / 9.6047 = 0.03900$ and $b_2 = .4332 / 16.3631 = 0.02647$. This happens to be the direction of the moisture gradient in March 1985 (page 79 of Unimodal Models).

The means and standard deviations of the x- and y-coordinates reported in the log-window of an RDA are the usual unweighted ones, unless you specified sample weights in the project. You may wish to check this by running the RDA without transforming the species data or by taking logs instead of square-roots; the means and standard deviations remain unchanged. The estimated angle changes of course, because the canonical coefficients in the solution file change.

☞ If you wish the means, standard deviations and correlations of variables to be calculated, specify the data file as Environmental Data File in CANOCO and ask for an RDA. If you do not have a natural Species Data File at hand, take a copy of the same data as Species Data File. If you would specify a CCA, weighted means are calculated.

8.2.7 Example SEASHORE - A vegetation succession study

Problem:	(1) Rate of land-uplift estimated from vegetation succession data along transects
	(2) Does vegetation succession track the land-uplift or does it lag behind?
Data:	Cramer & Hytteborn (1987)
Booklet:	pp 79 and Jongman et al. (1987: Exercise 5.6)
Directory:	\CANOCO\SAMPLES\UNIMODAL\SEASHORE
Illustration of:	<ul style="list-style-type: none"> • A DCCA with detrending by segments. • How to infer environmental change from the change in species composition. • How the sample weight affects the influence of samples in environmental space.

Files	Name	Description
Species	plantspe.dta	abundance of 68 plant species in 63 sites along 4 transects on a rising sea-shore, sampled in 1978 and 1984 (126 samples)
Environmental	elevyear.dta	elevation in 1984 and sampling year of 126 samples
Project	uplift.con	DCCA of plant species on elevation and year to estimate the rate of land-uplift

8.2.7.1 SEASHORE: Land-uplift estimated from vegetation succession data by DCCA

This example estimates environmental change from species data. The species composition of sites is sampled more than once, whereas the environmental information is available only from one sampling date. The estimation of environmental change is based on the assumption that the change in species composition is driven by the environmental variable, whose change is to be inferred.

The analysis by detrended CCA (DCCA) described on page 79 of Unimodal Models can be carried out with project uplift.con. Exercise 5.6 in Jongman et al. (1987) describes the way to estimate the uplift from the output. The 95% confidence interval cannot be easily constructed from the CANOCO output. Consult a statistician on this issue. More on the theory of the estimation can be found on pages 195 and 200 of Unimodal Models, where a similar model is used in RDA.

You may wish to check that the environmental data file contains the variables described on page 79, namely the elevation of each site in 1984 (the year of second sampling) and the year of sampling. The names of the samples can be used to check that samples in 1978 and 1984 from the same site have, indeed, the same value for elevation, which is essential for the method to work. For example, sample 1 and sample 24 have code names L1-03-78 and L1-03-84 and have the same elevation (-6.0 cm). Some of the sites are sampled only once, e.g. L2-01 in 1984 only. This is not problem in the analysis.

In the log-window, the samples 9 and 185 have large influences due to their position in environmental space. In this example, samples can be expected to be reported in pairs, because if an elevation of a site is extreme both samples from the site have an extreme elevation. This

would be true in RDA. Sample 9 and 185 are not from the same site, however. Sample 9 is from 1978 and has more influence in a CCA or DCCA than the corresponding sample from 1984 (sample 32 which has the same elevation), because sample 9 has a higher total abundance than sample 32 (19 vs 10, as can be found in the solution file under WEIGHT). Similarly, sample 185 is from 1984 and has a larger weight than the corresponding sample from 1978 because it has a higher total abundance.

There are two explanatory variables in the analysis and nevertheless the ordination summary does not report that the two axes explain 100% of the species-environment relation. Instead of 100%, 97.2% is explained in two dimensions. The reason for this slightly lower figure is that the analysis is a DCCA and not a CCA or RDA. This is something you do not need to be worried about. DCCA was used here to be able to interpret the succession in terms of SD (standard deviations of species turnover).

8.2.8 Example EPIALGAE - Conditional and marginal effects

Problem:	Effect of cooling-water discharge from a power plant on epilithic algal communities
Data:	Snoeijs & Prentice (1989)
Booklet:	pp 140-141 and 148-149
Directory:	\CANOCO\SAMPLES\UNIMODAL\EPIALGAE
Illustration of:	<ul style="list-style-type: none"> • The display of environmental variables by way of their canonical coefficients. • The difference between environmental biplot scores (BipE:) and canonical coefficients (Regr:),

Files	Name	Description
Species	epialgae.dta	88 epilithic algae in 181 samples from 11 sites
Environmental	sitetime.dta	indicators of site and month of sampling
	basinenv.dta	20 environmental variables at the 181 samples
Project	epi_cca.con	CCA with model "site+month" and the 20 environmental variables as Supplementary Environmental variables
	anomaly.con	idem, but with only the mean temperature anomaly and flow as Supplementary Environmental variables

8.2.8.1 EPIALGAE: Seasonal variation with a temperature anomaly in epilithic algae

The Forsmark Biotest Basin in Sweden is a shallow coastal ecosystem that receives brackish cooling-water discharge from a nuclear power plant. Snoeijs & Prentice (1989) investigated the effects of the cooling-water discharge on the epilithic algae communities. The communities were sampled "every third week throughout one year at 11 sites differentially affected by temperature and/or flow rate enhancement. The community variation was summarized in a CCA of species abundances as a function of site and date." The project epi_cca.con follows the analysis done in Snoeijs & Prentice (1989). The site and month points in Figure 10 of the paper are the centroids obtained by the project. The arrows for supplementary environmental variables are based on the biplot scores for environmental variables, labeled BipE: in the solution file. Figure 2 on page 141 of Unimodal Models shows the site points and the arrows for temperature anomaly and flow only. The solid arrows are the ones from Snoeijs & Prentice (1989). (The arrows may differ slightly as they were obtained using unweighted regression). It is argued on page 148 that these arrows do not display the real effect of the temperature anomaly and flow rate: given the flow rate, the arrow for temperature should run more to the west, as is clear from the symbols indicating flow and temperature of the sites. To display the effect of temperature given the flow rate, the arrow for temperature must be based on its scores in the table of regression coefficients (Regr:). The regression coefficients of temperature in the project epi_cca.con are conditional on all other 19 environmental variables. Its variance inflation factor is rather high (6.7). Therefore a second analysis was done (project anomaly.con), in which only flow and temperature were retained. On comparing the solution file of both projects, we see that the environmental biplot scores (BipE:) for flow and temperature are identical, but the regression coefficients differ somewhat. This is because the BipE: scores (being based on simple regressions) disregard the other environmental variables in the project, whereas the Regr: scores (being based on multiple

regressions) are partial coefficients that may change when other environmental variables are included.

8.2.9 Example STREAMS - Partial CCA on macro-invertebrates

Problem:	Effect of intensive agricultural land-use on macro-invertebrates in streams
Data:	Higler & Repko (1981), Ter Braak & Verdonschot (1995)
Booklet:	pp 156-175
Directory:	\CANOCO\SAMPLES\UNIMODAL\STREAMS
Illustration of:	<ul style="list-style-type: none"> • A standard CCA with both quantitative and qualitative explanatory variables. • Forward selection in CCA.

Files	Name	Description
Species	species.dta	197 macro-invertebrate taxa in 40 samples from the Leuvenum (L) and Uddel stream (U)
Environmental	environm.dta	29 environmental variables from all 40 samples, among which is the month of sampling
Derived	forward.txt	result of the project forward.con, as produced with Canoco for Windows (via the Clipboard copy)
Project	p_cca.con	partial CCA of taxa on Source Distance, EC, Discharge and the factor indicating Shrubs along the stream, adjusted for sampling month, yielding Fig. 3 of the paper. (p. 165 in Unimodal Models)
	forward.con	automatic forward selection yielding Table 3 (except all "-" entries) (p. 174 in Unimodal Models)

8.2.9.1 STREAMS: Partial CCA on macro-invertebrates

The example data are described on page 156 and page 157 (Table 1) of Unimodal Models. The project p_cca.con can be used to generate Fig. 3 of the paper. Note that the sampling months are specified as covariables. The selection of taxa displayed in that figure are all species in which the effective number of occurrences (N2 in the solution file) is greater than 4 and which also have a small N2-adjusted root mean square tolerance, as described in the section on species tolerance and sample heterogeneity (section 6.3.11.3 on page 178). Such selection can be achieved with CanoDraw for Windows using a species group defined by a combination of two rules. But if you wish to display the selected species only, a quicker way is to use the *Project / Suppress / Species* command. See page 163-171 of Unimodal Models for the interpretation of the analysis and the ordination diagram.

Table 3 on page 174 can be obtained with the project *forward.con*. The result is given as example file forward.txt. In the console version of CANOCO you must collect the entries of the table from the output file forward.log. The entries labeled "-" can be obtained most easily using the manual forward selection.

8.2.10 Example VEGCHANG - A study of change in time and space

Problem:	What caused the vegetation change in a wetland subject to water extraction and acidification?
Data:	Both & Van Wirdum (1981), Farjon & Wiertz (1988)
Booklet:	pp 189-200
Directory:	\CANOCO\SAMPLES\UNIMODAL\VEGCHANG
Illustration of:	<ul style="list-style-type: none"> • Models with interaction effects. • Decomposition of variance in space and time components. • Grid permutation. • Environmental change inferred from vegetation change.

Files	Name	Description
Species	grid_spe.dta	100 plant species in 20 plots arranged in a 5*4 grid, sampled in 1977 and 1988
Environmental Project	grid_env.dta	plot, strip and year indicators, pH and water depth
	dca.con	DCA of the species data with environmental interpretation
	pca.con	PCA of the species data with environmental interpretation
	rda0.con	RDA of the species data with model "plot*year", equivalent with pca.con
	fig2.con	RDA with model "plot+year+strip.year" used to produce fig.2 of the paper, which shows the spatial and temporal variation (pp 194)
	fig3.con	partial RDA focusing on how the change in time depends on space (strip; pp 196)
	yeartest.con	RDA with model "year" and covariables "plot" to test the change in species composition between 1977 and 1988 (pp 195)
	waterxyr.con	RDA with model "waterdepth88 . year" and covariables "plot and year" to test the interaction water depth.year, i.e. whether the change was constant against the alternative hypothesis that it depended on water depth
	watrxyr2.con	RDA with model waterdepth * year and covariables "plot and year"
	stripxyr.con	RDA to test the interaction strip . year
	ph_water.con	RDA to test the component "pH and water depth"
	yr_water.con	RDA on water depth and year to infer the change in water depth ($\Delta_{pH}=0$)
	yr_ph.con	RDA on pH and year to infer the change in pH ($\Delta_{waterdepth}=0$)
	yr_ph_wa.con	RDA with model "year+pH+Waterdepth" to infer the joint change (pp 196)

8.2.10.1 VEGCHANG: Spatial dependent vegetation change in permanent quadrates

The 40 vegetation samples of this example are from 20 plots that are arranged in a 4 by 5 grid (Fig. 1 on page 191 of Unimodal models). Each plot was sampled twice, namely in 1977 and 1988. As stated on page 191, a preliminary Detrended Correspondence Analysis was done which showed that the gradients are below 2 SD (project dca.con). Because the gradients are short, subsequent analyses use linear methods. A Principal Component Analysis showed that the vegetation change depended largely on the row label in fig. 2, referred to as strip in the paper. To see this, make an ordination diagram of the samples, based on the project pca.con, with CanoDraw. The interpretation is facilitated by the fact that the sample names are made as informative as possible. To see in the diagram the sample names, you must go into the dialog

displayed by the *Project / Settings* command, and in the *Appearance* page change the *Sample labels* option in the top row from *Indices* to *Names*. A further interpretation can be obtained by also plotting the environmental variables of this project. In CanoDraw, mark the variables strip (A,B,C,D,E) * year (1977,1988) as nominal variables within the dialog shown by the *Project / Nominal variables / Environmental variables* command. Then use the *Create / Scatter Plots / Environm. variables* command. The resulting plot of centroids and arrows looks pretty much like fig. 2 of the paper, except that the first axis is mirrored (and this can be changed using the *Flip axes: Horizontal* option in the *Project / Settings / Contents* dialog page. The correlations of pH and Waterdepth with the PCA axes are -0.77 and -0.91 (see the table CorE: in the solution file or the correlation matrix in the log-window). The PCA results could also be obtained by an RDA with model “plot*year” (project rda0.con), which has 39 parameters (40-1). For example, the sum of canonical eigenvalues of the project rda0.con is 1 (everything explained); also the species scores in both solution files are the same. This RDA-model is equivalent to “plot + year + plot.year”. Fig. 2 on page 194 of *Unimodal models* is based on a model with fewer degrees of freedom, namely “plot + year + strip.year” (pp 193; fig2.con), which has only slightly lower eigenvalues, thus confirming our initial observation that the change of the vegetation is approximately constant within strips. In fig2.con, Water depth and pH in 1988 are added as supplementary variables.

8.2.10.2 VEGCHANG: Statistical tests and decomposition of variance

This section gives details of the statistical tests described on page 195 of *Unimodal models* and of the decomposition of variance in Table 3 on page 196.

The significance of the change in species composition between 1977 and 1988 uses permutations within plots (project yeartest.con). The sum of squares for Time (year) in Table 3 is the trace of the test (.159). The interaction Waterdepth.Year and Strip.Year uses this permutation type (projects waterxyr.con and stripxyr.con). Again, the sum of squares for strip.year in Table 3 is the trace of the strip.year test.

The test of the joint effect of pH and water depth in 1988 as described on page 195 can be obtained via the permutation type of the “split-plot design” followed by “grid” permutation of the whole-plots. (project ph_water.con), as described in detail in section 8.3.5.1. This project also gives the sum of squares for pH and water depth (0.23). The sum of squares for Space (plot) can be obtained by entering all 20 plot indicators as environmental variables. To get this sum of squares, it does not matter whether or not year is taken as a covariable because the sampling design is balanced (all plots sampled twice). The residual of Space is obtained by subtraction. The sum of squares for Space.Time is the remainder after fitting Space and Time and can be obtained as $1 - 0.59 - 0.16 = 0.25$. The residual sum of squares for Space.time is obtained by subtraction ($0.25 - 0.08$).

Technical Note: In testing the interaction Waterdepth.year, it does not matter whether we use the variable Waterdep88 (Waterdepth in 1988 for all samples) or the variable Waterdep (compare the eigenvalues, F-ratios and P-values in projects waterxyr.con and watrxyr2.con). The reason for this is that Waterdep assumes a constant change of 27 cm in water depth between 1977 and 1988; the variable Waterdep is thus a linear combination of year and Waterdep88. Each linear combination of Waterd88 and year can thus be written as a linear combination of Waterdep and year by adjusting the coefficient for year. In consequence the sample and species scores in the solution files are identical, except for a change in sign. Also, the canonical weights for the two product variables waterdepth * year on each axis should be identical. The apparent difference in the solution file (under Regr:) disappears when the standard deviations of the variables are taken into account. For the Waterd88*year variable the canonical coefficient of the first axis is $2.2196 / 65.8138 = 0.0337$ and for Waterdep*year $-1.7696 / 52.4720 = -0.0337$.

8.2.10.3 VEGCHANG: Environmental change inferred from vegetation change

The projects yr_*.con are based on the change model of equation (1) on page 192 of Unimodal models. For example, the change in waterdepth can be inferred from the vegetation change and the 1988 measurements of water depth (in cm) using the project yr_water.con. From the solution file, the standardized canonical coefficients of year 1977 and waterd88 on the first axis from the solution file are -0.2936 and 0.3889, respectively. These coefficients need to be destandardized by division by their standard deviations in the log-window (.5000 and 12.6823), yielding -0.5872 and 0.0306647. From equation (2) on page 192, the inferred change is thus $0.5872/0.0306647 = 19$ cm (the unit of water depth in the data file).

8.2.11 Example DISEASES - Reduced-rank regression

Problem:	How do the effects of SES on standardized mortality rates (smr) change with time?
Data:	Kunst et al. (1990), Ter Braak & Looman (1994)
Booklet:	pp 239-256
Directory:	\CANOCO\SAMPLES\UNIMODAL\DISEASES
Illustration of:	<ul style="list-style-type: none"> • A regression biplot based on reduced-rank regression (RDA). • A t-value biplot.

Files	Name	Description
Species	mortality.dta	11 causes of death (smr's) in 39 Dutch regions in 4 periods (decades)
Environmental	explanat.dta	period indicators (P1-P4), SES, Urbanization (URB) and religion (CAT)
Project	stomca.con	multiple regression project with model Period + Period.(SES+URB+CAT) to obtain the t-ratios of Table 2 for Stomach Cancer (StomCa)
	fig1.con	RDA project with model Period + Period.(SES+URB+CAT) from which figs.1 - 3 are constructed
	fig4.con	partial RDA project with model Period.SES, adjusted for Period + Period.(URB+CAT), to obtain Fig. 4
	sesurbca.con	RDA with main effects only (SES, URB,CAT) to show that there are no outliers in these variables

8.2.11.1 DISEASES: An example of a reduced-rank regression biplot

This is an example of multivariate multiple regression and reduced-rank regression. As in the paper, we start with the univariate results, which are summarized in Table 2 (page 241 of Unimodal Models). The t-ratios and R in this table are easy to obtain with CANOCO, as illustrated for one variable, Stomach Cancer, in the project stomca.con. This project defines an RDA for StomCa only with model Period + Period.(SES+URB+CAT). For this, all other causes of death are deleted. The interaction terms of the model are defined by products of each of the period indicators (P1 - P4) with SES, URB and CAT. Note that the variables SES, URB and CAT are deleted as environmental variables. The t-values for StomCa in Table 2 for SES.Period can be found in the solution file in the table of t-values (tVal:) in the rows for SES*P1 - SES*P4, except for their sign. The difference in sign is caused by the fact that the species score for StomCa is -1 (see section 8.4.4 for further details). The value of R is the square-root of the percentage explained (%Expl) for StomCA in the table of cumulative fits (Cfit:). The regression coefficients themselves can also be obtained from this project by some post-calculations (section 8.4.4).

In log-window of the project, there is a long list of samples that have a high influence for particular variables, e.g. sample 1 has 8.2 times the influence when variable 8 would be the only predictor. This means that the value of variable 8 (SES*P1) is extreme. Remarkably, no sample is found to have a large influence in the column "Environment space influence", i.e. in the full predictor space. For multiple regression, it is this column that counts. The influence of individual variables is given to help you detect which variable may cause the influence. In this case, the alarm is false because it is an interaction variable with a dummy variable: the variable SES*P1 has the value 0 when the sample is from other periods and is equal to the SES of the

region when the sample is from period 1. The variable SES*P1 thus has numerous zeroes and, contrasted with these, the value in period 1 may be extreme. You may wish to check that CANOCO does not detect outliers in the variable SES itself (project sesurbca.con).

A reduced-rank regression biplot represents the table of regression coefficients. Figure 1 of the paper is based on the project fig1.con, which uses the same model as in the project stomca.con. The 11 causes of death are weighted inversely to their error variance, by selecting "Standardize by error variance" in the "Centering and Standardization" wizard page. The resulting relative weights can be found in the log-window, but also in the solution file alongside the table of species scores (e.g. StomCa is given weight 1.26). The reduced-rank models for rank 1, 2, 3 and 4 account for 56, 71, 74 and 76 of the total sum of squares as given in the summary of the ordination in the log-window.

The arrows for the causes of death in Fig. 1 can be obtained with CanoDraw for Windows. For the points of the periods P1 - P4, CanoDraw uses the centroids in the solution file (CenE:), if P1 - P4 are specified as nominal variables, whereas Fig. 1 uses the centered regression coefficients for periods. Because of the balance in the data, the difference is minor. The arrows of SES, URB and CAT with period in Fig. 1 of the paper can be obtained with CanoDraw as a part of the regression biplot, where the canonical coefficients (listed under Regr: in the solution file) of environmental variables are used. In other graphs, CanoDraw uses the environmental biplot scores (listed under BipE: in the solution file). The difference between these two types of scores is explained in Figure 2 on page 141 and on page 148 of Unimodal Models.

With project fig1.con, some of Fig. 2 of the paper can be obtained with CanoDraw. The *T-values Biplot* command under *Biplots and Joint Plots* in *Create* menu gives the positions on the arrows that mark the transition between solid and dashed parts.

The project fig4.con does a partial RDA to focus on the period-dependent effects of SES. The focus is obtained by defining Period, Period*URB and Period*CAT as covariables, and Period*SES as environmental variables. The contribution of SES to the explained variance is only 4.1% (see log-window). It is this small, but significant part that is of interest in this study. Fig. 4 shows the period-dependent SES effects. The arrow heads for diseases in Fig. 4 are the species scores (Spec:), the mark on the arrow gives the coordinates of the t-value biplot (StBi:, Table 8.5). The arrow heads of SES₁, ..., SES₄ are the canonical coefficients (Regr:) and the marks on the arrows are the environmental coordinates for t-value biplot (EtBi:). Note that the column %(E) in Table 8.5 gives the pure SES.Period component, in the sense of section 8.3.1.2, for each separate response variable.

The columns in Table 8.5 and Table 8.6 for the third and fourth axes contain zeroes because the t-value biplot is optimized for two dimensions. This default can be changed in the initialization file CANOCO.INI (option (09) in the Table 7.1).

8.3 Examples of significance tests

This section illustrates the extensive facilities for significance testing in CANOCO 4.5 using real-life examples. All tests use Monte Carlo permutations. The examples show how to specify the appropriate permutation type for a number of commonly applied research designs. In addition, we explain the distinction between "simple tests" and "partial tests" and between "design-based permutation" and "model-based permutation". The examples are arranged in order of increasing sophistication.

In this section we also illustrate the decomposition of the variance into different components, as popularized by Borcard et al. (1992), and the Principal Response Curves method (Van den Brink & Ter Braak 1998, 1999) used to display time-dependent treatment effects in a repeated measurement design.

8.3.1 Example DUNETEST - Simple and partial tests

Problem:	Determine the significance of differences in vegetation between management types
Data:	Batterink & Wijffels (1983)
Directory:	\CANOCO\SAMPLES\PERMUTIO\DUNETEST
Illustration of:	<ul style="list-style-type: none"> • A permutation test. • A partial test. • How to account for other variables in a permutation test. • How an F-ratio is calculated. • How to decompose the total variance into different components

Files	Name	Description
Species	table01.dta	see DUNEBOOK, page 241
Environmental	table02.dta	see DUNEBOOK, page 241
Project	manage1.con	test of differences among management types
	manage2.con	test of differences among management types after accounting for soil characteristics
	manage3.con	test of differences among management types after accounting for soil characteristics and manure
	soil1.con	test of effect of soil characteristics (A1 thickness and moisture) on the vegetation
	soil2.con	test of effect of soil characteristics after accounting for management types
	manasoil.con	test of effect of management and soil jointly

8.3.1.1 DUNETEST: Overall and partial tests using unrestricted permutation

The dune meadow data were collected to investigate the differences in vegetation among dune meadows that have been subjected to different management regimes, namely standard farming (SF), biodynamical farming (BF), hobby farming (HF) and nature management (NM). To investigate whether the observed differences in vegetation could be accounted for by pure chance, we can apply a Monte Carlo permutation test. By analyzing the data with the project manage1.con, the samples in the species data are randomly permuted (199 times). In nearly all

Table 8.5 Species coordinates of the t-value biplot (StBi):

Coordinates are from the solution file of a partial RDA obtained with project fig4.con, giving the marks on the disease arrows in Fig.4 on page 253 of Unimodal models. $\%(E+C)$ = percentage variance explained by both the environmental and covariables; $\%(E)$ = percentage variance explained by the environmental variables after adjustment for the covariables .

N	NAME	AX1	AX2	AX3	AX4	%(E+C)	%(E)
	EIG	0.0249	0.0139	0.0015	0.0006		
1	Stom Ca	0.6896	0.1861	0	0	82.02	1.27
2	Colo Ca	-0.5232	-0.0093	0	0	55.72	6.64
3	Lung Ca	-0.1585	0.0618	0	0	93.08	6.87
4	Pros Ca	-0.8007	-0.1008	0	0	49.43	2.29
5	Diab Me	-0.4616	0.6081	0	0	46.12	2.91
6	Isch ea	-0.0683	0.3231	0	0	90.74	2.46
7	OthH ea	0.5644	2.7425	0	0	67.07	0.39
8	Arte Di	0.29	-0.8101	0	0	61.99	1.91
9	COLD	1.0463	-0.12	0	0	59.25	1.2
10	Traf fic	0.1637	0.1396	0	0	83.01	10.52
11	NonT raf	0.4066	-1.041	0	0	65.91	0.94

Table 8.6 Environmental coordinates for the t-value biplot (EtBi):

Coordinates are from the solution file of a partial RDA obtained with project fig4.con, giving the marks on the SES arrows in Fig.4 on page 253 of Unimodal models.

N	NAME	AX1	AX2	AX3	AX4
	EIG	0.0249	0.0139	0.0015	0.0006
8	SES*P2	-0.5446	0.0847	0	0
9	SES*P3	-0.4563	-0.5331	0	0
10	SES*P4	-0.2938	-0.6043	0	0
11	SES*P1	-0.6394	0.586	0	0

permuted data sets, the F-ratio based on the trace (= sum of the canonical eigenvalues) is lower than in the data as observed (observed F-ratio = 2.13). In consequence, the reported P-value is 0.01. It is therefore concluded that the differences in vegetation among the management types are statistically significant.

Here is an example of how to calculate the observed F-ratio from the values reported in the summary of the ordination. The F-ratio is formally defined by equation (3.5). The numerator is equal to the regression sum of squares divided by the number of parameters tested. In CANOCO, the regression sum of squares is equal to the sum of the canonical eigenvalues or trace (0.604). The number of parameters tested is 3 in this case (number of classes -1). The denominator is the residual sum of squares divided by number of degrees of freedom ($n-p-q = 20-1-3 = 16$). The residual sum of squares is the sum of the unconstrained eigenvalues (2.115) minus the sum of the canonical eigenvalues (0.604). In full, the F-ratio is thus $(0.604/3) / \{(2.115-0.604)/(20-1-3)\} = 2.13$.

If you repeat the analysis with other initial seeds for the random number generator, e.g. by clicking Randomize.. in the Permutation Type page of the wizard, you may see that the reported significance level may vary between analyses. The variability can be seen more clearly by decreasing the number of permutations to 19, say. The reason is that other permuted data sets are used if other initial seeds are used. The variability in the P-value decreases with the number of permutations. Apart from this variability, the test is exact. The same P-value would have been obtained if the test statistic was simply the regression sum of squares. The reason is that this is a simple test, as there are no covariables (see section 3.7.4).

In CANOCO, you can also choose another test statistic, namely an F-ratio based on the first eigenvalue. This statistic has more power against 1-dimensional alternative hypotheses (i.e. if the effects of the management types can be represented by single ordination axis). This case also yields an exact test, but, even after many permutations, it does not need to yield the same P-value. In the example, the first eigenvalue is 0.32 and the resulting F-ratio is $0.319 / \{(2.115-0.319)/16\} = 2.84$. This test statistic gives in this example a somewhat higher P-value than the overall test.

Another option which you can set is: "Permutation under .. Reduced model or Full model". Recent research shows that there is little reason to ever change the default which permutes the residuals of the Reduced (null) model (Anderson & Legendre, 1999). The default in this simple case is equivalent to the permutation of the raw (transformed) species data and results in an exact test. The test using permutation under the Full model would not be exact for small n.

One may argue that the difference in vegetation among management types is not caused by management type but by the differences in the (initial?) soil characteristics of the meadows. The standard farms are on the driest places, the nature management meadows on the wettest places. With the project soil1.con we can check that the effect of the two soil variables, A1 and Moisture, on the vegetation is highly significant ($P < 0.01$). The question is thus whether the differences caused by the soil characteristics can account for the differences among management types, or phrased differently, whether there is still a difference in vegetation among management types after accounting for the effect of the soil characteristics. This can be investigated with CANOCO by specifying the soil characteristics as Covariables, as is done in the project manage2.con. The test so performed is called a **partial test** (based on a partial CCA). In partial tests, it is essential to use an F-ratio statistic, as CANOCO does, to ensure a good level-accuracy for the test (i.e. to ensure that the reported P-value is accurate). In the example, the new sum of canonical eigenvalues is .47, the resulting F-ratio is 1.98 and $P < 0.01$. In conclusion, there remains systematic differences in vegetation among management types after accounting for the effects of the soil characteristics A1 and Moisture.

Management type is perhaps determined mainly by the amount of manure that is applied. If we also account for the effect of manure (manage3.con), the remaining differences among

management types are no longer statistically significant. Phrased differently, the variables A1, Moisture and Manure are in this data set sufficient to explain the differences in vegetation among management types.

8.3.1.2 DUNETEST: Decomposition of variance

In a balanced designed experiment, each treatment factor explains a unique amount of variance. This is the basis of the usual analysis of variance table. In unbalanced situations, it depends on the other variables in the model how much variance a variable or factor explains. Table 8.7 (top half) gives an example: management type and soil each explain 29% and 25% of the total inertia when taken alone, but together they explain only 48%. (Inertia is the measure of variance in CA and CCA and is related to the chi-square statistic; see equation (6.38)). If management and soil were uncorrelated, we would have expected that management and soil explain together $29 + 25 = 54\%$. The difference ($54 - 48 = 6\%$) is their shared variance. To calculate the variance that can be uniquely attributed to the management, the soil variables must be taken as covariables. Adjusted for soil, management explains only 22%, whereas soil explains only 19% after adjustment for management (Table 8.7, lower half). Table 8.8 shows a decomposition of the total inertia into terms that sum up to the total inertia. This decomposition can be applied with any number of factors or variables (Whittaker, 1984). The decomposition was introduced in ecology by Borcard et al. (1992) to decompose the variance explained by environmental as opposed to spatial variables.

Table 8.8 also show an alternative way to obtain the shared variance (using the results from two CANOCO projects). Notice that the shared variance can be negative. An example was given by Baar & Ter Braak (1996).

Table 8.7 Variance explained by management type and soil characteristics in the dune meadow data.

Analyses without covariables

File	Source	Explained variance
managel.log	Management ignoring Soil	0.604 (29%)
soil1.log	Soil ignoring Management	0.535 (25%)
manasoil.log	Management and Soil	1.006 (48%)
	Shared: $0.604 + 0.535 - 1.006 =$	0.133 (6%)
	total inertia	2.115

Analyses adjusted for covariables

manage2.log	Management adjusted for Soil	0.471 (22%)
soil2.log	Soil adjusted for Management	0.402 (19%)

Explained variance = sum of all canonical eigenvalues; the total inertia = 2.115

Table 8.8 Variance decomposition of the effect of management and soil on dune meadow vegetation.

Component	Source	Calculation	Variance	%
a	Pure Management		0.471	22
b	Shared	$0.604 - 0.471 =$	0.133	6
c	Pure Soil		0.402	19
d	Residual	$2.115 - 1.006 =$	1.109	53
Total			2.115	100

8.3.2 Example PLOUGH - A randomized block experiment

Problem:	Effect of ploughing time on weeds
Data:	B. Post (unpub.)
Directory:	C:\CANOCO\SAMPLES\PERMUTIO\PLOUGH
Illustration of:	<ul style="list-style-type: none"> • Analysis of a randomized block experiment. • Preparation of a data file that defines the design of the experiment. • Definition of permutation groups (blocks). • Unrestricted permutation within blocks. • Mimicking blocks by a split-plot design.

Files	Name	Description
Species	weeds83.dta	13 weeds in 12 plots of a randomized block experiment
Environmental	factors.dta	blocks and treatments coded by factor levels
	bloctime.dta	7 dummy variables indicating block and treatment
	design.dta	as bloctime.dta, but in condensed format
Project	plough.con	test of the effect of the treatment (ploughing time) on the weeds by permutation within blocks
	mimicble.con	as plough.con without the usage of the block indicator variables

8.3.2.1 PLOUGH: Unrestricted permutation within blocks

Post (1986) carried out an experiment to investigate, among other things, the effect of the time of ploughing on the subsequent weed vegetation composition (13 species) in summer barley. There were three ploughing times. The experiment was a randomized block experiment of 12 sample units laid out in four complete blocks of three sample units each and was carried out in 1983.

Experimental data are commonly analyzed by the analysis of variance. Because the interest was not directed on one particular weed species, a multivariate analysis of variance is required but it cannot be used in this case, because the number of response variables (13 species) is larger than the number of experimental units (12). In this section we show that partial redundancy analysis combined with Monte Carlo permutation tests is an attractive alternative escaping the restriction on the number of response variables.

The samples in the data files are arranged in blocks of 3 samples; the treatments within blocks are in a standard order, although this is not needed for CANOCO. This is shown in the file factors.dta by the two variables Block and Treatment that have values 1, 2, 3 and 4 and 1, 2, 3, respectively. The file can be read by CANOCO, but is not suited for our purpose: CANOCO would consider the variables as being quantitative whereas they are qualitative. For the analysis of the experiment, each block and each treatment must be indicated by a separate dummy (0/1) variable, as shown in the file bloctime.dta in full format. The same information is represented in the file design.dta in condensed format.

Table 8.9 Experimental design with blocks and treatments coded in three different ways.

- (1) by two factors, columns 2 and 3, as in file factors.dta [not useful for the analysis]
- (2) by 7 dummy variables in full format, columns 4-10 as in file bloctime.dta
- (3) by 7 dummy variables in condensed format, columns 11-15 as in file design.dta.
(var = variable number, val = value)

The seven dummy variables are named: 1 = block 1, 2 = block 2, 3 = block 3, 4 = block 4, and 5 = pltime 1, 6 = pltime 2, 7 = pltime 3.

1	2	3	column number							11	12	13	14	15
			4	5	6	7	8	9	10					
plot	block	treat	block				treatment			plot	var	val	var	val
			1	2	3	4	1	2	3					
1	1	1	1	0	0	0	1	0	0	1	1	1.	5	1.
2	1	2	1	0	0	0	0	1	0	2	1	1.	6	1.
3	1	3	1	0	0	0	0	0	1	3	1	1.	7	1.
4	2	1	0	1	0	0	1	0	0	4	2	1.	5	1.
5	2	2	0	1	0	0	0	1	0	5	2	1.	6	1.
6	2	3	0	1	0	0	0	0	1	6	2	1.	7	1.
7	3	1	0	0	1	0	1	0	0	7	3	1.	5	1.
8	3	2	0	0	1	0	0	1	0	8	3	1.	6	1.
9	3	3	0	0	1	0	0	0	1	9	3	1.	7	1.
10	4	1	0	0	0	1	1	0	0	10	4	1.	5	1.
11	4	2	0	0	0	1	0	1	0	11	4	1.	6	1.
12	4	3	0	0	0	1	0	0	1	12	4	1.	7	1.

We are interested in the effect of ploughing time and want to eliminate the possible effects of blocks. The block variables (variables 1-4) should thus be entered in CANOCO as Covariables, and the ploughing time variables (variables 5-7) as Environmental Variables. This is achieved by specifying the file bloctime.dta both as the Environmental Data file and as the Covariable Data file and by deleting the superfluous environmental variables (namely, the blocks) and the superfluous covariables (namely the ploughing times). We further ask for an RDA on log-transformed species data and for a global permutation test. To obtain permutations of samples within blocks we check "Unrestricted permutations" and "Blocks defined by covariables". Subsequently all four block variables are selected to define the blocks. In the log-window of the analysis (from project plough.con), we can see that the resulting P-value is 0.02 showing that ploughing time has an effect on the weed composition. You can check whether CANOCO applied the intended permutation type from the report on the "Sample arrangement in the permutation test". This reports says, as intended, that samples 1, 2 and 3 are randomly permuted in block 1, samples 4, 5 and 6 in block 2 etc.

RDA was chosen because an initial DCCA showed that the gradient is very short. See also section 8.2.6.1. The RDA ordination diagram can be made with CanoDraw and shows, for example, that *Chenopodium album* and *Spergula arvensis* are most abundant after ploughing on March 9 (Pltime 1) whereas *Capsella bursa-pastoris* is most abundant after ploughing on March 23 (Pltime 2).

In spring 1984 the plots were all cultivated by rotary tillage on a single day (to obtain a more even distribution of seeds in the seed bank). The counts made thereafter in May 1984 were subjected to the same analysis as the 1983 counts ($\lambda_1 = 0.069$) but failed to show significant differences ($P = 0.45$, test on first eigenvalue). Apparently the one single date of rotary tillage canceled the effects of the previous treatments and/or recruited the same seedlings from the field seed bank.

In the example, the blocks are of equal size and the plots are arranged in a systematic way in the data file. In this special case there is an alternative way to specify permutations within blocks, which we mention for completeness. This alternative is exact for tests of simple hypotheses, but is approximate for partial tests. The alternative might be useful if you did not code the blocks in your data file, for example because there may be too many blocks, as may occur in a paired plots design (blocks of size 2). This alternative way is shown in the project `mimicblc.con`. For a simple hypothesis test, as in the example, the resulting P-values are the same, despite the fact that the F-ratios differ. The F-ratios differ because the residual sums of squares differ; the block effects are not subtracted from the residual sum of squares in the project `mimicblc.con`. The two ways of testing are no longer equivalent if there are other (non-block) covariables in the analysis, i.e. if the test is a partial test.

8.3.3 Example E40 - A multifactorial experiment with fixed factors

Problem:	Effect of N- and P-addition on the undergrowth in pine forest
Data:	Van Dobben, Ter Braak & Dirkse (1999)
Directory:	C:\CANOCO\SAMPLES\PERMUTIO\E40
Illustration of:	<ul style="list-style-type: none"> • How to analyze a factorial experiment. • How to construct a (multivariate) analysis of variance table in CANOCO. • How to test an interaction effect. • How to test a main effect. • How to specify permutation blocks from covariables in the console version of CANOCO.

Files	Name	Description
Species	E40_spec.dta	103 plant species (cover percentages) in 32 plots in experiment E40
Environmental	E40_dsgn.dta	dummy variables coding for 4 blocks, 4 levels of N and 2 levels of P of the N*P factorial experiment in 4 blocks
Project	e40_nxp.con	RDA of treatment groups, model N*P adjusted for blocks
	e40_np.con	test of the interaction effect N.P, adjusted for N, P and blocks
	e40_n.con	test of effect of N, adjusted for P and blocks
	e40_p.con	test of effect of P, adjusted for N and blocks
	e40_nap.con	parsimonious model, N + P, adjusted for blocks
	n_design.con	design-based permutation à la Edgington to test the N effect
	p_design.con	idem to test the P effect
	e40_err.con	(error) project of the console version of CANOCO in which covariables are in the wrong order to define blocks; this error cannot occur in the windows version

8.3.3.1 E40: A multifactorial experiment with fixed factors

The data of this example come from a factorial experiment with code name E40, that was carried out in pine forest in Lisselbo (Sweden) by Tamm et al. (1974) as an optimal nutrition experiment for pine tree growth. There were 32 plots, in 4 complete, randomized blocks of 8 plots each. In each block, all combinations of 4 levels of N (supplied as ammonium nitrate) and 2 levels of P (supplied as compound PK fertilizer) were applied. In the design file, e40_dsgn.dta, the levels of N-addition are coded by the dummy variables N1, N2, N3 and N0 (no additions) and the levels of P-addition by the dummy variables P1 and P0 (no addition). Block indicators Block1, ..., Block4 complete the design file. The design file is in condensed format: there are three couplets (pairs of columns) indicating that each plot receives one level of N, one level of P and is in one Block only. In 1987 the undergrowth of the plots was surveyed. The cover percentage of plants (including tree saplings, bryophytes and lichens) was estimated and scored on a ten-point scale. There were 103 species found. The scale points are re-expressed as mid-cover-percentages in the species data file.

Table 8.10 Analysis of variance table of all species simultaneously for experiment E40 obtained by RDA on log-transformed cover percentage data. (df = degrees of freedom, total SS = sum of squares totaled across species, F = F-ratio, P = Monte Carlo significance level, 199 permutations, project = name of project from which results are taken).

Source	df	total SS	F-ratio	P-value	project
blocks	3	0.083			e40_nxp
N	3	0.544	13.527	0.005	e40_n
P	1	0.051	3.814	0.010	e40_p
N.P	3	0.035	0.857	0.645	e40_np
Residual	21	0.286			e40_nxp
Total	31	1.000			

The published ordination diagram of this experiment (van Dobben et al, 1999), reproduced as Figure 8-2, is based on the species scores and environmental centroids in the solution file of project e40_nxp.con. In this project, the design file is specified both as Environmental data file and as Covariable Data file. The block indicators are the only variables kept as covariables. All (!) environmental variables are deleted. After checking the interaction box, 8 interaction variables are created, namely the products of the 4 levels of N with P1 and with P0. This uniquely identifies all 8 treatment combinations, whose centroids are plotted as begin- and end-points of the arrows for N-levels in Figure 8-2, the begin-point representing P0 and the end-point P1. The chosen scaling focuses on inter-sample distances because all environmental variables are nominal. The diagram explains 92% of the variance of the fitted abundance values of the full model; λ = eigenvalues of the axes.

Table 8.10 shows the analysis of variance table for the experiment as compiled on the basis of four different analyses by (partial) RDA. The figures in each row are taken from the log-window of the project named in the last column. For example, to test the effect of N-addition on the undergrowth, Blocks and P are specified as Covariables, and N as Environmental variables in the project file e40_n.con. The permutation type is unrestricted permutation within blocks. The sum of squares in Table 8.10 for N, P and N.P is the sum of canonical eigenvalues of each respective analysis and is also given under the name 'Trace' along with the F-ratio and P-value of the Monte Carlo permutation test. Although not fully conventional, the F-ratio of N and P neglects the N.P interaction; its denominator pools the N.P interaction with the residual given in the table (the pooled residual SS is 0.322). The reason for this is that the "pure" interaction cannot easily be specified as covariable in CANOCO. Because testing main effects is usually meaningful only if the interaction is not significant, this deviation from the conventional ANOVA table does no harm.

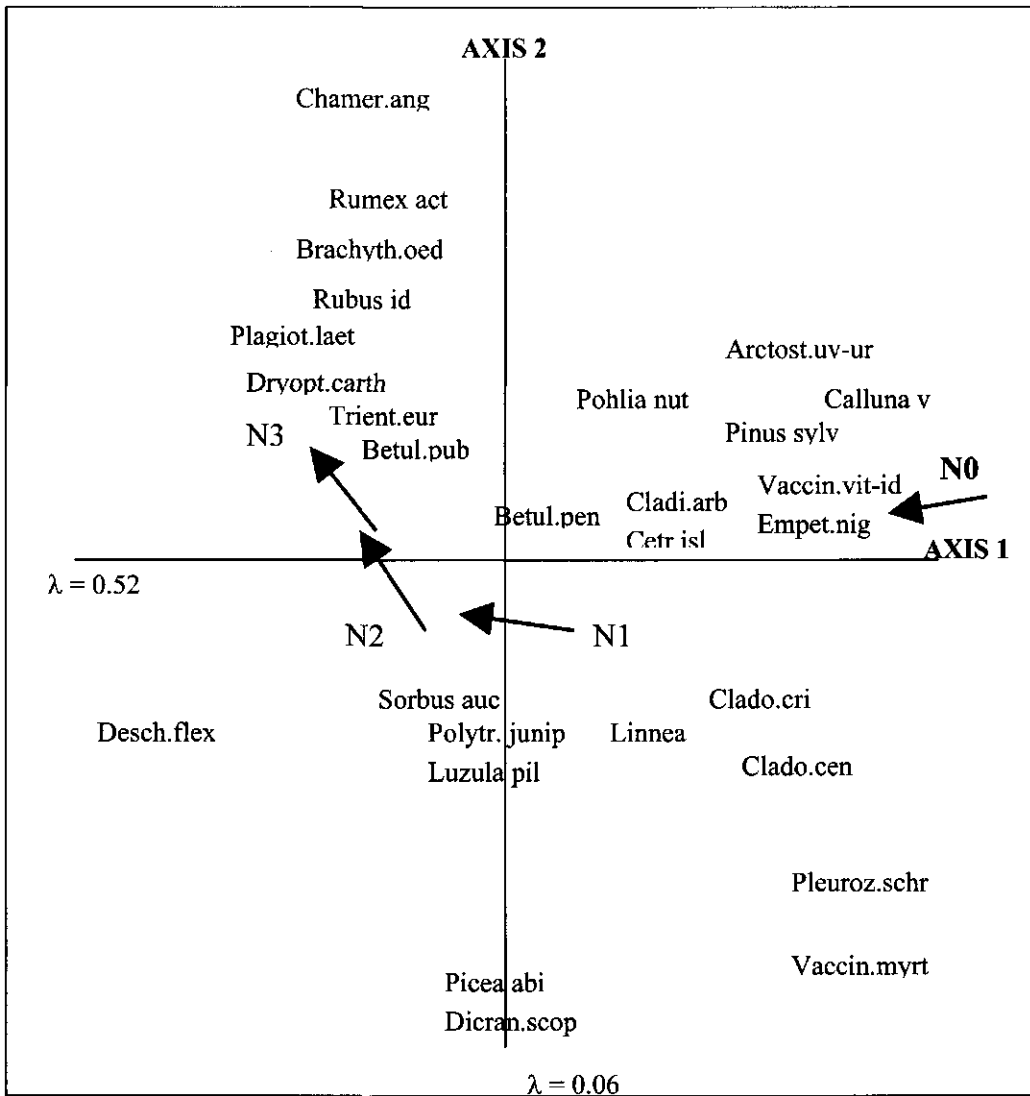


Figure 8-2 Ordination diagram based on RDA with model block+N*PK in E40

Explanation of abbreviated species names: *Arctost.uv-ur*, *Arctostaphylos uva-ursi* (L.)Sprengel; *Betul.pen*, *Betula pendula* Roth; *Betul.pub*, *Betula pubescens* Ehrhart; *Brachyth.oed*, *Brachythecium oedipodium* (Mitt.)Jaeg.; *Calluna v*, *Calluna vulgaris* (L.)Hull; *Cetr.isl*, *Cetraria islandica* (L.)Ach.; *Chamer.ang*, *Chamerion angustifolium* (L.)Hohub; *Cladi.arb*, *Cladina arbuscula* (Wallr.)Hale&Culb.; *Clado.cen*, *Cladonia cenotea* (Ach.)Schaerer; *Clado.cri*, *Cladonia crispata* (Ach.)Flotow; *Desch.flex*, *Deschampsia flexuosa* (L.)Trinius; *Dicran.scop*, *Dicranum scoparium* Hedw.; *Dryopt.carth*, *Dryopteris carthusiana* (Villars)H.P.Fuchs; *Empet.nig*, *Empetrum nigrum* L.; *Hyloc.spl*, *Hylocomium splendens* (Hedw.)Schimp.; *Linnea bor*, *Linnaea borealis* L.; *Luzula pil*, *Luzula pilosa* (L.)Willdenow; *Picea abi*, *Picea abies* (L.)Karsten; *Pinus sylv*, *Pinus sylvestris* L.; *Plagiot.laet*, *Plagiothecium laetum* Schimp.; *Pleuroz.schr*, *Pleurozium schreberi* (Brid.)Mitt.; *Pohlia nut*, *Pohlia nutans* (Hedw.)Lindb.; *Polytr.junip*, *Polytrichum juniperinum* Hedw.; *Rubus id*, *Rubus idaeus* L.; *Rumex act*, *Rumex acetosella* L.; *Sorbus auc*, *Sorbus aucuparia* L.; *Trient.eur*, *Trientalis europaea* L.; *Vaccin.myrt*, *Vaccinium myrtillus* L.; *Vaccin.vit-id*, *Vaccinium vitis-idaea* L..

The figures in the rows for blocks and residual in Table 8.10 require further explanation. The sum of squares for the blocks is equal to the total variance (1.0) minus the sum of all unconstrained eigenvalues (0.917) with blocks as covariables. The latter sum is the variance that remains after fitting blocks. Of this sum, 0.631 (the sum of the canonical eigenvalues) can be

explained by the treatments combinations N*P. The sum of squares for the residual is thus $0.917 - 0.631 = 0.286$.

Conceptually, the figures in Table 8.10 could be obtained by carrying out an ANOVA on each of the 103 observed species, totaling the sums of squares of each ANOVA-source across the 103 species, and dividing the resulting sums by the sum for the bottom row (Total).

Table 8.10 shows that the N.P interaction is not significant, whereas both main effects are significant. An ordination diagram based on main effects only would thus sufficiently describe the data. If N and P are entered as Environmental variables only, we obtain centroids for the N- and P- levels only. A plot of the sample scores, yields 8 distinct points, one for each treatment combination. Connecting these points to form arrows as in Figure 8-2, gives a diagram in which the arrows run parallel. A nice feature of the arrows in Figure 8-2 (with interaction) is that the arrows for lower N-levels each point approximately in the direction of the next higher N-level, suggesting that P and N can replace one another to some extent in this undergrowth. This finding contributes to the discussion of multiple versus single element limitation.

If there is reason to, CCA can be used instead of RDA in the above projects, resulting in an "analysis of inertia". Such an analysis would stress the relative abundances of species in each sample. The SS in the row "Total" is then unequal to one; it is the total inertia. Optionally, each SS can be expressed as a percentage of the total inertia.

We conclude the example with some more advanced issues. Because E40 is an orthogonal designed experiment, each sum of squares is unique. You may wish to decompose the variance as in section 8.3.1.2 to find out that the shared variance is zero. If some samples are deleted or weighted, the design is no longer orthogonal and the shared variance is no longer exactly zero. An unbalanced design does not necessitate another procedure, however. The resulting ANOVA table contains the pure effects only (cf Table 8.8), as desired. With CCA, the shared variance is not precisely equal to zero, even in an orthogonal experiment, because the implicit sample weights destroy the orthogonality somewhat. As in RDA of an unbalanced experiment, this imbalance does not necessitate another procedure to obtain the ANOVA table.

One might think that another way to obtain centroid points for the N*P treatment combinations in the main effects model is to enter the design file as a Supplementary Data file also and to use the interaction option of supplementary environmental variables to define the N*P product variables (project e40_nap.con). However, the points calculated in this way do not coincide exactly with those calculated above and the resulting arrows for N-levels would not run exactly parallel. The reason is that the centroids are derived from the species-derived sample scores, whereas the environment-derived sample scores are the scores that yield parallel arrows.

For users of the console version of CANOCO, the project files illustrate one additional point. In the projects, the block covariables are selected first, before the nutrient indicators. For example, in the project file e40_np.con, all covariables are required in the analysis (all covariables are selected: first all blocks, then all nutrients). One might think that one could simply answer "no delete/select". However, if this is done, as in e40_err.con, the block covariables cannot be selected anymore in the permutation test, as is visible at the end of e40_err.log. This error cannot occur in project files generated by Canoco for Windows.

8.3.3.2 E40: Design-based versus model-based permutation

This section contains a discussion of design-based versus model-based permutation, which is of interest to statisticians using CANOCO.

The permutation tests reported in Table 8.10 are model-based: there is no exact permutational argument for their validity. The level-accuracy of these partial tests hinges on the additive model and the usage of the F-ratio as (asymptotic pivotal) test statistic. In contrast,

design-based permutation methods have an exact permutational argument for their validity (Edgington, 1995). The permutation setup in CANOCO allows you to carry out design-based tests, if they exist. The crux is to use all covariables to define blocks. For example, in the design-based test of the N-effect on the undergrowth, the samples that are given different N-levels are permuted **within** each level of P and each block. This is achieved in CANOCO by selecting P0, P1, Blocks1, ..., Blocks 4 as block-defining covariables (project n_design.con). For example, in the resulting "Sample arrangement in the permutation test", the first block is reported to consist of the samples 1, 3, 6 and 7. These are all the samples from the first block in the experiment that received P (Block1, P1). The second block in the arrangement (samples 2, 4, 5, 8) consists of the remaining samples from the first block that did not receive P (i.e. Block1, P0). The resulting P-value changes little, if at all. The project p_design.con yields the design-based test for the effect of P.

According to Edgington (1995) there exists no design-based permutation test for interaction. However, in special cases an exact test of interaction is still possible (Welch, 1990) and the N.P interaction in our case is such a special case. How to obtain the Welch permutation test of interaction with CANOCO is postponed to section 8.3.13, as it requires a reordering of the present data and the whole-plot permutation options of CANOCO discussed later. To carry out this test with CANOCO, we must arrange the data within blocks in standard matrix order, e.g. NOP0, NOP1, N1P0, N1P1 etc. This is not the case for the data files in this directory.

In summary, design-based methods are exact, but have a limited domain of application, mainly designed experiments with one or two factors. Model-based permutation methods are almost exact and are more widely applicable, also in observational studies. If both methods are applicable, the question arises which method to choose. For large sample sizes there is little reason to prefer one over the other, because the model-based permutation tests in CANOCO have an extremely good level-accuracy. For small sample sizes, the number of distinct permutations may become very small with a design-based method. If there are fewer than 20 distinct values of the test statistic, the minimum achievable P-value is greater than 5%, so that effects, however large, cannot be shown to be significant. In this case, model-based methods are to be preferred as they have a good level-accuracy, even for small sample sizes.

8.3.4 Example LINES - Line transect(s) across the seashore

Problem:	Effect of environmental variables on plant composition along a transect
Data:	Cramer & Hytteborn (1987)
Directory:	\CANOCO\SAMPLES\PERMUTIO\LINES
Illustration of:	<ul style="list-style-type: none"> • How to test environmental effects in data from one or more line transects. • The minimum achievable P-value and the number of samples along a line transect. • How to select samples for analysis in Canoco for Windows.

Files	Name	Description
Species	plantspe.dta	abundance of 68 plant species in 63 sites along 4 transects on a rising seashore, sampled in 1978 and 1984 (126 samples, sea-shore only)
Environmental	line_env.dta	8 environmental variables and 4 transect indicators (238 samples, sea-shore+forest)
Project	line1.con	forward selection of environmental variables with cyclic permutation tests on the data from transect 1 in 1984 only (9 active samples)
	line1dis.con	as line1.con with mirror image disabled
	lines.con	forward selection of environmental variables with cyclic permutation tests on the data from all 4 transects in 1984 (63 active samples; transects are Blocks)

8.3.4.1 LINES: Permutation tests for line transect(s)

This example uses the species data of the Example 8.2.7 together with data on additional environmental variables. The effect of these variables on the seashore vegetation is examined by forward selection. The permutation tests used in each step of the selection account for the fact that the samples lie along line transects, while assuming that each pair of consecutive samples is equally correlated.

The project line1.con analyzes the 1984 data of transect 1 by deleting all other samples in the data files in the analysis. In Canoco for Windows, this is achieved most easily by first moving all samples from the "Source pool" box to the "To be deleted" box and then moving the 1984 samples of line transect 1 back to the "Source pool" box. These are the 23 samples from L1-03-84 to L1-45-84. Because only the first 9 samples up to the forest edge are in the species data file, the ordination (by CCA) is on 9 samples only. Table 8.11 shows the results of automatic forward selection with permutation tests. The permutation type is "restricted" for "line transects", resulting in cyclic random shifts along the transect. Because there are 9 samples, the number of different shifts is 2^9 ; the "2" arises because there are also 9 shifts from the mirror image of the transect. The minimum achievable P-value is thus $1/18 = 0.0556$. The P-value for the first, best variable, Elevation, is given as 0.054. The F-ratio for the elevation in the data is thus the maximum of all F-ratios that are calculated from the 18 different permuted data sets. This is no surprise, as Elevation is expected to be well correlated with the zonation across the sea shore. The marginal effect of the variable Inundation (ignoring Elevation) is about equally strong. The surprise is that Inundation has also a significant effect after accounting for Elevation ($P=0.056$). The additional effects of the other variables are not statistically significant.

Just for demonstration, one may wish to verify the minimum achievable P-value is only $1/9 = 0.111$ if the box "Disable random shifts of mirror image" is checked (project line1dis.con). There is little reason to do so in practice.

The full data are from four different line transects, placed in the data files one after the other. Of all samples, 9, 28, 20 and 6 are seashore samples that are in the species data file. To carry out the line permutation tests, CANOCO must know which transect each sample belongs to. This is achieved by specifying each transect as a Block, i.e. by entering the file line_env.dta also as the Covariable file, and by selecting the dummy variables Transct1 - Transct4 as block-defining covariables (project lines.con).

Table 8.11 Results of the automatic forward selection with line permutation tests for transect 1 (project line1.con) using 9999 permutations for each test.

Variable	Conditional Effects		P-value	F-value
	Var.Num.	Lambda-A		
Elevatio	1	0.69	0.054	3.80
Inundati	7	0.34	0.056	2.20
Loglight	2	0.34	0.116	2.77
Stones	3	0.19	0.226	1.90
Drift	5	0.14	0.330	1.61
Mosses	4	0.12	0.280	1.64
Litter	6	0.07	0.724	0.76

8.3.5 Example GRID - Samples in a rectangular spatial layout

Problem:	Effect of pH and water-depth on plant species in a wetland
Data:	Both & Van Wirdum (1981), Farjon & Wiertz (1988)
Directory:	\\CANOCO\SAMPLES\PERMUTIO\GRID
Illustration of:	<ul style="list-style-type: none"> • How to obtain a permutation test that accounts for spatial autocorrelation when sampling points are on a grid. • How to check whether the grid dimensions are 5*4 or 4*5. • How to specify a permutation of plots (i.e. pairs of samples). • Using a CANOCO.INI file to change an advanced option.

Files	Name	Description
Species	grid_spe.dta	100 plant species in 20 plots arranged in a 5*4 grid, sampled in 1977 and 1988
Environmental	grid_env.dta	plot, strip and year indicators, pH and water depth
Project	grid88.con	Toroidal permutation test of the pH effect on the samples from 1988 only (5*4 grid)
	grid_err.con	as above but with wrong grid dimensions (4*5)
	grid7788.con	Toroidal permutation test of paired samples to test the pH effect on all samples (5*4 grid)
	whol_err.con	Wrong attempt to test the pH effect, namely by permutation within sample pairs

8.3.5.1 GRID: Samples in a rectangular spatial layout

This section uses the same data as the Example VEGCHANG in section 8.2.10. The layout of the data is given on page 191 of Unimodal models. The position of each sample is clear from its name (letter with digit, e.g. B2 for row B column 2). For illustration, we wish to test the relation between pH and the vegetation in 1988. Because the samples lie on a rectangular grid, there is a danger that autocorrelation between samples makes the test too liberal if random permutations are used. CANOCO allows you to account for autocorrelation by carrying out a permutation type that has been designed for rectangular grids. This is achieved by selecting "Restricted permutation" for data on a "Rectangular spatial grid" (project grid88.con in which all 1977 samples are deleted). We are then asked to specify the dimensions of the grid. From the field layout, it is known that the grid is 5 rows by 4 columns or 4 rows by 5 columns, but it depends on the order of the samples in the data which one is correct. Let's be lazy and select 5 rows by 4 columns. After the permutation test has been carried out, the sample arrangement is reported in the output (log-window) as being:

```

Row    1 consists of the plots:
  21    22    23    24
Row    2 consists of the plots:
  25    26    27    28
Row    3 consists of the plots:
  29    30    31    32
Row    4 consists of the plots:
  33    34    35    36
Row    5 consists of the plots:
  37    38    39    40

```

The 20 plots are permuted using toroidal shifts

CANOCO reports here sample identification numbers in the data file. To see the corresponding sample names, either inspect the solution file or switch back to the Project View, Options, Delete samples. By either method, you can verify that the samples of the first row (21-24) are A1_88 to A4_88, the samples of the second row (25-28) are B1_88 to B4_88, and so on. All seems correct. If we had chosen 4 rows * 5 columns, the first row was reported to consist of the samples 21 to 25, which are A1_88 to B1_88, which is clearly wrong. Instead of being lazy, you may want to make the correct choice from the start. For this, click the Help button (or read Q58 of the console version of CANOCO) and inspect the order of the samples in the data file by returning to the wizard page on Deleting samples.

As judged by the P-value obtained with project grid88.con, the vegetation is significantly related to pH ($P = 0.03$), also after accounting for spatial autocorrelation. It should be added here that any variable that changes gradually across the field is judged significant by this method. In the example data, such a variable is Waterdepth in 1988. Section 3.7.3, subsection Rectangular Grids, warns against taking spatial trends for environmental effects. Although other explanations cannot be excluded on statistical grounds, the significance is interpreted here that pH and/or Water depth causes the spatial trend in the vegetation.

On page 195 of Unimodal models, both the 1977 and the 1988 data are used to test the component "pH and water depth", while taking into account the spatial and temporal structure of the data. It is proposed to account for the autocorrelation

- in time by permuting plots instead of individual samples: paired samples are permuted together.
- in space, by wrapping the plots around a torus and then applying a toroidal shift.

In CANOCO 4.5, this can be achieved by "restricted permutation" using the "split-plot design" option. The "split-plots" are the individual samples. The "whole-plots" are the plots from which the individual samples are taken. The number of split-plots per whole-plot is thus 2. In the data file, the samples are arranged per year, first all 20 samples from 1977, then all 20 samples from 1988. The rule to find the two samples of the plot A1 is thus "take 1, skip 19". With this rule all pairs of samples can be found. The whole-plots lie on a spatial grid of 5 rows by 4 columns as discussed above. The start and end of the report on the sample arrangement of the permutation test so specified reads like

```

Row      1 consists of the whole plots:
Whole plot  1 :
  1    21
Whole plot  2 :
  2    22
Whole plot  3 :
  3    23
Whole plot  4 :
  4    24
Row      2 consists of the whole plots:

```

....

```

Row      5 consists of the whole plots:
Whole plot 17 :
 17    37
Whole plot 18 :
 18    38
Whole plot 19 :
 19    39
Whole plot 20 :
 20    40

```

```

These    20 whole plots are permuted using toroidal shifts
The      2 split plots are not permuted

```

By inspecting the sample names, you may wish to verify that this arrangement is as intended. The resulting P-value is again ca. 0.025. This is an exact Monte Carlo test by design (there are no covariables; it is not a partial test; cf section 8.3.3.2).

To test a spatial component such as pH in these data, plots must be permuted. It is perhaps instructive to see what happens if split plots are permuted and whole-plots are not (project whol_err.con). In all the permutations so obtained the *F*-ratio is the same as in the data (8.717), as can be seen in the file whol_err.scr, obtained with the console version of CANOCO); the resulting P-value is 1.0, i.e. no significance is found.

We close this example with some remarks on the number of possible permutations. On page 195 of Unimodal Models, where analyses of the same data are reported, the number is calculated to be 80. This number can be reached only if CANOCO is allowed to generate toroidal shifts from each of the four grids displayed in section 3.7.3. For this, option (21) in the initialization file must be changed from 0 to 1.

This has been done in the CANOCO.INI file in the directory of this example. The change is appropriate if the bivariate autocovariance function is symmetric (see the "Sample arrangement in the permutation test").

8.3.6 Example SPLITPLT - A split-plot analysis

Problem:	Ectomycorrhizal fungi occurrence as affected by the manipulation of litter and humus in Scots pine stands of different age.
Data:	Baar & Ter Braak (1996)
Directory:	\CANOCO\SAMPLES\PERMUTIO\SPLITPLT
Illustration of:	<ul style="list-style-type: none"> • What a nested design or split-plot design is. • What a whole-plot variable is and what a split-plot variable is. • The requirements to be able to perform permutation tests in a split-plot design. • How to test for whole-plot variables. • Two ways to test for split-plot variables. • How to test the interaction between a split-plot variable and a whole-plot variable.

Files	Name	Description
Species	fungi.dta	numbers of sporocarps of 33 ectomycorrhizal fungi in 64 plots in 6 Scots pine stands
	fungi2.dta	as fungi.dta with two extra variables (total numbers of sporocarps and of species in each sample). Added as supplementary species in fig1.con
Environmental	design.dta	6 Stand and 3 treatment (S,A,C) indicators, Age and Soil type (Podzol, Arenosol) of the stands
	nutrient.dta	nutrient concentrations and pH in the ectorganic layers (log-transformed). Used in fig1.con as supplementary environmental file
Derived Project	treatmnt.dta	as design.dta but without the stand indicators
	stand1.con	project to show that age and soil are constant within stands (=whole-plot)
	stand2.con	idem by regression of Age and Soil on Stand
	age_soil.con	test of the relation between the fungi and the two stand (=whole-plot) variables Age and Soil type in a balanced subset of 6*8 = 48 plots (all A-plots deleted)
	age.con	idem, but now for Age adjusted for Soil type
	lh.con	test of the effects of the treatment of the litter and humus layer (LH, a split-plot factor) on the fungi on all data, using block permutation within stands
	lh_split.con	as above but on the balanced subset of the data, and using "within whole-plot" permutation of split-plots
	lh_x_age.con	test of the interaction of the LH treatment and age on all data
	agetrial.con	attempt to test Age on all data (wrong P-value)
	agsl_err.con	wrong test of Age and Soil (whole-plots not permuted)
fig1.con	project upon which fig.1 of the paper is based	

8.3.6.1 SPLITPLT: The experimental design and data: split-plot or nested design

In this example we study the effect of sod-cutting and sod-addition on the number of sporocarps of ectomycorrhizal fungi in Pine stands differing in age and soil type using data from an experiment reported in Baar & Ter Braak (1996). The experimental layout is as follows. Six stands (St1 - St6) with Scots pine of different age and soil type were selected. Within each stand, plots were laid out, with size of 15 m x 15 m. The treatments applied to plots were:

- S: Sod-cutting in which the litter and humus layers and the herbaceous vegetation were removed
- A: Sod-addition in which removed litter and humus were added on to existing litter and humus layers
- C: Control in which the litter and humus layers were left unchanged

Treatment A (Sod-addition) was not applied in the two oldest stands (St5, St6) as their litter and humus layers were already quite thick. Each treatment was replicated four times, giving $4 \times 3 = 12$ plots in the younger stands St1-St4 and $4 \times 2 = 8$ plots in the oldest stands (St5, St6). Within each stand the treatments were completely randomized. The treatments S, A and C are collectively indicated as the LH-treatment (litter and humus).

The variables in the file design.dta are given in Table 8.12.

Table 8.12 The variables in the file design.dta.

variable	explanation:	Age (year)	Soil type
St1_Aren	stand 1	3	Arenosol
St2_Podz	stand 2	10	Podzol
St3_Podz	stand 3	16	Podzol
St4_Aren	stand 4	27	Arenosol
St5_Aren	stand 5	50	Arenosol
St6_Aren	stand 6	66	Arenosol
S	Sod-Cut plot		
A	Sod-Added plot		
C	Control plot		
Age	log(Age of stand)		
Podzol	Podzol soil		
Areno	Arenosol soil		

The experimental layout is a split-plot or nested design: plots are selected within selected stands. In the terminology of the split-plot design, stands are called “whole-plots” and plots within stands “split-plots”. The split-plots are the samples in the data files. With the projects stand1.con and stand2.con or by inspecting the design data file, it can be verified that the variables Age and Soil type vary only between stands. They could therefore be called “whole-plot factors” or “whole-plot variables” because age is a quantitative variable.

The species data (in the file fungi.dta) are the number of sporocarps (in 1993) of 40 ectomycorrhizal fungal species of which 7 turned out be absent in all plots. In addition, nutrient concentrations and acidity were determined in the ectorganic layers (humus in C and A, litter in S) with the data in the file nutrient.env (log-transformed except pH).

Because the lengths of gradient as determined by DCA on log-transformed numbers of sporocarps are small (<4 SD) and because we are interested in absolute amounts, we analyze the data by redundancy analysis (RDA).

We close this subsection with some information on the projects Stand1.con and Stand2.con. Stand1.con is a project in which the 6 Stand indicators, Age and Soil type are the only environmental variables. In the log-window it can be seen that Age and Soil type are both found to be collinear with the Stand indicator variables; both have an inflation factor of infinity

(reported by CANOCO as being 0). Stand2.con is a project in which Age and Soil type are regressed on the Stand indicators by using RDA. The sum of all canonical eigenvalues is 1, meaning that the stand indicators explain all the variation in Age and Soil type. Note that in the latter project, Age and Soil are specified as Species Data. For this to be possible, the file treatment.dta is used instead of design.dta, because CANOCO 4.0 does not allow you to specify the same file as Species Data and as Environmental or Covariable Data.

➤ To treat one subset of variables in a data file as response variables and another as explanatory variables, copy the file and specify one copy as Species file and the other as Environmental or Covariable file. Use the Delete boxes to retain the correct variables in each subset.

8.3.6.2 SPLITPLT: Testing the effect of whole-plot variables in a split-plot design or a nested design

The split-plot design options of CANOCO allow you to determine the significance of whole-plot factors or variables, but only if whole-plots have equal numbers of split-plots (samples). This is not the case in the example data. The design can be made balanced either by deleting all samples from Stands 5 and 6 or by deleting all sod-added plots. We choose to delete the sod-added plots because we are testing whole-plot factors here and thus want to retain as many stands as possible. After deletion, there are 6 whole-plots with 8 split-plots per whole-plot.

To determine the significance of the stand variables Age and Soil Type, select “Restricted” permutation (without blocks) and “Split-plot design”, specify the number of split-plots per whole-plot (8) and accept the default “take and skip” rule because the samples of the same stand are consecutive in the data file (project age_soil.con). In the wizard-page “Split-plot design II”, specify that whole-plots are freely exchangeable and that split-plots do not have to be permuted. The resulting P-value is ca. 0.09.

It may be instructive to see what happens if whole-plots are not permuted and split-plots are (project agsl_err.con). CANOCO does not complain. The resulting P-value is 1.0000, because the F-ratio of each permuted data set is equal to the observed F-ratio (See the file agsl_err.scr, which is the screen file obtained from the console version of CANOCO).

The effect of age adjusted for soil type can be tested by specifying soil type as a covariable (the project age.con). **Attention:** in CANOCO 4.5, such partial tests of whole-plot factors may be too liberal, i.e. the real P-value may be higher (Anderson & Ter Braak, 2002). This happens if both the environmental data and the covariable data are constant within whole-plots (see section 3.7.6).

With restricted permutation types, Canoco for Windows does not always allow you to specify the design, if the sample numbers are not consecutive or if you have deleted some samples. The console version of CANOCO is more tolerant here. Fortunately, this is not the problem in the example data.

It might be instructive to see what happens if we attempt to test Age using all data (project agetrial.con). We again specify “Restricted” permutation and “Split-plot design”, but what about the number of split-plots per whole-plot? Note that there are 64 samples in the data files. We cannot choose 12 (4×3), because $64/12$ is not a whole number. We can choose 8 ($64/8$) because the result is a whole number. Canoco for Windows accepts 8 and runs, but from the sample arrangement listed in the log-window we see that CANOCO used 8 whole-plots, whereas there are only 6 stands. The lesson is to always check the sample arrangement in the log-window.

☞ If Canoco for Windows does not accept your specification of the design, check whether the sample numbers are consecutive in the data files. If not, renumber them or use the console version of CANOCO.

8.3.6.3 SPLITPLT: Testing the effect of split-plot variables

Testing the effect of split-plot variables is best carried out using unrestricted permutations within blocks. For this, the whole-plots must be specified as covariables and as blocks.

The project `lh.con` gives an example in which the significance of the treatment of litter and humus is tested. Note that it is no problem that the number of samples differs between blocks: 12 in blocks 1-4 (Stand 1-4) and 8 in blocks 5-6 (Stand 5-6). Similarly, in the project `lh_x_age.con`, it is tested whether the effect of the treatment of litter and humus differs among stands of different age (adjusted for possible LH.soil interaction). The interaction is judged significant.

For illustration only, the project `lh_split.con` shows, by using the data file `treatmnt.dta`, how the LH effect can be tested without using any stand indicator variables: split-plots are permuted, whereas whole-plots are not.

We close this example with the project `fig1.con` used to produce Figure 1 of Baar & Ter Braak (1995). The number of species and the number of sporocarps per sample are defined as supplementary species. The nutrient concentrations in the soil are used as supplementary environmental variables.

8.3.7 Example BAC1SPE - A univariate, unreplicated BACI analysis

Problem:	Does spraying with deltamethrin affect <i>Oedothorax apicatus</i> females?
Data:	Everts (1990), van der Voet (1987), Everts et al. (1989)
Directory:	\CANOCO\SAMPLES\PERMUTIO\BAC1SPE
Illustration of:	<ul style="list-style-type: none"> • The univariate analysis of a BACI design by permutation methods. • Univariate randomized intervention analysis (Carpenter et al. 1989). • How to account for autocorrelation in time in an unreplicated BACI experiment. • Exploiting the facilities of split-plot permutations in CANOCO. • Dependent permutations. • Specifying a permutation file.

Files	Name	Description
Species	oedo_api.dta	counts of the spider <i>Oedothorax apicatus</i> from a control site and an impact site, sampled 22 weeks
Environmental	toxicant.dta	indicator whether spraying has occurred at a site, and a putative impact size proxy (9 minus time-since-spraying)
Covariables	timesite.dta	time and site indicators
Other	permutio.dta	example of a permutation file
Project	baci1.con	permutation test of “no impact” against a “constant impact” model using random permutation of differences
	baci2.con	permutation test of “no impact” against a “constant impact” model using dependent time shifts
	baci3.con	permutation test of “no impact” against an “instantaneous effect with extinction” model using dependent time shifts
	baci4.con	as baci3.con with a general impact model
	permfile.con	example project that uses a permutation file

8.3.7.1 BAC1SPE: Testing a putative impact on a single species in an unreplicated BACI experiment

This is an ecotoxicological example. There are two sites which are each sampled in 22 consecutive weeks. One site is sprayed with deltamethrin after week 13 (Impact site). The other site acts as the Control site. In each week the number of females of the species *Oedothorax apicatus* at each site is counted.

Stewart-Oaten et al. (1986) proposed to judge the effect of the spraying on the basis of the differences in the log-counts, calculated at each time, between the impacted and the control site. From these differences, $\{d_t\}[t = 1 \dots n]$ say, the significance of the effect can then be determined by a two-sample t-test testing the mean difference between Before and After Impact times. Carpenter et al. (1989) extended their proposal by using a permutation test. These authors also examined the effect of autocorrelation on the test. In this example, it is shown how this permutation test can be obtained with CANOCO (project baci1.con) and how it can be extended to account for autocorrelation (projects baci2.con - baci4.con). The input data for CANOCO can be the original counts at the two sites. This has the advantage that the proposed analysis can easily be generalized to multi-species responses and multi-site situations.

The BACI model for the original counts is (cf Unimodal Models, page 205)

$$(8.1) \quad y = \text{site} + \text{time} + \text{impact} + \text{error}$$

with $y = \log(\text{count}+1)$. See Stewart-Oaten et al. (1986) for the assumptions made in this model and Underwood (1992, 1994) for extensions. The interest focuses on the impact. This is achieved by specifying the site and time indicators (file timesite.dta) as Covariable Data and the spraying-treatment (file toxicant.dta) as Environmental data (project baci1.con). The t-value of the regression coefficient (tVal: in the solution file) is precisely the t-ratio used in the Stewart-Oaten et al. t-test.

The randomized intervention test of Carpenter et al. (1989) consists of permuting the differences $\{d_i\}$. These permutations are obtained by asking for a restricted permutation test using a split-plot design. In this test, the sites are whole-plots which are not permuted and the samples are split-plots that are freely exchangeable, except that the samples taken at the same time travel together. This is achieved by asking for dependent permutations across whole-plots. From a randomization point of view it also makes sense to freely exchange the sites, as - ideally - the spraying treatment is randomly assigned to one of the two sites⁵. This means that also the values $\{-d_i\}$ are being permuted among themselves. Because the test statistic in CANOCO is unsigned, this has no effect in this case, as you may wish to check empirically. In the project baci1.con, we freely exchange both sites and times and we checked the "Dependent across whole-plots" box.

Carpenter et al (1989) point out that the differences may be autocorrelated in time. CANOCO allows you to account for possible autocorrelation by permutations for "time series". If these are made "dependent across whole-plots", the test is based on cyclic shifts of the differences (project baci2.con). The dependent time-shift permutation test circumvents the need for detailed time-series modeling (van der Voet, 1987) at the expense of some power of the test. Moreover, the test is easily extended to the multi-species case as we show in the next subsection.

There is not much difference between the above two permutation approaches for these data - both yield a P-value of ca 0.02 - presumably because the autocorrelation of the differences is low (at lag 1, $r = -0.23$ and at lag 2, $r = -0.39$; Van der Voet, 1987). Note that the autocorrelations in the original counts are much higher (at lag one, $r = 0.8$).

These projects test the null hypothesis of no impact against the alternative model of an instantaneous impact that is constant after the spraying. In project baci3.con, the alternative model is made more realistic: the impact has an instantaneous effect that dies out linearly on the log-scale, i.e. exponentially in terms of the original counts. The variable "Sprayed" accounts for the instantaneous effect, the variable "Isize" for the extinction of the impact effect. The virtue of this linear model is that it is parsimonious and much more flexible than the constant model.

In project baci4.con, the alternative model is left completely free with as many impact parameters as there are After-Impact times. This is achieved by defining products of the Impact site variable and the After-Impact times. The test in project baci4.con has less power than that in baci3.con, if the impact changes smoothly over time. For the example data, $P = 0.12$. The test statistic in project baci4.con is precisely the F-statistic for the site.time interaction in an analysis of variance; only the method to determine its significance level is nonparametric in CANOCO and, in project baci4.con, accounts for autocorrelation.

The file permutio.dta enumerates all 88 different permutations of the dependent time-shift permutation test in which also the whole-plots are permuted. This file is used in project permfile.con as an example of the option "Read from file" in the Permutation Type wizard page.

⁵ Even if not randomized, the sites should be exchangeable under the model of proportional population change.

Note that the number of permutations in the file (88) must be specified in the previous wizard page.

8.3.8 Example BACIMSPE - A multivariate, unreplicated BACI analysis

Problem:	Testing the effect of a toxicant on a species assemblage (BACI design)
Data:	R.P.A. Van Wijngaarden (unpubl.)
Directory:	\CANOCO\SAMPLES\PERMUTIO\BACIMSPE
Illustration of:	<ul style="list-style-type: none"> • The multivariate analysis of a BACI design by permutation methods. • Multivariate randomized intervention analysis.

Files	Name	Description
Species	species.dta	57 species from a control pond and an impact pond, each sampled for 10 consecutive months
Environmental	toxicant.dta	indicator of the impact (a toxicant) after month six at the impact pond
Covariables	pondtime.dta	pond and time indicators
Project	baci_rda.con	permutation test of “no impact” against a “constant impact” model using dependent time shifts
	logratio.con	as above but using a log-ratios per sample
	baci_cca.con	as above but using CCA instead of RDA

8.3.8.1 BACIMSPE: Testing a putative impact on a species assemblage in a unreplicated BACI experiment

As the previous example, this example also has an unreplicated BACI design. The important difference is that the response is not univariate, namely the abundance of a single species, but multivariate, namely the abundances of an assemblage of 57 species. For each species we assume the model of equation (8.1) with possibly different site, time and impact parameters for each species. Because the pseudo-F statistic used in CANOCO simply adds sums of squares across species, the analysis proceeds precisely as in the previous section. The project `baci_rda.con` uses the same permutation type as the project `baci2.con` of the previous section. The only difference is that the species file contains 57 species instead of one.

With an assemblage of species as response, there is the additional possibility to focus the impact assessment on the ratios of species within each sample, i.e. to carry out a log-ratio analysis (project `logratio.con`). This is attractive if one has already done an separate univariate BACI analysis on the total abundance, or if the total abundance in a sample is defined by the sampling method. The only difference of the log-ratio analysis with that in `baci_rda.con` is that centering by samples is specified. The model for the analysis can be written as

$$(8.2) \quad y_{itk} = c_{it} + \text{site}_{ik} + \text{time}_{tk} + \text{impact}_{tk} + \text{error}$$

with $y_{itk} = \log(\text{count})$ or $\log(\text{count}+1)$ of species k at time t at site i . This equation excludes one component of the site \times time interaction (namely c_{it}) from the impact assessment. This component is related to the abundance total in each sample. The subscripts attached to the impact term in equation (8.2) indicate the general impact model. In the project `logratio.con`, however, all impact terms are taken as equal in time ($\text{impact}_{tk} = \text{constant impact}_k$), by having

just one impact indicator as Environmental variable in the analysis. As in the previous section, this “constant impact” model can be made more flexible by adding more explanatory variables.

Instead of a log-ratio analysis one may perhaps wish to switch to a CCA (project baci_cca.con). The CCA is then best carried out on log-transformed data so as to ensure that the model continues to be multiplicative in terms of the original counts.

8.3.9 Example BACI3SIT - A multivariate BACI analysis with three sites

Problem:	Testing the effect of a toxicant on a species assemblage in a three-site BACI design
Data:	R.P.A. van Wijngaarden (unpubl.)
Directory:	\CANOCO\SAMPLES\PERMUTIO\BACI3SIT
Illustration of:	<ul style="list-style-type: none"> • The multivariate analysis of a three-site BACI design by permutation methods. • How to interpret the fit-diagnostics of species in the solution file. • Details of the Take and Skip rule when some samples are deleted.

Files	Name	Description
Species	species.dta	59 species in three ponds sampled in consecutive months
Environmental	toxicant.dta	treatment indicators of the pond (Control, Low and High dose) and coded as 0,1,2 in the variable Toxlevel
Covariables	pondtime.dta	pond and time indicators
Project	bacirda1.con	permutation test of “no impact” against a “level of impact” model (linear in Toxlevel using 1 degree of freedom)
	bacirda2.con	permutation test of “no impact” against an “impact per dosage class” model (2 degrees of freedom)
	bacirda3.con	as bacirda1.con but after deleting a species that had extreme influence on the analysis (species 8, clo dip)
	bac_i_c_h.con	re-analysis of Example in 8.3.8.1 by deleting the data from the Low-impact pond

8.3.9.1 BACI3SIT: Testing a putative impact on a species assemblage in a BACI experiment with three sites

With two sites (a Control site and an Impact site), the permutation of sites is unimportant. The statistical test necessarily focuses on the permutations of the time points, even if sites are permuted. If the impact assessment can be based on data from more than two sites, the permutation of sites (whole-plots) becomes gradually more important. The permutation of the time points is still needed if there are few more sites. In observational studies there may be one Impact site and a small (2-4) number of Control sites (Underwood, 1992, 1994). If there are a few more Impact sites, they may lie along an impact gradient from putatively strong to low impact. The ecotoxicological example that we analyze here is of this form. There are three sites (ponds): one Control pond, one pond with a low dose of the toxicant and one pond with a high dosage. In bacirda1.con, “no impact” is tested against a “level of impact” model which is proportional to the putative impact: 0 for the Control, 1 for the low dosage pond and 2 for the high dosage pond. In bacirda2.con, the impact model is a little bit less restrictive in that the impact level is coded by classes (Control, Low, High).

As in the previous section, the RDA analysis can be easily modified to a log-ratio analysis by “centering by samples”.

A check of the fit-diagnostics in the solution file indicates that there is one species with an extremely high variance in the data: the species “clo dip” (number 8) has a variance of 14.22 whereas most species have a variance lower than 1. The percentage fit of the species is not

extreme (%EXPL = 13.55 in bacirda1.sol). Because the implicit weight of a species in the analysis is equal to the product of the last two columns (headed VAR(y) and %EXPL), it is of interest to see whether the impact is still significant if this species is deleted from the analysis (project bacirda3.con).

The data of Example 8.3.8 are a subset of the current data, in that the data of the Low dosage pond have been deleted. The project baci_c_h.con shows how the analysis of the previous section (in BACIMSPE\baci_rda.con) can be obtained by using the file with the full data and deleting the samples from the low-impact pond from the analysis. Note that the “Take and Skip” rule is Take 1 and Skip 1 (as in the previous section) instead of being Take 1 and Skip 2 (as in this section), because the “Take and Skip” rule is applied to the sample sequence as given in the solution file rather than to the sequence in the data file. Samples that are deleted are also deleted from that sequence, and samples that are made supplementary, are placed at the end of the sequence.

8.3.10 Example BACI_REP - A multivariate, replicated BACI analysis

Problem:	Testing the effect of liming on nematodes in a replicated BACI experiment
Data:	Manger & Schouten (1989)
Directory:	\CANOCO\SAMPLES\PERMUTIO\BACI_REP
Illustration of:	<ul style="list-style-type: none"> • The analysis of a replicated BACI experiment carried out in blocks. • Using the split-plot options in combination with blocks.

Files	Name	Description
Species	nematode.dta	4 food groups of nematodes in 72 samples (3 forests * 6 plots * 4 sampling times)
Environmental	treatmnt.dta	3 treatment classes for no liming, 3 and 9 ton/ha (L0, L3 and L9) and lime quantity (lime), as applied after the first sampling time
Covariable	plotime.dta	3 forests (D,E,G), 4 time (0-3) and 18 plot indicators (d1-d6, e1-e6, g1-g6)
Project	baci_rda.con	Test of the liming effect by a BACI analysis using permutation of plots within forests
	logratio.con	as above, using log-ratio analysis

8.3.10.1 BACI_REP: Testing a putative impact in a replicated BACI experiment

If the impact assessment can be based on data from many sites, the permutation of sites (whole-plots) becomes the essential part of the analysis, and even with just one Before and one After time, valid tests of the impact can be obtained. In the example that we analyze here, there are 18 sites (arranged in three blocks) and each site is recorded four times. The test on the impact effect uses permutation of sites only. Optionally the time points could be permuted also, but this will provide little additional information in the example and, with many sites, has a less secure basis than the permutation (randomization) of sites.

The example is a liming experiment carried out in three forests (labeled D, E and G). In each forest, there are six plots, recorded one time before and three times after the treatments were applied. Recorded are the abundances of nematodes in four food-groups. The treatment is the application of three doses of lime: 0, 3 and 9 ton/ha lime. The code names of the samples are best explained by an example. The sample Dt2L3p4 stands for the 4th plot (pl) in forest D that received 3 ton/ha lime(L) and was sampled at time (t) 2.

The project `baci_rda.con` shows that the forest, time and plot indicators are covariables. A "Restricted permutation" for "Split-plot designs" is requested with "Blocks" defined by the three forest indicators. The number of split-plots is 4 (for the four samples of each plot). The samples of each plot are consecutive in the data file, so that the default "Take and Skip" rule can be used. In the console version of CANOCO we also must specify that this layout holds true within each block. The resulting "Sample arrangement in the permutation test" in the log-window clearly indicates both the blocks, the whole-plots per block and the samples that form each whole-plot. The resulting P-value is 0.62.

The test does not reveal a liming effect, or, said otherwise, the strong time and plot effects in the data are not mistaken for being an effect of the liming.

8.3.11 Example PRC_SIM - Displaying time-dependent effects by PRC

Problem:	How do treatment effects change over time and how to display them?
Data:	Van den Brink & Ter Braak (1999: simulated data of appendix II)
Directory:	\CANOCO\SAMPLES\PERMUTIO\PRC_SIM
Illustration of:	<ul style="list-style-type: none"> • How to obtain the PRC diagram from the CANOCO output. • How interpret the PRC diagram quantitatively.

Files	Name	Description
Species	species.dta	Counts of six species (S1 -S6) in 5 cosms sampled 4 times each
Environmental	design.dta	Treatment (Control, Low, High) and time (Wk0-4) indicators. There are two Control cosms, one Low dosage cosm and two High dosage cosms
Derived	tm_x_tr.dta	4*3 (time * treatment) indicators (CW0, LW0, ..., HW3)
	prc_sim.xls	Excel workbook to calculate and display the first PRC
	pcr_sim.ppt	Powerpoint 4.0 file with PRC diagram (based on prc_sim.xls)
Project	prc.con	PRC analysis of Table 4 of the paper with the product variables of time and treatment created in the project using design.dta as both Environmental and Covariable data
	prc1.con	as above with product variables ordered per treatment
	prc_alt.con	as above with deletion of the pre-treatment variables (LW0, HW0)
	prc_x.con	PRC analysis of Table 4 of the paper using the file tm_x_tr.dta as Environmental file and design.dta as Covariable file

Table 8.13 Data tables used as input files for the PRC analysis in CANOCO: species.dta (columns S1-S6) and design.dta (columns C, L H and W0 - W3).

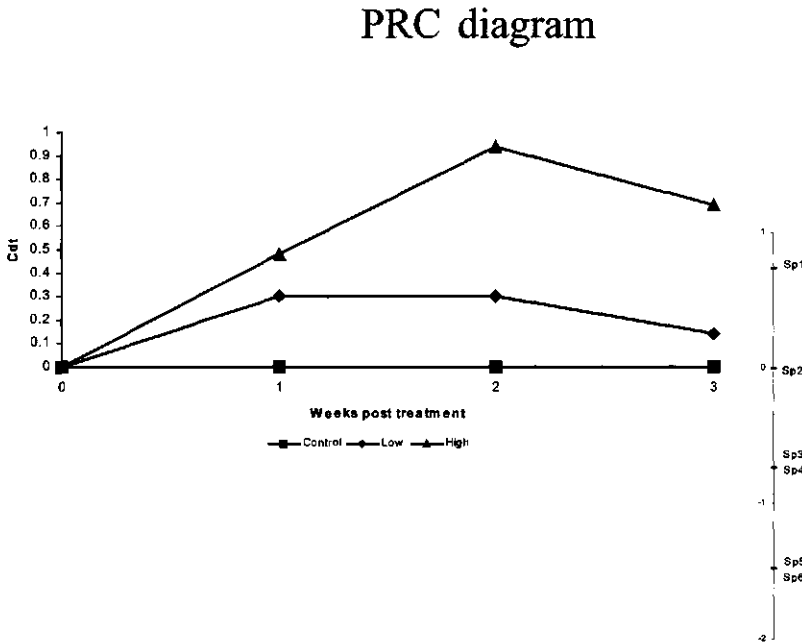
Sample	S1	S2	S3	S4	S5	S6	C	L	H	W0	W1	W2	W3
C1-W0	100	100	100	100	200	100	1	0	0	1	0	0	0
C2-W0	100	100	100	100	200	100	1	0	0	1	0	0	0
L-W0	100	100	100	100	200	100	0	1	0	1	0	0	0
H1-W0	100	100	100	100	200	100	0	0	1	1	0	0	0
H2-W0	100	100	100	100	200	100	0	0	1	1	0	0	0
C1-W1	110	90	120	90	240	130	1	0	0	0	1	0	0
C2-W1	110	90	120	90	240	130	1	0	0	0	1	0	0
L-W1	138	90	96	72	154	83	0	1	0	0	1	0	0
H1-W1	157	90	84	63	118	64	0	0	1	0	1	0	0
H2-W1	157	90	84	63	118	64	0	0	1	0	1	0	0
C1-W2	120	80	140	80	240	70	1	0	0	0	0	1	0
C2-W2	120	80	140	80	240	70	1	0	0	0	0	1	0
L-W2	150	80	112	64	154	45	0	1	0	0	0	1	0
H1-W2	240	80	70	40	60	18	0	0	1	0	0	1	0
H2-W2	240	80	70	40	60	18	0	0	1	0	0	1	0
C1-W3	130	100	160	70	200	100	1	0	0	0	0	0	1
C2-W3	130	100	160	70	200	100	1	0	0	0	0	0	1
L-W3	144	100	144	63	162	81	0	1	0	0	0	0	1
H1-W3	217	100	96	42	72	36	0	0	1	0	0	0	1
H2-W3	217	100	96	42	72	36	0	0	1	0	0	0	1

Table 8.14 Output of CANOCO for obtaining the PRC's.

Standardized canonical coefficients (Regr:AX1), standard deviations of environmental variables (sd_env) and species scores (Spec:Ax1). The treatment scores of the first PRC diagram (PRC1) is obtained as $(TAU * \text{Regr:AX1}) / \text{sd_env}$ with $TAU = .364917$, the total standard deviation in the species data.

N	Name	Regr:AX1	sd_env	PRC1	N	Species	Spec:AX1
8	Wk0*Low	0.0000	0.2179	0.0000	1	Sp1	0.7385
9	Wk0*High	0.0000	0.3000	0.0000	2	Sp2	0.0000
10	Wk1*Low	0.1805	0.2179	0.3022	3	Sp3	-0.7385
11	Wk1*High	0.3970	0.3000	0.4829	4	Sp4	-0.7385
12	Wk2*Low	0.1805	0.2179	0.3022	5	Sp5	-1.4771
13	Wk2*High	0.7716	0.3000	0.9385	6	Sp6	-1.4771
14	Wk3*Low	0.0852	0.2179	0.1426			
15	Wk3*High	0.5686	0.3000	0.6916			

Figure 8-3 PRC diagram of the simulated data.



8.3.11.1 Principal Response Curves analysis (PRC)

The Principal Response Curves analysis, a novel multivariate method for the analysis of repeated measurement designs, is designed to test and display treatment effects that change across time. The method is based on a reduced rank regression that is adjusted for changes across time in the control treatment. This allows the method to focus on the time-dependent treatment effects. The principal component thereof is plotted against time in the PRC diagram. The method has been developed for the analysis of ecotoxicological studies, an example of

which is given below, but may also prove useful in other disciplines. The theory and further examples are given in Van den Brink & Ter Braak (1997, 1998, 1999).

8.3.11.2 PRC_SIM: How to obtain the PRC curves

We use a small simulated cosm experiment (cosm = microcosm/mesocosm = experimental ecosystem) to show how to obtain the PRC diagram with CANOCO. The example cosm experiment consists of 3 treatments: Control (C), Low dosage (L) and High dosage (H). The treatments C and H are applied to two cosms each, whereas L is applied to a single cosm. All cosms are sampled 4 times ($t = 1, \dots, 4$) indicated by W0, W1, W2 and W3 for week 0, 1, 2 and 3. The treatment dosages are applied just after the sampling in week 0. Table 8.13 shows artificial, noise-free count data for 6 species (the columns labelled S1 through S6) in all 5×4 combinations of cosms and weeks, which are the rows of Table 8.13. The file with count data, species.dta, are entered as Species data in CANOCO. The remaining columns of Table 8.13 are indicator variables for treatment and sampling week. These data are in the file design.dta, which is entered as both Environmental and Covariable data.

In the project prc.con, the RDA option is chosen and the count data are log-transformed. All environmental variables are deleted but $2 \times 4 = 8$ interaction terms are added, namely, all products of the variables Low and High and the variables W0 - W3. The week variables are retained as covariables by deleting the treatment variables Control, Low and High.

If the product variables are available as a data file, this file can be entered as Environmental data, instead of defining the product variables in the project. All 3×4 product variables are in the file tm_x_tr.dta. So we need to delete in CANOCO the products involving the Control, when tm_x_tr.dta file is used as Environmental Data File (project prc_x.con). The results of the projects prc.con and prc_x.con can be seen to be identical.

The required output items of CANOCO for the formation of the PRC diagram are the species scores, the regression/canonical coefficients for standardized environmental variables (both in the solution file) and the total standard deviation of the species data (TAU) and the standard deviations of environmental variables (both in the log-window). The output for the example data is listed in Table 8.14 together with the formula and result for the treatment scores of the first PRC diagram. The formula for the treatment scores differs from that given in Van den Brink & Ter Braak (1999). They used CANOCO 3.14. The difference arises because CANOCO 4.5 reports TAU and uses a different scaling of the species scores. See section 8.4.4 for the same formula in a multiple regression context. Figure 8-3 shows the resulting PRC diagram.

CanoDraw for Windows assists with the combination of the canonical coefficients values with the information obtained from the analysis log (the TAU value and the standard deviations of individual environmental variables), in the *Project / Import variables / Setup PRC scores* command (see sections 12.4.8.3 and 14.8).

On the logarithmic scale, the inferred changes with respect to the control are calculated as treatment score * species score. The log-change of species 4 in week 3 in treatment High is $-0.7385 \times 0.6916 = -0.5107$, i.e. the species decreases by ca. 50% compared to the control in that week. To be precise, the predicted count in treatment High is $\exp(-0.5107) = 0.60$ times the count in the Control (70), i.e. $70 \times 0.60 = 42$, which fits precisely the observed count (Table 8.13), because here we are analyzing noise-free simulated data.

In the projects prc.con and prc_x.con and Table 8.14, the explanatory product variables are arranged per week. For plotting the PRC diagram, it is more convenient to order the explanatory product variables per treatment (project prc1.con). The results are the same. Note that the sign of PRC treatment scores and species weight may be interchanged. This is done in the Excel

worksheet prc_sim.xls, so as to stress that most of the species (4 out of 6) decreases in abundance with higher dosage. Figure 8-3 is made with Powerpoint (file prc_sim.ppt)

Week 0 is pre-treatment, so that no treatment effects are expected in week 0. Therefore it is more logical to delete the variables W0*Low and W0*High from the analysis (project prc_alt.con).

For completeness and illustration, the projects are supplemented with permutation tests. In the permutation test, we permute the cosms, not the samples. There are 4 samples per cosm (W0-W3), i.e. the number of split-plots is 4. The samples in the data file are arranged per week (Table 8.13), so that the "Take and Skip" rule is Take 1 and Skip 4. In Table 8.13, this rule brings one from, for example, C2-W0 to C2-W1, to C2-W2 and finally to C2-W3.

Because these are simulated, noiseless species data, the F-ratios are so large that they are sometimes printed as *****. A more realistic example of the permutation test is given in the next section.

8.3.12 Example PRC - Testing time-dependent effects by PRC

Problem:	What is the response across time of an invertebrate community to different dosages of an insecticide?
Data:	Van den Brink & Ter Braak (1999)
Directory:	\CANOCO\SAMPLES\PERMUTIO\PRC
Illustration of:	<ul style="list-style-type: none"> • Testing the significance of the first PRC diagram. • Testing the significance of the second PRC diagram.

Files	Name	Description
Species	species.dta	log-counts of invertebrate species in 12 experimental ecosystems (mesocosms)
Environmental	design.dta	design data file with 5 treatment (ds0-ds4) and 11 time indicators of which two Before (Wk-4 - Wk-1) and 9 After (Wk0.1 - Wk24) treatment at Wk0
Derived	dsgnax1.dta	design file with the first environmental axis of prctest1.sol added
Project	prc.con	PRC of Fig.3 of the paper with a test of first axis (first PRC diagram)
	prctest1.con	as above with deletion of the pre-treatment variables
	prctest2.con	PRC test of the second axis (second PRC diagram) using dsgnax1.dta
	prctst2.con	first, automatic part of the PRC test of the second axis using the console version of CANOCO
	prctst2b.con	second part of PRC test of the second axis using the console version of CANOCO
	prctst2c.con	first and second part to give a PRC test of the first and second axis using the console version of CANOCO

8.3.12.1 PRC: Description of the experiment

The example data are the invertebrate data set in Van den Brink & Ter Braak (1999). This data set was obtained from an experiment in outdoor experimental ditches. Twelve mesocosms were allocated at random to treatments; four served as controls and the remaining eight were treated once with the insecticide chlorpyrifos, applied as Dursban® 4E, with nominal dose levels of 0.1, 0.9, 6 and 44 µg/L in two mesocosms each. The dose levels are coded as ds0, ds1, ds2, ds3, and ds4 in the design data file, ds0 being the control. Sampling was done 11 times, from Week -4 pre-treatment through Week 24 post-treatment, giving in total 132 samples (12 mesocosms times 11 sampling dates) in the statistical analyses. A total of 189 different taxa were identified and counted in these samples. The responses and recovery of the invertebrate community after chlorpyrifos treatment were analysed in time using RDA in Van den Brink et al. (1996) and by PRC in Van den Brink & Ter Braak (1999).

In the data files, samples are arranged by sampling date. The sample code name "w2,c4", for example, stands for a sample from cosm 4 at week 2. This sample has identification number 52 (see the solution file) and received the third treatment number (dosage ds2), as can be seen in the file design.dta. The species data file (species.dta) contains $\ln(10x+1)$ -transformed counts.

8.3.12.2 PRC: Testing the significance of the first PRC diagram

The project prc_fig3.con does a default PRC analysis on the data. This project forms the basis of the first PRC diagram in Van den Brink & Ter Braak (1999). The difference among

weeks account for $1 - 0.781 = 0.219$ (21.9%) of the total variance; the treatment regime (i.e. the remaining week*treatment interaction) accounts for 33.5%. The first two eigenvalues are 0.087 and 0.029, showing that the first axis dominates the second. The first axis explains $0.087/0.335 = 26.1\%$ of the variance captured by the treatment regime, the second axis 8.6% (together 34.7%).

The first axis is significant ($P \leq 0.005$). The test in project `prc_fig3.con` does not use the Before-After aspect of the data because the Before weeks * treatment terms are also included. Any permutation of the time points would yield the same significance level (if the number of permutations is large).

In the project `prctest1.con`, the products involving Before weeks are not included. Now the data have a small BACI aspect and one would perhaps win a little power by permuting the time point also (not done in the project).

The second axis can be tested as described in section 8.2.4.2. The new point here is that there are already some covariables. For this reason the first environment-derived sample axis needs to be added to the covariable file first. This will yield a file like `dsgnax1.dta`. This file is used in project `prctest2.con` to test the second PRC diagram, which is not significant ($P = 0.5790$). As a check, note that the second axis of `prctest1.sol` is the same as the first of `prctest2.sol`, as required. Van den Brink & Ter Braak (1998) give an example in which the second axis is significant. They also show how to visualize the joint effect of the two PRCs.

It is perhaps easier to test the second axis with the console version of CANOCO. For this, the project `prctest1.con` is copied to `prctst2.con` and the last two last lines are deleted. In a Command-box (DOS-box), type CANOCO and enter at the first question a 1 and specify `prctst2.con` as answer file. Eventually CANOCO asks you to specify the number of permutations (still for the first axis), answer 1, then ask for

- more ordination axes, namely 1
- more analyses with the same data
- a Monte Carlo test of the first axis
- specify the permutation type

The answers of this second part are listed in project `prctst2b.con`. This second part is appended to the first part (`prctst2.con`) to obtain `prctst2c.con`, which is a complete project with which both the first axis and second axis can be tested using the command line

```
CANOCO <prctst2c.con
```

8.3.13 Example WELCH - A design-based test of interaction

Problem:	Testing the N.P interaction in experiment E40 by the Welch (1990) method
Data:	Van Dobben, Ter Braak & Dirkse (1999)
Directory:	\CANOCO\SAMPLES\PERMUTIO\WELCH
Illustration of:	<ul style="list-style-type: none"> • The Welch (1990) permutation test of interaction. • Permuting rows and columns of a data table by using the split-plot option. • Design-based versus model-based permutations.

Files	Name	Description
Species	species.dta	As Example in 8.3.3 (E40) but with the plots arranged in a standard order per block
Environmental	design.dta	As Example in 8.3.3 (E40) but with the plots arranged in a standard order per block
Project	welch1.con	permutation test of N.P interaction by permuting rows (P) and columns (N) by using split-plot permutation of whole-plots consisting of the same N-level and dependent split-plots consisting of the same P-level
	welch2.con	permutation test of N.P interaction by permuting rows (P) and columns (N) by using split-plot permutation of whole-plots consisting of the same P-level and dependent split-plots consisting of the same N-level

Table 8.15 Layout of the N*P experiment E40: 4 blocks with each 4 N-levels by 2 P-levels.

	N0	N1	N2	N3	Blk2	N0	N1	N2	N3	Blk3	N0	N1	N2	N3	Blk4	N0	N1	N2	N3
P0	x	x	x	x	P0	x	x	x	x	P0	x	x	x	x	P0	x	x	x	x
P1	x	x	x	x	P1	x	x	x	x	P1	x	x	x	x	P1	x	x	x	x

8.3.13.1 WELCH: Design-based test of an interaction as proposed by Welch (1990)

A factorial design in blocks can be displayed as in Table 8.15. Welch (1990) proposed to test for the interaction effects by randomly permuting the rows and columns of the 4*2 table per block. The main effects of the experiment, i.e. the row- and column-effects are eliminated by specifying the N- and P-indicator variables as covariables. If N*P products are specified as Environmental data, the pure interaction effects remain. Such effects are of the form $(P_i N_k - P_j N_l) - (P_j N_k - P_i N_l)$, where $P_i N_k$ indicates the plot which received level i of P and level k of N. Under the null hypothesis these effects are exchangeable, i.e. in the original data one might permute both the N-levels and the P-levels.

For this to be possible in CANOCO, the data need to be arranged so that, at least per block, the order of the P-levels is the same for each N-level. The data of experiment E40 were not in this form. Therefore the data files that appear in this directory are just a reordering of the original data.

In project welch1.con, the whole-plots are the N-levels (N0,N1,N2,N3), each consisting of two samples (with P-level P0 and P1). Both whole-plots and split-plots are randomly permuted, the split-plot permutation being the same per N-level by requesting dependent permutations across whole-plots. In this way, the row and columns of the table are permuted.

Of course, we can define rows and columns the other way round without changing the test. This is done in project `welch2.con`, where the whole-plots are the P-levels (P0 and P1) and the split-plots are the N-levels. Both whole-plots and split-plots are randomly permuted, the split-plot permutation being the same per P-level. In both cases the P-value is .85 (999 permutations). For comparison, the model-based permutations of section 8.3.3.2 yielded $P = 0.66$ (999 permutations under the reduced model).

8.4 Other ordination methods that are also available in CANOCO

8.4.1 Introduction

This section describes a number of statistical techniques that can be obtained with CANOCO as special cases of the six primary ordination methods in CANOCO (PCA, RDA, CA, DCA, CCA and DCCA).

8.4.2 Example LOGRATIO - Log-ratio analysis of compositional data

Problem:	How does the chemical composition of rock relate to depth and its porosity?
Data:	Aitchison (1984a: the Coxite data)
Directory:	\CANOCO\SAMPLES\METHODS\LOGRATIO
Illustration of:	<ul style="list-style-type: none"> • How to obtain a log-ratio analysis in CANOCO using PCA and RDA. • How to regress composition on explanatory variables. • Problems of the $\log(Ay+B)$-transformation with $B = 0$.

Files	Name	Description
Species	coxite.dta	chemical composition of rock (5 chemical species in 25 samples)
Environmental	depthpor.dta	depth and porosity of the samples
Derived	coxite2.dta	percentages changed to fractions except for samples S1 and S2
	cox_zero.dta	species C in sample S1 set to 0.00 in an all-fractions file
Project	pca_lgrt.con	Log-ratio PCA of compositional data (log-contrasts)
	pca_chk.con	Log-ratio PCA on coxite2.dta
	rda_lgrt.con	Log-ratio RDA of compositional data (with permutation test)
	rda_forw.con	manual forward selection of variables with permutation tests that account for the spatial arrangement of the samples
	error.con	Log-ratio PCA on cox_zero.dta with $\log(y)$ transformation
	corr_err.con	Log-ratio PCA on cox_zero.dta with $\log(100y+1)$ transformation

8.4.2.1 LOGRATIO: Log-ratio analysis of compositional data (generalized logit analysis)

Compositional data *sensu stricto* are data scaled so that each sample total sums to 1 or 100% or data with unequal sample totals in which the sum is arbitrary. Biplots of compositional data are briefly discussed on pages 144-145 of Unimodal models. See also section 3.9.2 for the theory. In this subsection we illustrate how to obtain such biplots in CANOCO. As an example we use the coxite data set presented by Aitchison (1984a: pp. 535-536). These (artificial) data consist of the percentages of five chemical constituents in 25 samples of rock taken at different depths (file coxite.dta). All percentage values are strictly greater than 0.

A biplot of log-ratios can be obtained with the project pca_lgrt.con, which does a log-ratio PCA or, as Aitchison (1984b) calls it, a loglinear-contrast PCA. Note that the species data are

log-transformed (without any added constant: $B=0$) and that the data are centered by samples as well as by species. The scaling may focus either on the samples or on species; there is no need to post-transform the species scores. Also note that the data are percentages, not ratios.

To illustrate that the absolute values are unimportant in log-ratio analysis, the coxite data are expressed as fractions, except for the samples S1 and S2. In the resulting file, `coxite2.dta`, the sample totals for S1 and S2 are 100 and for the remaining samples 1.00. The project `pca_chk.con` does a log-ratio PCA on `coxite2.dta`. You may wish to verify that the ordination summary remains as in project `pca_lgrt.con`. Note also that the total standard deviations (after site- and species-centering) are identical. The solution files are identical within numerical precision. Both projects thus result in the same biplot. This shows that log-ratio analysis focuses on relative abundances only.

In contrast with the analysis of compositions and relative abundances by CCA, the weight of samples S1 and S2 (with total 100) in the analysis is equal to that of the other samples (with total 1). Log-ratio analysis is based on the idea that each fraction is measured equally precisely. In contrast, CA and CCA use the idea that, for counts, the precision increases with the sample total, hence the different weighting scheme.

Log-ratio analysis can also be carried out with predictor variables. The composition is regressed on the explanatory variable Depth in the project `rda_lgrt.con`. Depth explains only 6% of the variance and is not judged significant. The project `rda_forw.con` specifies a manual forward selection (for Canoco for Windows only) with permutation tests based on cyclic shifts. The variable Porosity explains 39% and is significant ($P=0.03$).

If the data contain some zero values, it is wrong to use the transformation $\log(A*y+B)$ with $A=1$ and $B=0$. See section 5.6.2. This is illustrated with the data file `cox_zero.dta` (all fractions with one zero value) which is analyzed in the project `error.con`. Please note that Canoco for Windows gives warnings as before. However, the results are non-sensical. The biplot of this project suggests that sample S1 has a very high percentage of component C compared to the other samples and other components. However, it was this sample and this component that were set to zero! Any reasonable biplot should show that sample S1 has an extremely low percentage of C. The error is corrected in `corr_err.con` in which the data are $\log(100y+1)$ -transformed. The 100 means that all data (fractions!) are transformed to percentages, to which 1 is added. The resulting biplot correctly shows that there is an outlying low value, the original zero.

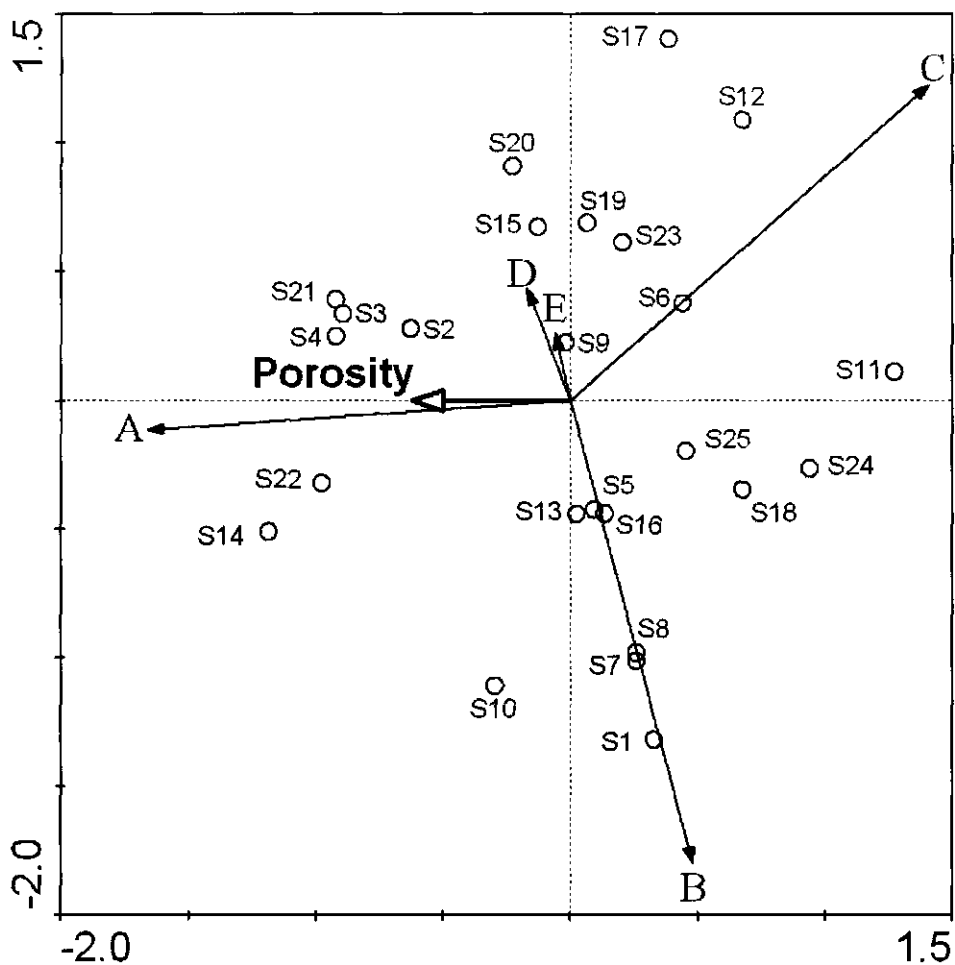


Figure 8-4 Ordination diagram based on the redundancy analysis of the coxite data (Aitchison 1984a).

Figure 8-4 shows the ordination diagram of RDA on porosity; the first axis ($\lambda_1 = 0.39$) displays the relation of composition with porosity; the second axis displays the residual variation ($\lambda_2 = 0.48$). Porous rocks lie on the left hand side of the diagram and contain the largest percentages of constituent A. The least porous rocks lie on the right hand and contain the largest percentages of constituent C.

8.4.3 Example CVA - Canonical Variates Analysis

Problem:	How do three <i>Iris</i> species differ in flower morphology?
Data:	Fisher's <i>Iris</i> data, e.g. in Mardia, Kent & Bibby (1979) or the Minitab example files
Directory:	\CANOCO\SAMPLES\METHODS\CVA
Illustration of:	<ul style="list-style-type: none"> • A standard CVA with triplot showing group overlap and group means. • Permutation test of differences between groups using CVA. • Forward selection in CVA. • How to obtain means and correlations of the variables (for all data or per group). • How to obtain a pooled within-group correlation matrix.

Files	Name	Description
Species	iris_spe.dta	Indicators to which species (<i>Iris setosa</i> , <i>I. versicolor</i> and <i>I. virginica</i>) each of the 150 specimens of <i>Iris</i> belongs (50 per group)
Environmental	iris_flw.dta	4 measurements on the flower of each specimen (length and width of the sepal and of the petal)
Derived	stmeans.dta	standardized group means (spec_env.tab of the CVA)
	iris_fl2.dta	copy of iris_flw.dta for the trick to obtain group-means and the within-group covariance matrix
Project	cva.con	default CVA on the <i>Iris</i> data
	cva_f.con	as cva.con but with manually modified solution file cva_f.sol
	cva_forw.con	forward selection of variables
	means_v.con	project giving means and correlations for <i>I. versicolor</i> only (using a standard trick)
	withinco.con	project giving the pooled within correlation matrix by specifying the groups as covariables

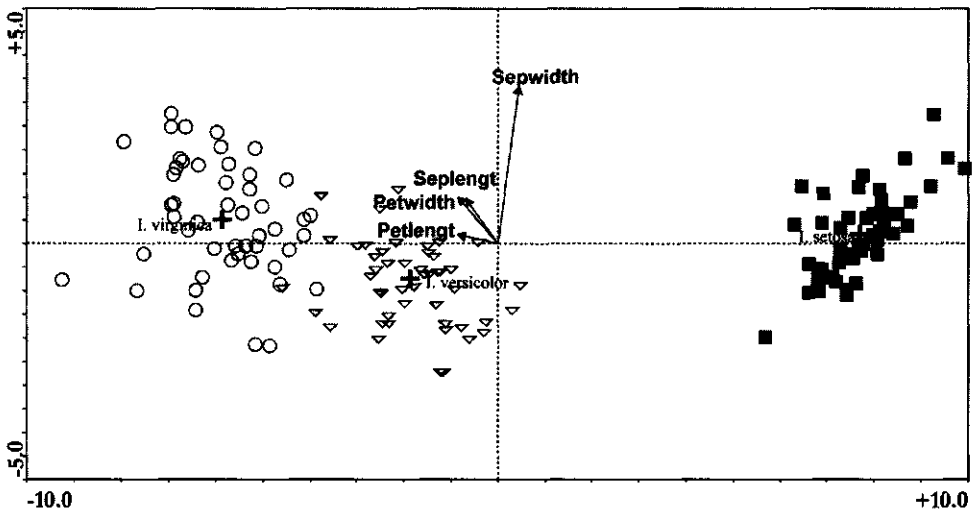


Figure 8-5 Triplot based on a CVA of the Fisher's Iris data.

8.4.3.1 CVA: Canonical Variate Analysis (Discriminant Analysis)

We illustrate CVA using Fisher's famous *Iris* data. These data concern three *Iris* species, each represented by 50 specimens. For each specimen, four measurements of its flower were recorded, namely sepal length, sepal width, petal length and petal width. The question which CVA addresses is which linear combination of the flower measurements discriminate best between the three species as based on the analysis of variance F-ratio (e.g. Mardia, Kent & Bibby, 1979, Jongman et al. 1987: pp 148-149). The result is the first axis. After this, a second axis is derived that best discriminates and that is uncorrelated with the first.

In the example the statistical units are the specimens (or flowers if you wish), so that, for CANOCO, the flowers are the samples. For CVA, the groups, which are here the three *Iris* species, must form the Species data in CANOCO (file *iris_spe.dta*). The measurements of which we want the best linear combinations to discriminate the groups must be the Environmental data (file *iris_flw.dta*). The other options for obtaining a standard CVA are: CCA with Hill's scaling and focus on inter-species distances (project *cva.con*).

The triplot (Figure 8-5) shows that *I. setosa* specimens are very different from those of the other species, the inter-species distance being in this scaling the Mahalanobis distance. Also the sample distributions of *I. virginica* and *I. versicolor* hardly overlap. The arrows allow an inference of the (means of the) original measurements: *setosa* has smaller flowers except for sepal width. The long, nearly vertical arrow for sepal width shows that it differs also considerably within species. The output file *spec_env.tab* (*stmeans.dta*) contains the means after standardization of each variable to zero mean and unit variance. The table of species tolerances (Tol:) contains the standard deviations along the axes of Figure 8-5 for the groups. They are reasonably equal among species and axes, as required. The plot of species centroids and specimen scores is given by Krzanowski (1988) with the first axis mirrored. The plot shown in Mardia et al (1979) is schematic. The addition of the arrows for the measurements was first proposed by Gabriel (1981).

Unfortunately, there is a problem with the CANOCO environmental biplot scores for use in CVA. The problem is that the lengths of the environmental vectors (based on BipE scores; section 6.3.9) do not give the right impression about the relative importance of the environmental variables for group separation. The reason is that in CANOCO each environmental variable is standardized to unit variance. This is the unit of measurement for the biplot scores. A better impression about the relative importance would be obtained by standardizing each environmental variable to unit within-group variance. The squared length of the biplot vector is then a measure of the F-ratio of between-to-within-variance of the corresponding environmental variable. Together with the group centroids, so scaled environmental biplot scores continue to display the approximate group means, but in a different unit of measurement. The new unit of measurement is the within-group standard deviation. To obtain this unit of measurement, we need to calculate for each variable a factor *f* which is the ratio of the total variance to the within-group variance and multiply each of the BipE scores with the square-root of *f*. If the within-group variance is relatively small compared to the total variance or, equivalently, if the between-group variance is relatively large compared to the total variance, the vector for that variable will become relatively larger.

With Canoco for Windows the factor *f* can be determined by calculating the between-groups variance using the project *Betweenco.con*. In this project the flower measurements are specified as species data and the groups (the *Iris* species) as environmental data. Using these data, a redundancy analysis is carried out. Import the solution file *betweenco.sol* in Excel and search for the table with title CFit (cf. Table 6.54). The values under the heading %EXPL are the

percentages variance explained by the groups. The factor f (the ratio of the total variance to the within-group variance) is now $100 / (100 - \%EXPL)$. For the four variables we obtain:

	Variable	f	\sqrt{f}
1	Sep length	2.622	1.619
2	Sep width	1.669	1.292
3	Petal length	17.065	4.131
4	Petal width	14.065	3.750

The importance of Petal length and Petal width for the separation among the Iris species is thus much higher than suggested by Figure 8.5. To create the modified figure with CanoDraw is somewhat tricky. A safe way to go is to open *cva.con* in Canoco for Windows under the name of *cva_f.con* and to modify the name of the solution file in this project to *cva_f.sol*. After this, manually modify the BipE scores in solution file *cva.sol*, for example using the Notepad program, and save the file under the existing name *cva_f.sol*. Make sure to retain the same layout of the number as in the original *cva.sol*. Now invoke CanoDraw from the Start Menu, open *cva_f.con* as a new project and click *Create / Simple Ordination Plot / Triplot*.

The eigenvalues issued by CANOCO are somewhat atypical for a CVA (see section 3.11); they are best reported as $\theta = \lambda / (1 - \lambda)$. For the *Iris* data, CANOCO reports as eigenvalues 0.9699 and 0.2220. The standard CVA eigenvalues are thus $0.9699 / (1 - 0.9699) = 32.2$ and 0.28. The canonical coefficients given by CANOCO are standard; recall that they must be divided by the standard deviation to express them in terms of the original measurements. The first canonical variate is

$$0.84 \text{ Sepal length} + 1.55 \text{ Sepal width} - 2.22 \text{ Petal length} - 2.84 \text{ Petal width}$$

The canonical coefficients are numerically somewhat unstable because of variance inflation factors of 31 and 16. In contrast, the triplot is numerically stable.

The project *cva_forw.con* does a forward selection of variables with permutation tests. The variables are added in the sequence: petal length, sepal width, petal width and finally sepal length. The last variable is not significant given the other three variables.

The means and variances per group can be obtained with CANOCO by applying a standard trick: carry out a PCA with the measurement variables specified both as Species data and as Environmental data. For this trick, you must make a copy of the data (*iris_fl2.dta*). The means per group are then obtained by deleting the samples (specimens) that do not belong to the group (e.g. project *means_v.con*).

The pooled within-group correlation matrix can be obtained by specifying the group indicators as covariables (project *withinco.con*). The data entered as Species data are arbitrary. Two variance inflation factors are higher than 20, showing that the measurements are also highly correlated within-groups. For predictive purposes (determination of new specimens) at least one variable must be deleted, as was found also in the forward selection.

Insert legend into created diagrams

This simple **on-off** option determines whether a legend is created for the ordination and attribute diagrams. All the remaining options below this one are available only if this option is **on** (checked).

Legend is generally composed from one or more legend sections, which collect information about the appearance of graphical attributes shared by the diagram items of a similar kind. Examples of legend sections may be marks showing types of species symbols, corresponding to species of different classes, lines of different color corresponding to different series from an active series collection of samples, or square patches of fill patterns corresponding to pie-wedges of sample pie-symbols, representing different classes of species. Legend sections therefore contain zero, one, or more items of similar type. Legend section can be optimally introduced by a heading showing its name.

Legend position

The choices for this option determine the actual position of the legend area (rectangle) – they determine to which window (paper) edge the legend is closest.

Sections layout

This item determines the arrangement of whole legend sections across the legend area. Legend sections may be arranged horizontally, filling rows until the "wrap quota" is reached. Then the starting position of the next legend section moves to the left side of the next row. Alternatively, legend sections may have a vertical layout, where each new legend section is below the preceding one, until the "wrap" quota is reached.

Wrap sections after

This value determines how many legend sections are put into one row (column), before the next section is placed in the leftmost position of the next row (or in the topmost position of the next column). For example, if the section layout is "Horizontal" and following **Wrap** field has value 2, the second legend section is positioned to the right of the first one, and the third section (if any) has its left side aligned with the left side of the first section and is placed below it.

Items layout in sections

This layout option is similar to the *Section layout* option described above, but concerns the nested hierarchical level – arrangement of items within a particular legend section. Additionally, the two options of horizontal and vertical layout are extended by the optional presence of the section heading ("w. heading").

Wrap items after

How many section items are laid out in the horizontal (for horizontal section items layout) or vertical direction before the next row (or column) in the legend section is entered.

8.4.4 Example MULTREGR - Multiple regression with CANOCO

Problem:	To relate y to predictors x1, ..., x5
Data:	M. J. Anderson (unpubl.)
Illustration of:	<ul style="list-style-type: none"> • Multiple linear regression by using RDA. • The partial tests and partial regression coefficients.

Files	Name	Description
Species	y_var.dta	one quantitative response variable (y) in 9 units
Environmental	x_vars.dta	5 predictor variables (x1-x5)
Derived	yx_vars.dta	all data (y, x1- x5)
Project	multregr.con	multiple regression of y on x1-x5
	mltrgr.con	idem, using yx_vars.dta as Species data
	partialx.con	regression of y on x1 with x2-x5 partialled out (i.e. with covariables x2-x5)
	forward.con	forward selection with permutation tests

Table 8.16 Multiple regression of y on x1 - x5: standardized regression coefficients (Regr:) and associated t-ratios (Tval:) as given by CANOCO.

The usual regression coefficients (b) are the Regr-values divided by the standard deviation of the predictors (Sd env) times 2.329 (TAU, the standard deviation of y). If the species score of y is -1, b must be replaced by -b. The constant of the regression with respect to centered predictors is the mean of y (7.3209).

name	Regr:AX1	Tval:	sd env	b
x1	0.1365	0.828	0.8988	0.353775
x2	0.4543	2.809	0.695	1.522702
x3	0.6998	3.4579	1.0029	1.625449
x4	0.255	1.4978	0.6221	0.954854
x5	0.5978	3.6421	0.6898	2.018784

8.4.4.1 MULTREGR: Univariate analysis by multiple regression

There are many computer packages that provide more extensive facilities for multiple linear regression than CANOCO. Multiple regression with CANOCO is nevertheless a good choice if

- you want to use permutation tests because you do not want to rely on the assumptions of the normal distribution in significance testing.
- your data are already in a CANOCO-format and you want to do a quick exploratory analysis

The example in this subsection also has an instructive purpose. Some aspects of canonical ordination are best explained in the simple context of multiple regression.

The project multregr.con does a multiple linear regression of a response variable y on five predictor variables x1 - x5, via RDA with focus on inter-sample distances. This focus is chosen because the species scores for y are equal to 1 or -1 in this scaling; this makes the output of CANOCO more easily interpretable in a quantitative way. The first and only canonical

eigenvalue (and thus the sum of all canonical eigenvalues) of the analysis is the fraction of explained variance. The species-environment correlation of the first axis is the multiple correlation coefficient, usually denoted by R . The second eigenvalue is the fraction of unexplained variance. The log-window also contains the means, standard deviations and correlations of the predictor variables. For the example data, CANOCO does not detect any high influence points.

Regression coefficients and associated t-ratios are listed in the solution file. The regression coefficients (Regr:) given by CANOCO are standardized regression coefficients: they apply to the normalized variables $y, x_1 \dots x_5$, i.e. in which each has zero mean and unit variance (with divisor n instead of $n-1$). If the species score for y is -1 , all signs must be changed. The t-values of regression coefficient are the usual ones (estimate divided by the standard error of estimate). Because we are carrying out univariate linear regression here, we may conclude from the t-values that x_2, x_3 and x_5 are significant at the 5% level (at least if the errors are independent and approximately normal). To obtain the regression coefficients in terms of the original variables we need to divide the standardized coefficients by the standard deviations of the predictors $x_1 \dots x_5$, and multiply the result by the standard deviation of y (Table 8.16). The latter is given under name of "total standard deviation in the species data TAU" in the log-window, as y is the only variable in the Species data. The constant of the regression (with centered predictors) is the mean of y , which is $-\text{ORIGIN} * \text{TAU}$ where ORIGIN is given among the species-derived sample scores. In the example, the constant is $3.1427 * 2.329 = 7.32$. An alternative way to obtain the mean of y is to specify a copy of the y -data among the Supplementary Environmental data ($yxvars.dta$), as is done in the project.

With one response variable, the species-derived and environment-derived samples scores are identical for the first axis. The first axis of the sample scores is the fitted value of the normalized response variable, whereas the second axis of (species-derived) sample scores is the residual. A plot of the sample scores in CanoDraw thus shows "residuals against fitted values", a standard plot to check for homoscedasticity of the error variance. To obtain the usual fitted values, use

$$(8.3) \quad \text{Fitted values} = \text{mean}(y) + \text{TAU} * \text{Sample scores} * \text{Species score}$$

where all scores are from the first axis, and the species score of y is either $+1$ or -1 . Note that (8.3) is the usual algebraic formula for the biplot rule in one dimension. To obtain the usual residuals, multiply the second axis scores by TAU.

The project `mltrgr.con` uses the data file with all variables, `yx_vars.dta`, as the Species data file. This project shows that it is possible to carry out a univariate regression by deleting all superfluous variables in the project. Deleting very many variables from the Species data file may, however, occasionally be numerically unstable. To be on the safe side, always put the response data in a separate data file.

A test for a partial regression coefficient is obtained by specifying the other predictors in the model as Covariable data. An example is given in project `partialx.con`. The resulting, permutation based P-value for x_1 is 0.46, which is about the same as the P-value based on the t-value of x_1 . The permutation test is obtained using forward selection for reasons of efficiency only: in CANOCO, testing the significance of single variables (1 degree of freedom per test) is done most efficiently with forward selection.

Note that the regression coefficient for x_1 in the Regr: table in the solution file `partialx.sol` is the same as that in `multreg.sol`. This happens because there is just one response variable. With more than one response variable, the coefficients will differ numerically, but their meaning remains unchanged in the sense that the coefficients give the conditional effect of the variable,

i.e. the effect adjusted for the effects of all other variables in the model (both covariables and environmental variables).

In the project forward.con, the predictors are selected by forward selection. In each step of the selection a permutation test is carried out. The variable x1 is added last; the resulting P-value is the same as in the previous project, provided the number of permutations is large.

9. Program PrCoord

9.1 Working with PrCoord

The PrCoord program implements principal coordinate analysis (PCO or PCoA) for non-negative (semi-)quantitative or presence-absence data. Canoco for Windows is able to calculate only a limited number of the most important axes (principal coordinates). To obtain the full solution, you must use the PrCoord program. This is needed for example for the distance-based redundancy analysis (Legendre & Anderson, 1999), as illustrated in the section 9.3 of this chapter. The user interface of the PrCoord program is integrated into a simple application window, illustrated in Figure 9-1.

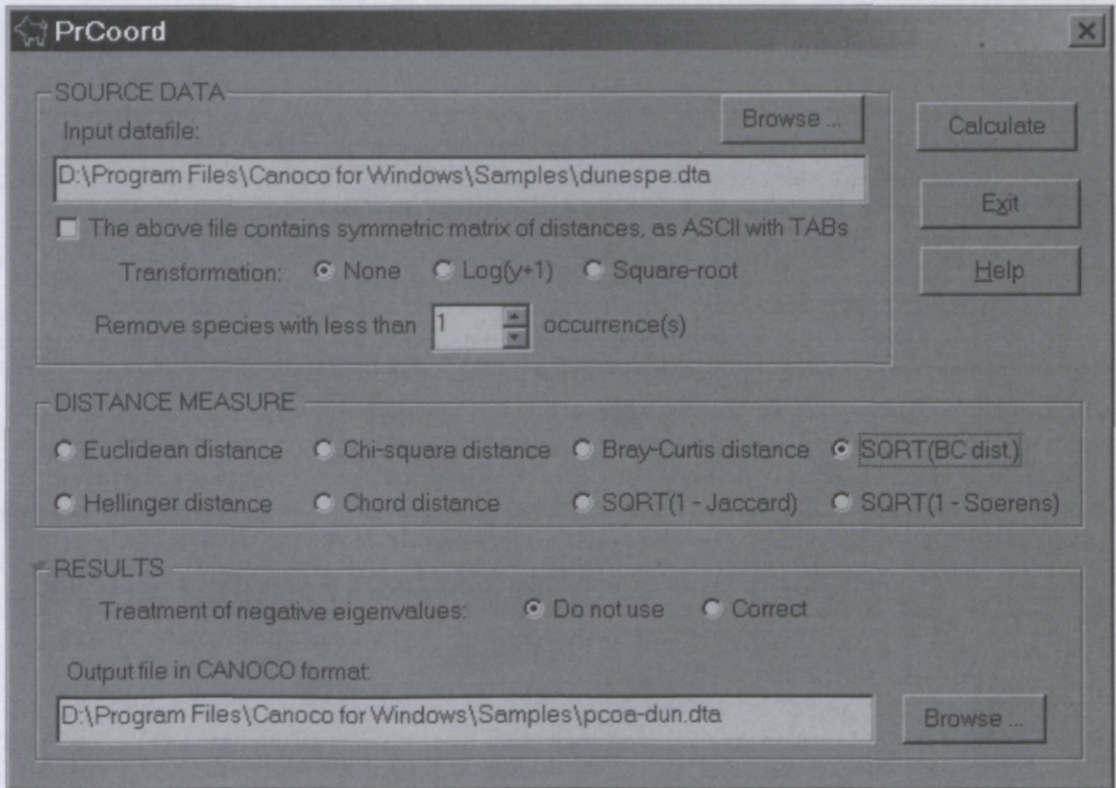


Figure 9-1 User interface of PrCoord program

To calculate the PCO solution with the PrCoord program, you must proceed with the following steps:

1. Specify input data

- * You can specify an existing Canoco data file as the input data and then select one of the available distance measures (see section 9.2 for additional details). Note that PrCoord does not accept a data matrix which contains empty samples (with zero sum of values) and also does not accept a data matrix with negative values, except when the *Euclidean distance* was chosen.

- * Alternatively, you can specify as input data file a text file representing the full matrix of distances, with each row of the matrix on a separate row of the text file and the individual columns separated by the TAB characters. This possibility allows you to calculate any kind of dissimilarity measure outside the PrCoord program and use it here, if the file contents meet the following requirements. The first row must contain the names of individual samples, separated by TAB characters. The following rows do not start with the row name. Therefore, if you have a dissimilarity matrix for N samples, the input file in TAB-separated format must have N+1 rows and N columns. To specify to PrCoord program that the input data represent a matrix of distances, you must check the option named *The above file contains symmetric matrix ...*
2. Decide about optional transformation and reduction of the input data, unless the input file represents a matrix with distances. You can omit species (variables) with less than N occurrences. The default value of 1 selects all the species present in the data. The data can be also optionally log-transformed (with the value 1.0 added before the transformation) or square-root transformed. Any negative values are set to zeros when the log-transformation or the square-root transformation is applied.
 3. Select the distance measure to be used, unless the input already represents a distance matrix. Beside the Euclidean distance and Chi-square distance, the other available distance measure include the frequently used Bray-Curtis distance, square-root of the Bray-Curtis distance, two distances related to Euclidean distance: the Hellinger distance (advocated for example by Legendre & Gallagher, 2001) and Chord distance. Additionally, two dissimilarity measures based on the presence and absence of species (variables) in the samples are available. These distances are calculated as square-rooted complements of either Jaccard or Soerensen similarity coefficients. Note that if one of these last two measures is selected, any quantitative input data are implicitly transformed to 0/1 values during the calculations. Additional information about the distance measures supported by PrCoord can be found in Legendre & Legendre (1998, Chapter 7).
 4. Decide (for any kind of input data) about the treatment of negative eigenvalues. PrCoord can either ignore the principal coordinates with negative eigenvalues or it can use the Lingoes correction method (see Legendre & Legendre 1998, p. 434). Note that the negative eigenvalues occur only if the (non-transformed) Bray-Curtis distance was selected (as this measure is not metric), or if a matrix of non-metric distances was specified as the program input. Note that the correctness of "correcting" the distance matrix to prevent occurrence of negative eigenvalues is questionable (see McArdle & Anderson, 2001).
 5. Specify the name and location of the output file which will be created using the Canoco full format. The sample scores on individual axes are represented by variables named *Ax1*, *Ax2*, etc.

After you obtained the PCO sample scores using PrCoord, you can display the PCO results (the sample scores) using Canodraw for Windows. For this, the output file produced by PrCoord must be entered as species data in Canoco for Windows. Then, specify a principal components analysis (PCA) with the scaling focused on inter-sample distances and species scores not post-transformed. The resulting PCA scores are equal (except a constant rescaling factor) to the PCO scores and can be visualized in a scatter of samples with CanoDraw.

You can also use the PCO sample scores as the response variables in a constrained analysis in Canoco, producing effectively the solution of distance-based redundancy analysis (db-RDA, see section 9.3).

9.2 Implementation details

The PrCoord program implements the Gower (1966) method of calculating the PCO solution. The matrix of distances $\{D_{ij}\}$ is transformed to $\{-0.5 D_{ij}^2\}$ and then double centred (e.g. Jongman et al. 1987: equation 5.17) and then submitted to the singular value decomposition (SVD) routine *DSYEV* from the LAPACK library (Anderson et al. 1999).

PrCoord can read an input data file with up to 25 000 samples and 5000 variables (species). Similarly, the input matrix of distances can refer up to 25 000 samples. Note, however, that practical limitations of the number of samples will be usually much lower on most computers. Input of a symmetrical matrix of distances for 10 000 samples, for example, requires allocation of dynamic memory with the size exceeding one gigabyte at one stage of the algorithm. The read or calculated distance matrix is stored using the "single-precision" floating-point representation, the transformed matrix A passed to the SVD routine uses the "double-precision" floating-point values.

Following paragraphs show the formulae used to calculate the individual distance metrics in the PrCoord program. In the formulae, y_{ij} is the value of j -th species (variable) in the i -th sample, y_{i+} is the sum of the (species) values in the i -th sample, y_{+j} is the sum of j -th species values over all samples, y_{++} is the total sum of values in the data matrix, m is the number of species (variables), a is the number of species occurring in both compared samples, b and c is the number of species occurring, respectively, only in the first or only in the second sample.

Euclidean distance between samples 1 and 2 is calculated as:

$$(9 - 1) \quad D_{12} = \sqrt{\sum_{j=1}^m (y_{1j} - y_{2j})^2}$$

Chi-square distance between samples 1 and 2 is calculated as:

$$(9 - 2) \quad D_{12} = \sqrt{y_{++}} * \sqrt{\sum_{j=1}^m \frac{(y_{1j} - y_{2j})^2}{y_{1+} y_{2+} y_{+j}}}$$

Bray-Curtis distance between samples 1 and 2 is calculated as:

$$(9 - 3) \quad D_{12} = \frac{\sum_{j=1}^m |y_{1j} - y_{2j}|}{\sum_{j=1}^m (y_{1j} + y_{2j})}$$

Hellinger distance between samples 1 and 2 is calculated as:

$$(9 - 4) \quad D_{12} = \sqrt{\sum_{j=1}^m \left(\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}} \right)^2}$$

Chord distance between samples 1 and 2 is calculated as:

$$(9 - 5) \quad D_{12} = \sqrt{2 * \left(1 - \frac{\sum_{j=1}^m y_{1j} y_{2j}}{\sqrt{\sum_{j=1}^m y_{1j}^2} \sqrt{\sum_{j=1}^m y_{2j}^2}} \right)}$$

Jaccard similarity between samples 1 and 2 is calculated as:

$$(9 - 6) S_{12} = \frac{a}{a + b + c}$$

Note that the S_{12} value is then subtracted from 1.0 and then the square-root of the result is taken.

Soerensen similarity between samples 1 and 2 is calculated as:

$$(9 - 7) S_{12} = \frac{2a}{2a + b + c}$$

Note that the S_{12} value is then subtracted from 1.0 and then the square-root of the result is taken.

Also note that the Chord distance is implied in PCA and RDA methods in the CANOCO program, when standardization by sample norm was selected. The transformation implied by the use of Hellinger distance cannot be directly achieved in the CANOCO program - the original data values are replaced by the square roots of the relative contribution of individual species to the sample total.

9.3 How to calculate db-RDA with Canoco software

In our tutorial, we will use the dune meadow data, described in the DUNEBOOK example in section 8.2.5. The aim of our analysis is comparable to the analysis represented by the *cca_bipl.con* project in the `\CANOCO\Samples\Unimodal\Dunebook` directory, but we will base our analysis on the Bray-Curtis distances among the samples.

Start the PrCoord program and select the *table01.dta* as the **Input datafile**, using the **Browse** button. We will not transform the species data and we will not omit any species, so use the default values of the other settings in the **SOURCE DATA** area. Select **Bray-Curtis distance** in the **DISTANCE MEASURE** area and keep the **Do not use** choice for the **Treatment of negative eigenvalues**. In the **Output file in CANOCO format** field specify `\CANOCO\Samples\Unimodal\Dunebook\pcoa-dun.dta`, where *CANOCO* should be replaced by your actual Canoco install directory.

After you click the **Calculate** button, the PCO solution is calculated, and the *Analysis Report* window appears. You can see there that the analysis of the matrix with Bray-Curtis distances among 20 samples leads to 20 principal coordinates. Because the *Do not use* box is checked, only the 14 principal coordinates with positive eigenvalues are written to the output file *pcoa-dun.dta*. Close the window and then also the PrCoord program, using the **Exit** button.

Open the Canoco for Windows program and create a new project. In the Project Setup Wizard specify on the first page that you have *Species and environment data available*, but do also check the *Supplementary environment data available* option. In the same page, keep the default setting of *direct gradient analysis*.

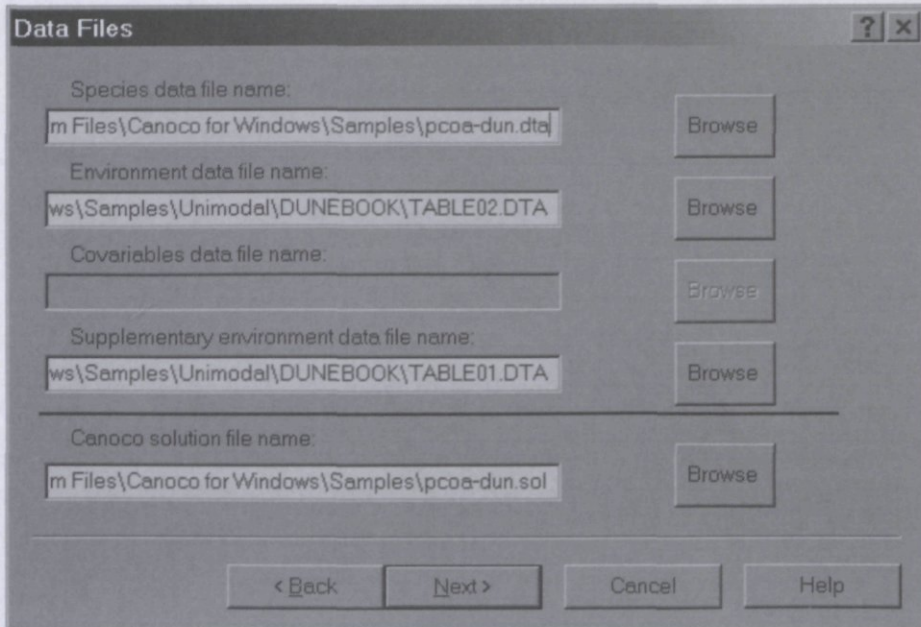


Figure 9-2 Specifying input data for db-RDA in Canoco

In the next page (see Figure 9-2 for its final look), specify the *pcoa-dun.dta* file produced by the PrCoord program as your species data, specify the *Table02.dta* (in the `\CANOCO\Samples\Unimodal\Dunebook` directory) as the environment data file, and specify the *Table01.dta* (which was already used as the input of PrCoord program) as the file with supplementary environment data. You can use the *pcoa-dun.sol* as the name of the Canoco solution file.

Specify RDA on the next wizard page and keep default settings for the remaining options. Save the resulting Canoco project under the name *pcoa-dun.con* and click the Analyze button. In the Canoco log view you can see that the first two axes explain 51% of the total variability among the Bray-Curtis distances (as represented by the 14 principal coordinates with positive eigenvalues, while ignoring the 5 coordinates with negative eigenvalues). The summary shown in the log view is actually the second summary which summarizes the relation of the ordination axes with the supplementary variables (in this summary the species-environmental correlations are given as 0.000, because the fit is perfect, there being more supplementary variables than samples). By scrolling up in the log view, you can find the first summary, which summarizes the RDA of principal coordinates to the real environmental variables. You can see for instance that the third eigenvalues (0.049) is much smaller than the second (0.177) and that all environmental variables together explain 64.76 % of the variability in the Bray-Curtis distances.

We included the actual species data as the supplementary variables in our analysis to enable visualization of species occurrences within the db-RDA ordination space. To do so, you should specify these variables as nominal variables in the CanoDraw program. The symbols for individual species then lay on the (weighted) centroids of samples, in which they occur.

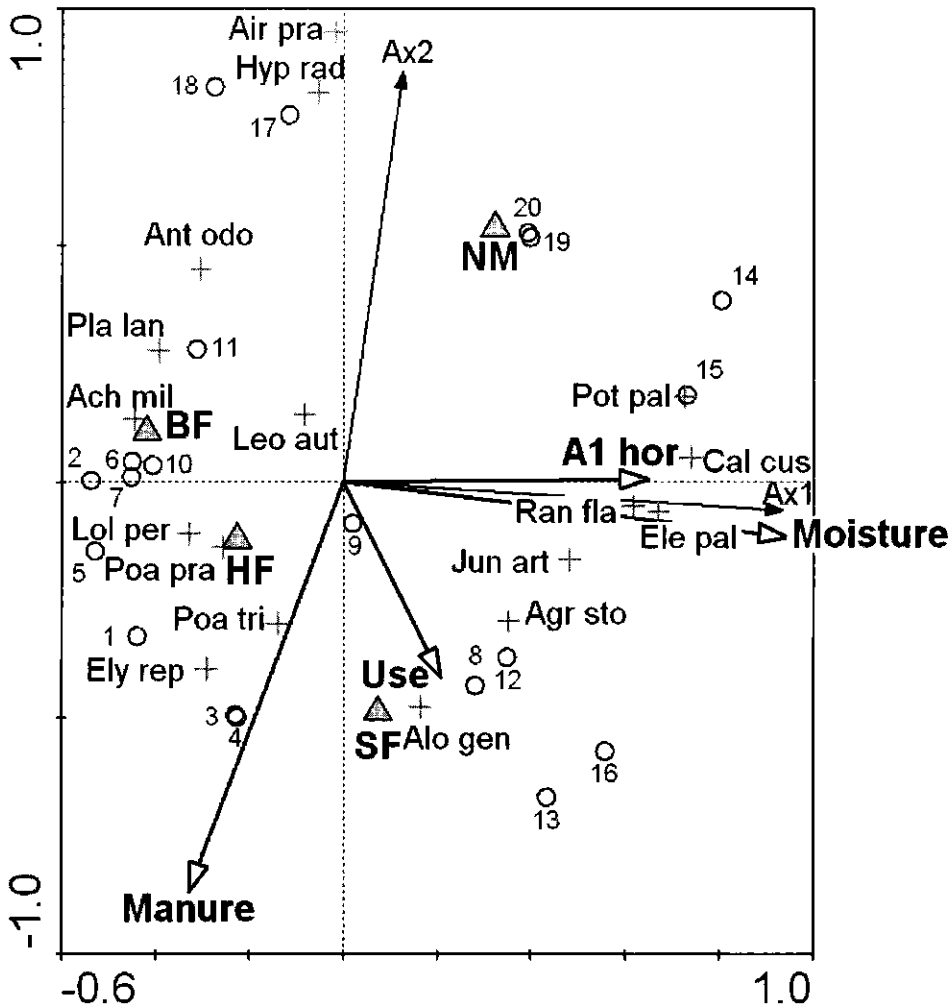


Figure 9-3 Diagram from the distance-based RDA, first two axes

The diagram in Figure 9-3 represents a triplot diagram with both the environmental variables and the supplementary variables (individual plant species). The thin arrows in this diagram correspond to individual axes of the analysis of principal coordinates. Note that only the "species" (PCO axes) with at least 10% of their variability explained by the first two axes of db-RDA are shown (*Project / Settings / Inclusion Rules*). The selection of *Ax1* and *Ax2* demonstrates clearly the coherence between the solutions from the unconstrained PCO and the constrained db-RDA. The plotted species are those with correlation with the ordination axes exceeding 0.5 in the absolute value (*Project / Settings / Inclusion Rules 2*, with values -0.5 and +0.5 placed in the two fields at the bottom).

CanoDraw for Windows User's Guide. Version 4

The code for Generalized additive modelling (GAM) in CanoDraw is partially based on a public domain version of GAIM software (Hastie & Tibshirani, 1990).

Loess model fitting is based on DLOESS code and in the following two paragraphs, its copyright notice is quoted:

1. *The authors of this software are Cleveland, Grosse, and Shyu. Copyright (c) 1989, 1992 by AT&T. Permission to use, copy, modify, and distribute this software for any purpose without fee is hereby granted, provided that this entire notice is included in all copies of any software which is or includes a copy or modification of this software and in all copies of the supporting documentation for such software.*
2. *This software is being provided "as is", without any express or implied warranty. In particular, neither the authors nor AT&T make any representation or warranty of any kind concerning the merchantability of this software or its fitness for any particular purpose.*

Export in the PNG format uses the PNG library:

libpng version 1.0.8 - July 24, 2000

Copyright (c) 1998, 1999, 2000 Glenn Randers-Pehrson

(Version 0.96 Copyright (c) 1996, 1997 Andreas Dilger)

(Version 0.88 Copyright (c) 1995, 1996 Guy Eric Schalnat, Group 42, Inc.)

My thanks go first and foremost to my wife Marie and to my daughters, for their patience with me. Petr Šmilauer

10. CanoDraw Introduction

CanoDraw for Windows is a program written specifically for users of Canoco for Windows and it is distributed with Canoco in a single package. The Canoco program provides the analytical engine, allowing you to address specific research questions by setting-up precisely specified analyses based on constrained and unconstrained ordination methods. CanoDraw focuses more on the exploratory and presentation aspects of data analysis. It allows you not only to present the basic ordination results using the ordination diagrams, but also to penetrate deeper into your complex datasets and to explore the research hypotheses, suggested by the ordination results.

Among the frequent applications of ordination methods, there are also many cases of misuse. Correctness of using an ordination method depends on several assumptions and these can be (and should be) checked either before or after the analysis. CanoDraw is able to support you in most of those checks, offering a wide variety of variables, which can be plotted, and of regression models which can be fitted within such plots.

Work with CanoDraw program is centred on projects. A project in CanoDraw closely corresponds to a Canoco project, which represents a single ordination model fitted to your data. This model might have either implied explanatory variables (in unconstrained ordination methods, like PCA or DCA) or there could be explicit explanatory variables called, in Canoco, *environmental variables* and / or *covariables*. Canoco takes the model specification and reads the source data (in one or more data-files) and then it produces results, which are stored in the Canoco *solution file* (typically with its name including the *.sol* extension). The analysis options are stored in the Canoco *project file* (typically using the *.con* extension).

To create a new CanoDraw project, you start by selecting an existing Canoco project. At this time, the project must have been already analysed by the Canoco program and the results stored in a Canoco solution file. CanoDraw needs you to specify the location of the Canoco project file and determines from its contents the location of the Canoco results as well as the source data-files used in the analysis. The Canoco results as well as the source data provide the starting information for the CanoDraw project, allowing the creation of most of the graphs needed for an efficient exploration of your data. But you do not need to stop there. You can use additional information about your data which can be imported into the CanoDraw project or you can compare results from two alternative Canoco analyses by importing results of one of them into the other one.

When exploring the analysis results, you should start with simple ordination diagrams to obtain a better understanding of your data. You do not need to save the graphs unless you find them to be worth presenting to a wider audience. CanoDraw maintains for each project a log, where the names and contents of the **saved** graphs are listed. You are invited to supplement the text, which CanoDraw puts into the log window, by your own comments, summarising, for example, the conclusions made from the created graphs. In fact, this text can provide a basis for writing your report or research paper.

CanoDraw also stores the information about the saved graphs within the project file, so you can re-open the graphs at any time, assuming that you keep the project information up-to-date (i.e. that you save your project file each time you are asked to do so). Similarly, each CanoDraw graph file (using the *.cdg* extension) maintains a link to the CanoDraw project from which it was created. This is important because it extends the possibilities for re-establishing links between the graphs and its parental projects. You can even ask a graph document, opened in CanoDraw, to locate and open the CanoDraw project file from which it was created. The links between the graphs and the projects are needed for a more substantial exploration of your data. With such a link in place, you can, for example, see summary information about individual samples or

variables shown in an ordination diagram or to select a particular variable and ask CanoDraw to create a new graph, summarising the pattern of the variable' values over the ordination space.

After you explored the analysis results and recorded your conclusions, based on the created graphs, into a log file, you can copy it to the Windows Clipboard and paste the log into a word-processor document. The text can be then supplemented with the graphs created by CanoDraw. To do so, you can either copy the graphs to the Clipboard and paste them from there, or export them in various formats. CanoDraw supports the Windows bitmap format as well as the PNG format (accepted by all the recent WWW browsers) and it is also able to export files in vector formats, namely the Windows enhanced metafile format and the Adobe Illustrator™ format. CanoDraw creates each graph as a separate entity and does not have any facilities to combine the graphs into more complex layouts. You must use other software to combine the individual graphs into composite illustrations.

The example 14.1 guides you through the basics of CanoDraw for Windows. Before working with that example, you are advised to read at least the Chapter 11.

11. CanoDraw Concepts

This chapter introduces some important concepts necessary for understanding this guide, but also for working efficiently with the CanoDraw program. It is assumed that you have a basic understanding of ordination methods, as implemented in the Canoco for Windows software and documented in the beginning part of this manual.

11.1 Types of items in Canoco and CanoDraw projects

All ordination methods work with primary data, often representing the composition of biological communities (such as terrestrial vegetation, birds assemblages in particular habitats, invertebrate species on river bottoms, etc). Individual records of community composition (taken at various places and / or at various times) are called **samples** in Canoco and CanoDraw, although this usage clashes with the traditional meaning of this term in statistics (where sample refers to the whole collection of recorded data). The presence or abundance of organisms is recorded separately for individual categories, typically representing taxa, most often at the species level. Therefore, these categories are called **species** in Canoco and CanoDraw. A species in this meaning represents a specific kind of variable. The primary data are represented by a rectangular table where individual rows correspond to samples and individual columns to species.

The **species data** are usually supplemented with other kinds of information used to interpret their variability. In this context, we can call the species the **response variables**, and the additional variables available for individual samples can be called the **explanatory variables**. Canoco and CanoDraw distinguish at least three types of such explanatory variables:

We use **environmental variables** as the main source of information for interpreting variation in the species data. Their name derives from the fact that these variables most often describe properties of the environment in which the communities were recorded (such as water or soil properties, landscape characteristics, etc). The **constrained** ordination methods (also called canonical ordination methods) focus on summarising the variability in the species data explainable by the available environmental variables. In **unconstrained** ordination methods, the results summarise the total variability in the species data, and the effects of optionally present environmental variables can be determined *a posteriori*.

The **covariables** (often called *covariates* in other statistical software) are another kind of explanatory variable, and they differ from the environmental variables in the context of their use. Their effects upon the variability in species data are accepted and are not interesting for the particular analysis. Therefore, the variability explained by the covariables is removed ("subtracted") from the total variability and only the additional (partial) variability, not explainable by covariables, is portrayed in the ordination results. Therefore, the information about the effects of covariables is never directly shown in the ordination results and the results are believed to be free of the effects represented by the covariables. The ordination methods where covariables are used are called **partial** ordinations.

The **supplementary variables** differ from environmental variables simply by representing a secondary set of explanatory variables, in addition to the environmental variables. The supplementary variables are never used to constrain the solution of an ordination method; they are always projected *a posteriori* into the calculated ordination space. The supplementary variables are most often used as an additional set of explanatory variables in a constrained ordination method.

To simplify notation, this guide sometimes refers collectively to environmental variables, supplementary variables, and covariables as *explanatory variables*. All the various kinds of items which can be plotted in the ordination diagrams (sample scores, species scores, environmental variable scores, and supplementary variable scores) are collectively referred to as various *types of items*.

11.2 Indices of items

In an ordination analysis computed with Canoco, each sample, species or explanatory variable is uniquely identified by a whole number, called an **index** in the CanoDraw documentation and user interface (identification number is used in the Canoco documentation). These indices are unique **within** each item type. Each such numbering usually starts from value 1 and goes through an increasing sequence (2, 3,...). But Canoco allows you to work with datasets where only some members of the sequence are present (a data-file may contain samples numbered 2, 4, 6 etc., but not 1, 3, 5 etc.) and you can also remove some samples, species, or explanatory variables during the analysis. In such cases, the results provided by Canoco contain non-contiguous sequences of indices for one or more item types.

In CanoDraw, the indices are often reported along with the labels attached to items in the original data files. Note that the labels, which can be up to eight characters long (see section 4.3), do not need to be unique for different items.

The item indices are most important when you are importing additional information into a CanoDraw project. If you do not have such information available for all the items present in the analysis (e.g. ecological traits are known just for a subset of species) or if this information is available for a superset of species (taken from a data-base), CanoDraw can extract the relevant part of information from the imported data (e.g. from a Clipboard or from a Canoco data file) by matching item indices.

11.3 Window types

CanoDraw can present information about a project or about a graph in more than one window type. The window types are summarised in Table 11-1 and also illustrated in Figure 11-1 and Figure 11-2.

Both projects and graphs have one window type mandatorily connected with them. If you close the window titled *Project* <project-name>, the whole project is closed (along with any additional project-related windows, if present). Similarly, the window titled *Graph* <graph-name> is always shown for each open graph and closing it implies closing of the graph document. On the other hand, the windows containing a tree-like structure (named *Project Details* <project-name> for CanoDraw projects and *Graph Contents* <graph-name> for graphs) are displayed optionally, and their presence is governed by commands in the *View* menu.

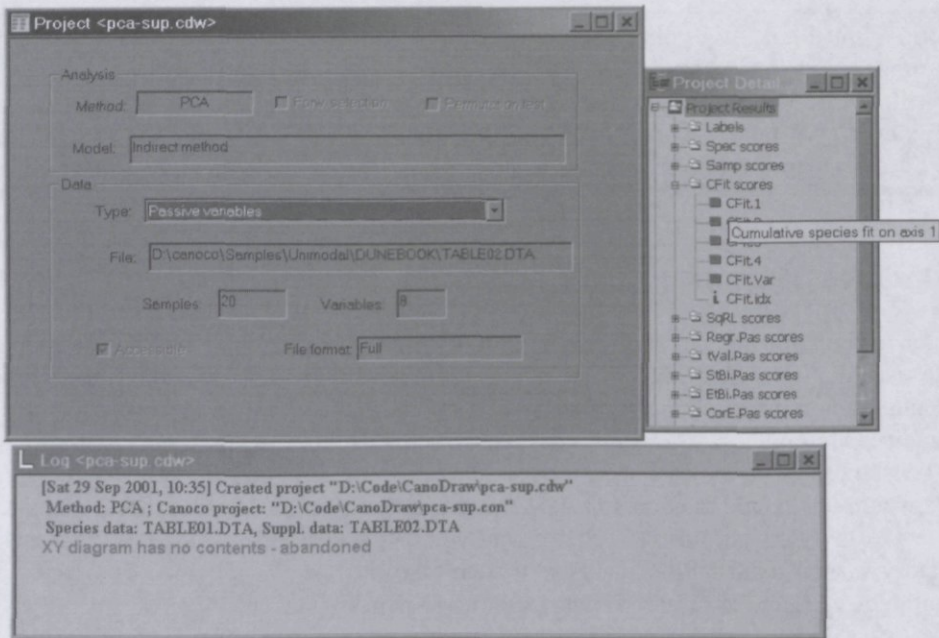


Figure 11-1 Three types of project windows

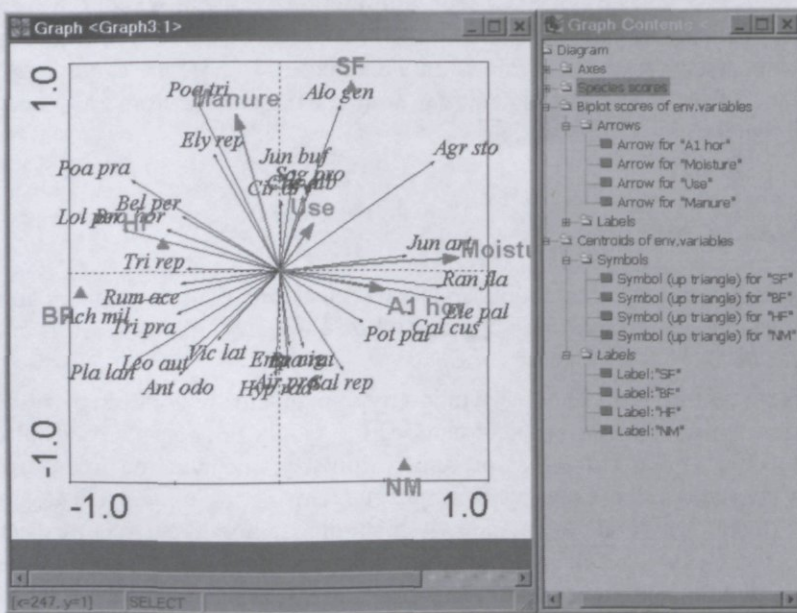


Figure 11-2 Two types of graph windows

The window titled *Log <project-name>* has a special behaviour. It cannot be closed and accumulates messages generated when new graphs are saved or problems and errors occur during the program execution. Log window contents can also persist across the sessions if you select so in the dialog invoked by *Workspace Settings* command (see section 12.3.3).

Window title	Window contents
<i>Project <X></i>	Shows summary of the ordination analysis on which this project is based; a constrained ordination is described with the notation similar to that used in software for multiple regression; the window also lists the data files used in the analysis, including the number of samples and variables
<i>Log <X></i>	Logs various kinds of messages: problems encountered while creating graphs, particularly when fitting regression models and calculating species data statistics; errors occurring while saving files; records about the name and contents of graphs saved to a permanent storage; etc.
<i>Project Details <X></i>	<p>Lists the variables available in a project. There are three main folders in the upper level of the hierarchically structured contents:</p> <ul style="list-style-type: none"> * <i>Project Results</i> folder contains variables with item scores and statistics, extracted from Canoco™ solution file, and arranged according their placement in individual solution file sections. These are usually available for all items of the particular type and some of them are available separately for individual ordination axes * <i>Source Data</i> folder contains variables available in the original Canoco™ data files. They are placed into the folder corresponding to individual data file types (<i>Species Data, Environmental Data, Passive Data</i>) * <i>Imported Variables</i> folder contains variables imported later into this CanoDraw project. They are placed into one or more subfolders, depending on which type of items they refer to <p>You can click on any of the terminal items (corresponding to one variable) with right mouse button to display the <i>Variable Summary</i> floating dialog (see section 13.5 for its description)</p>
<i>Graph <Y></i>	Shows the graph as it will appear when printed (except it is always shown in full colour here). You can change the magnification to see just a part of the graph in higher resolution and you can directly manipulate the graph contents here
<i>Graph Contents <Y></i>	Presents the graph contents in hierarchical manner, displaying organisation of its structure. In the window illustrated in Figure 11-2, we can see that the graph contains both species scores and the scores of environmental variables, the latter represented by biplot scores or by centroids. This window type supports only partial manipulation of graph contents: you can perform several types of selection of graph objects (direct selection and <i>Select Suchlike</i> and <i>Select Similar</i> commands), you can lock selected objects, and you can edit visual attributes of selected objects. Using this tree view, you can simply select a group of objects (by selecting a non-terminal tree item) and change its visual attributes. Selection of objects in this window is independent of the selection in the standard graph view.

Table 11-1 Contents of individual window types

Beside the window types described in the above table, CanoDraw can also display three types of floating windows described elsewhere: the *Properties* window that can be used to change the visual attributes of selected graph objects (see section 12.3.4), and the two floating windows summarising the values of the active variable (*Variable Summary* window) or the values of the active sample (*Summary of sample* window) – see section 13.5 for their description.

11.4 Graph types

Graphs produced by CanoDraw can be roughly classified into three types:

- * *Ordination diagrams* display a two-dimensional projection of the ordination space calculated during the analysis performed by the Canoco program. Ordination diagrams most often

attempt to summarise the original species (primary) data or their relation to explanatory variables. Contents of the ordination diagrams can be interpreted to provide an optimal approximation of either the original data or secondary tables derived from such data (e.g. matrix of correlations among the species, matrix of weighted averages of species to individual explanatory variables, etc.) – see Ter Braak (1994) and Ter Braak & Verdonschot (1995) for more details. CanoDraw for Windows optionally provides hints concerning the correct interpretation of a particular ordination diagram (see section 13.5 for description). Ordination diagrams are created with commands located in the upper part of the *Create* menu.

- * *XY and XYZ diagrams* display a joint distribution of values of two or three variables either in form of scatter plots (with points displayed at the coordinates representing the values of X and Y variables) or by displaying the fitted regression curves (representing the relation between the response variable Y and the explanatory variable X) or regression surfaces (representing the relation between the response variable Z – the attribute variable – and two explanatory variables X and Y). In both cases, CanoDraw provides a choice among three families of regression models: generalized linear models (including the traditional linear regression models), generalized additive models, and regression models based on the loess smoother. In XY diagrams, there can be more than one response (Y) variable at the same time, but such a display is typically useful only if the fitted regression models, not the original scatters of points are shown. See section 12.5.5.3 for more details about creating XY and XYZ diagrams, and section 13.6 for additional discussion about fitting regression models.
- * *Ordination-based attribute plots* stand halfway between the previous two categories. They show the patterns of values of a selected variable, an **attribute**, in the positions corresponding to the sample (or species, or environmental variable) scores in the ordination space. You can use as an attribute either an original variable from the source data, an imported variable, or any variable present in the file with the Canoco results. The visualisation of the attribute values can be done either by varying the size of symbols representing individual plotted items (usually the samples), or by fitting a regression model (generalized linear model, generalized additive model, or loess smoother model) using the horizontal and vertical axis as two predictors and the attribute as a response variable. In the latter case, the pattern of attribute values across the ordination space is shown as contour plot, representing the fitted "response surface". The ordination-based attribute plots are created using the menu commands *Data Attribute Plot* and *Results Attribute Plot*. See sections 12.5.5.1 and 12.5.5.2 for additional descriptions.

11.5 Graph object types

The contents of each CanoDraw graph can be viewed from two different points. First, you can concentrate on the **meaning of the graph contents** and see, for example, an ordination diagram with species and samples as presenting information about the changes of expected values of individual species along the ordination axes, about the expected co-occurrence of species in samples, and about the similarity of species composition among the samples. Alternatively, you can focus more on the **visual aspects of the graph**. In that case, you want to differentiate among the symbols representing individual samples and the labels connected to them, you are concerned with the length of tickmarks on the plotted axes, and you want the envelopes enclosing samples from different classes to be easily distinguishable by the viewer.

Certainly, the first view is the one of ultimate importance, but you need to pay attention also to more formal aspects of your graphs, to facilitate an easy interpretation of their contents. When

we consider a CanoDraw graph from the point of view of its visual attributes, we work with individual **graph objects**. Often, a single plotted item (such as one species or one sample) is represented by at least two graph objects (e.g. species arrow and its label), and additional graph objects combine into graph axes, legend, etc. CanoDraw graphs may contain up to seven different kinds of graph objects, listed in Table 11-2, together with a summary of their main use.

Graph object type	How this type is used in CanoDraw graphs
labels	to label individual items plotted in graphs (species, samples, explanatory variables) to label ordination axes – either at minimum and maximum values or at all the tickmark positions to name individual categories differentiated in graph legend
lines	to draw the coordinate system of each graph: the scale (axis) lines and the tickmark lines; the tickmarks can be transformed to a reference grid within the graph
arrows	to plot ordination items representing vectors, not positions (quantitative explanatory variables, species in linear ordination methods)
polylines	to represent isoline contours in contour diagrams to represent one or more ordered series of samples or species to visualise the area occupied by items of particular class (using <i>envelopes</i>)
symbols	to represent ordination items corresponding to positions in the diagram space
pie-symbols	to provide special representation for ordination items corresponding to sample or species positions in the diagram space: in addition to marking item location, the pie symbols visualise distribution of classes of complementary items (section 12.4.1.1)
rectangles (bars)	used only to allow manipulation of graph legend area

Table 11-2 Graph object types in CanoDraw graphs

11.6 Graph scaling and coordinate units

The size of a CanoDraw graph depends on the currently specified properties of the output media. The media properties (size and orientation of the output page) are selected in the *Print Setup* dialog. If the currently active window in CanoDraw corresponds to a project or there is no window opened in the CanoDraw workspace, the output media properties are used as defaults for all the newly opened or created graphs. If, on the other hand, the currently active window corresponds to a graph, then the page size and orientation is set just for this graph. In this way, you can change the original layout of the graph with respect to the output media. Note, however, that this change is **not** persistent across the CanoDraw sessions. When you open an existing CanoDraw graph file again, the graph is scaled to fit as well as possible onto the currently active output media format. This is the result of the way CanoDraw constructs the graphs: they are scaled so that they fit best into a *virtual coordinate space*.

The *virtual coordinate space* used by CanoDraw is a two-dimensional drawing space with isomorphic scaling in the horizontal and vertical directions (the same physical distance corresponds to an identical difference in coordinate space units in both directions). The main graph contents (all the plotted data items as well as the graph axes and their labels) are always fitted into the area spanning from coordinates [0,0] at the lower left corner to the coordinates [1,1] at the upper right corner. The graph may not (and usually does not) fill this whole area, unless it has the unit aspect ratio (i.e. the same height and width). The labels can reach over this unit area of the virtual coordinate space. For example, if a label is adjusted to be on the right side of a point with virtual coordinates $[x=0.990, y=0.5]$, its right edge is likely to reach over the

value of 1.0 in the horizontal direction. If a graph contains a legend, the legend area is placed on one of the edges of the output media, depending on user-specified settings (see section 12.4.1.2). If these settings specify the placement of the legend on the left or bottom edge of the output page, the graph is shifted correspondingly towards the right or upwards, so its right edge or top edge can reach over the coordinate value 1.0. Similarly, if the legend is placed on the right or top edge of the output page, the right or top edge of the legend frame is likely to reach over the coordinate value of 1.0.

The longer side of the main graph area maintains its length in virtual coordinate units equal to 1.0 minus the specified Outer Graph Margins (see section 12.3.1.2). The unit length in the virtual coordinates provides a reference for specifying the size of symbols or fonts, width of lines, and other dimensions in various places where CanoDraw options might be set.

When displaying or printing a graph, CanoDraw scales the original unit rectangle (extended on its right and / or the upper side if any graphs objects exceed the unit coordinate) into the printable area of the output media.

The actual virtual coordinates of the mouse pointer are displayed in the graph window status bar.

11.7 Limiting contents of graphs

If you analyse large data sets, with hundreds of variables and / or of samples or if you ask specific questions concerning, for example, just a subset of species with specific ecological properties, you might like to restrict the set of items displayed in a graph.

CanoDraw provides several methods for limiting the set of plotted items. The primary method is the selection of plotted items using their properties. There are several criteria, which can be active for a particular type of items (e.g. species) at the same time. These criteria are called **inclusion rules**. The primary inclusion rules are limited to the most important criteria (e.g. fit of species or samples into ordination space, see section 12.4.1.3, or correlation of explanatory variables with ordination axes, see section 12.4.1.4). But you can, in the case of sample and species, limit the plotted items to members of a group of items. Such a group can be selected based on a very wide range of criteria (e.g. group of all samples with an abundance of selected species above a certain limit or a group of species with a specified range of positions on the third ordination axis) and these criteria can be combined together (by creating a new group either by intersection or union of two existing groups). This allows you to define very complex criteria for the appearance of samples or species in an ordination diagram or any other type of CanoDraw graph.

Alternatively, you can override the inclusion rules by specifying directly which items should never be plotted (see section 12.4.6) or which items should be always plotted (see section 12.4.7).

Note that limitations imposed on item appearance in graphs are also used to select observations used to fit regression models in the attribute plots.

11.8 Application-wide and project-specific options

CanoDraw works as a "state machine". If you change the options concerning the creation of graphs, then any new graphs created from this time on are governed by the changed set of

options. Any existing graphs are not updated with the changed options, unless you explicitly request so (using the *Recreate graph* command, see section 12.5.6).

The options concerning the graph contents and appearance are separated into two layers:

- * **Application-wide options** are used independently of which CanoDraw project you are currently working with. These options are persistent across the individual sessions: they are loaded from the Windows registry (for Windows NT 4.0, Windows 2000, and Windows XP) or from the configuration file (for Windows 98 and Windows Me) and saved again at the end of each CanoDraw session. You can additionally store and re-load "snapshots" of the current visual graph settings, using the two commands in the *Visual Attributes* submenu in the *Files* menu. Application-wide options can be changed using the commands in the upper part of the *View* submenu.
- * **Project-specific options** are stored in the CanoDraw project files and represent all the choices, which are supposed to depend on the actual analysis properties. These options include classifications of items, definitions of series collections, or the information about which items should be explicitly excluded from the plots, and also the choices like using pie-symbols or plotting envelopes around the classes of items. The project-specific options can be modified from the dialog displayed by the *Settings* command in the *Project* menu.

There is little dependence between the options in these two layers, with just one exception. CanoDraw records the latest changes to project-specific options concerning the presence of a legend in the diagrams and the legend position on one of the output media edges, and defaults to those values when initially setting-up a new CanoDraw project.

12. Commands Reference

CanoDraw allows one to work with two types of documents:

- * **CanoDraw projects** have a one-to-one correspondence to Canoco projects and are initially created by importing a Canoco project file into the CanoDraw program (using *File / New Project* command). CanoDraw projects are stored in files with extension *cdw* and contain ordination results (imported from a Canoco solution file), original data files (“species, environmental, and covariable data” in Canoco terminology), and the project-specific settings (like the rules determining which variables should be plotted in the diagrams).
- * **CanoDraw graphs** represent the individual graphs (ordination diagrams, attribute plots, XY plots) produced with CanoDraw using information from a CanoDraw project. Graphs are stored in files with *cdg* extension and contain information linking them to their parental projects. So even if the graphs are opened independently of their parental project, they are attached to this project if it is opened. Similarly, each project keeps track of the graphs created from it, if they were saved. When opening a CanoDraw project file, a list of related graphs is optionally offered, so they can be opened alongside the project.

Each graph or project is represented by one or more windows within the CanoDraw application. At any time, only one window is active within the CanoDraw workspace and its identity (i.e. whether it belongs to a project or to an individual graph) determines which commands are available from the CanoDraw menu. Most menu commands are shared, but there are also some distinct differences, marked in the following text using this icon for menu commands available only for projects: **P** and this icon for commands available only for graphs: **G**.

12.1 File

Commands in this menu provide for the opening and saving of projects and graphs, exporting graphs in other file formats, storing “visual style” configuration, printing graphs, and closing the application.

12.1.1 New Project

Use this command to define a new CanoDraw project. Keyboard shortcut for this command is *Ctrl+N*. Each new project is based on an existing Canoco project. Therefore, you must start with locating the file representing such project. Canoco project file names usually have a *con* extension and provide sufficient information for CanoDraw to find both the source data files (having extensions like *dat*, *env*, *spe*, *cep*) and also the file with the analysis results (typically using the *sol* extension), which was produced by the Canoco program.

If the other files expected during the definition of project (data files and *.sol* file) cannot be found in the paths specified within the Canoco project, CanoDraw attempts to find them in the same directory where the *con* file (Canoco project file) is located. If found there, the user is asked to confirm the appropriateness of using the file found. This is a feature useful in situations where the location of a Canoco project was changed (e.g. by moving the files from one computer to another). Note that this feature works only if all the concerned files are located in the same directory when the analysis with the Canoco program was performed.

To create a new CanoDraw project, you need only the Canoco project (*.con*) file and the file with results (*.sol* file). CanoDraw is able to define a new project even if the original source data

are not available. Nevertheless, the absence of the source data limits the set of graphs that can be created in the CanoDraw program.

The source data files must be found by CanoDraw during the project setup to enable the full range of CanoDraw diagrams. They can be imported later, but then the range of graphs where they can be used (as “imported” variables) is more limited.

After the Canoco project file was specified, CanoDraw attempts to locate the related files, parses the analysis options as well as the source data, and - if successful - asks for the file name under which the new CanoDraw project should be saved. The suggested name for the CanoDraw project file is the same as the one used for the original Canoco project file, except the .con extension is changed to .cdw extension. Also, CanoDraw suggests placing the new project file in the same directory where the Canoco project file is located.

The dialog asking you to select a name for the new CanoDraw project follows immediately after the dialog box where you had to specify the source Canoco project file.

12.1.2 Open Project

Use this command to open an existing CanoDraw project. Keyboard shortcut for this command is *Ctrl+O*. File Open dialog appears, as shown in Figure 12-1.

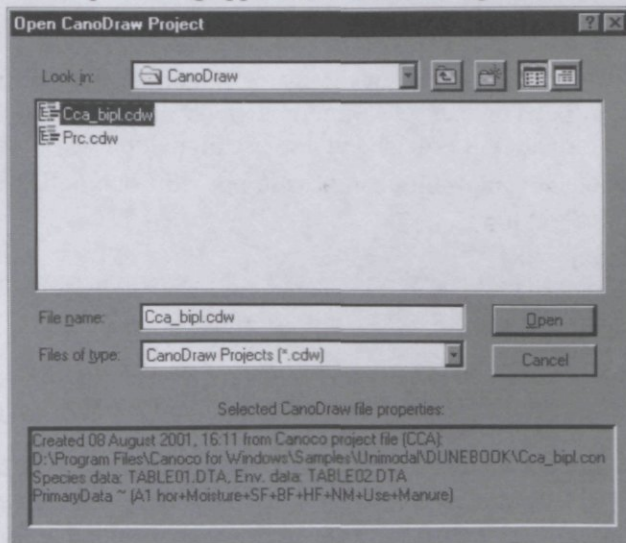


Figure 12-1 Open CanoDraw Project dialog box

Note that as you select a valid project file, a summary description of the project contents is shown at the dialog bottom. This summary box shows the project creation date and time, type of ordination analysis, name of the source Canoco project, names of the data files, and, for a constrained analysis, also the ordination model specification (explanatory variables and covariables). Only one project can be selected in this dialog, but several projects may be open at the same time in the CanoDraw program workspace.

Recently used CanoDraw project files can be also opened from the list of files at the bottom of the *File* menu.

12.1.3 Open Graph

Use this command to open an existing CanoDraw graph file. Keyboard shortcut for this command is *Ctrl+G*.

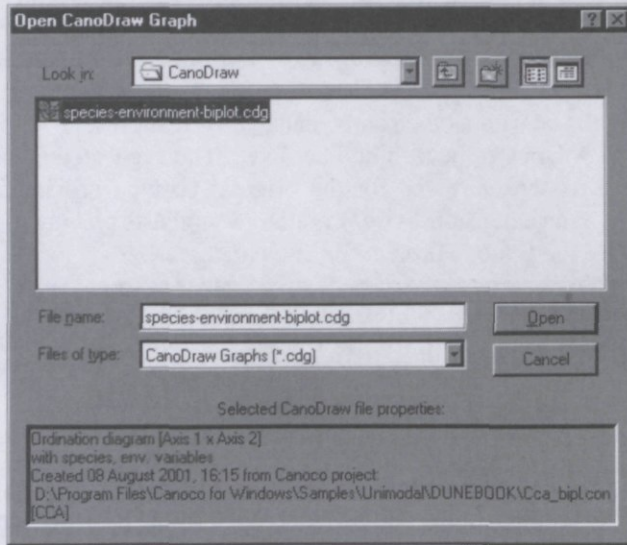


Figure 12-2 Open CanoDraw Graph dialog box

Note from Figure 12-2 that the box at the dialog bottom summarises the contents of the currently selected graph file. You can select multiple graph files at the same time in the dialog box, but no summary information is provided in the case of multiple selection.

Recently used CanoDraw graph files can be also opened from the list of files in the lower part of the *File* menu.

12.1.4 Close

Closes the window currently active in the CanoDraw workspace. If this is the last window for a particular graph or project, the document is closed, so you might be asked about saving any outstanding, non-recorded changes of the graph or project.

Note that a short-cut for accelerated closing of graphs of a particular project (optionally discarding their changes without prompt) is available from the Windows submenu.

12.1.5 Save

Saves the document corresponding to the currently active window in the CanoDraw workspace. Keyboard shortcut for this command is *Ctrl+S*. If the document was already saved before, it has a file name assigned and you are not asked about it. Otherwise the following command, *Save As*, is executed.

12.1.6 Save As

Saves the document corresponding to the currently active window, allowing you to specify a new name for it. The starting name is identical with the file name currently attached to the document. After CanoDraw successfully saves the document under the new name, it closes the older file (if it exists), without changing its contents, and then continues working with the new copy of the document.

12.1.7 Export G

These commands store the currently active CanoDraw graph in formats accepted by various graphical packages as well as word processor software. Note that this is an export action: it does not change the name of the document being exported and does not update the document file. It simply stores the current document state in a new file, which has a format of Windows bitmap, Windows metafile, Adobe Illustrator, or PNG format.

12.1.7.1 Bitmap G

Exports the active graph in the format of Windows bitmap. This is a common format for storing raster graphics on Microsoft Windows® operating systems. The target physical size of the image is based on the output page dimensions of the active printer format (see section 12.1.11 below) and the output resolution you specify in this dialog (see Figure 12-3). The image resolution is in DPI (dots-per-inch) units. For example, if your output page format has an approximate size of 8.5 x 11 inches like the **Letter** page format has, the dimensions of an image which would occupy a whole page are 850 x 1100 points when 100 DPI resolution is selected. Note however, that CanoDraw works with the **printable** area of the selected output media, which is always smaller than the target paper sheet. The image almost never fills the whole page, being constrained by its aspect ratio. The estimated physical size of the resulting bitmap image is displayed in the two *Image size* fields at the bottom of the dialog.

You can also specify *Color depth* (color resolution) of the target image.

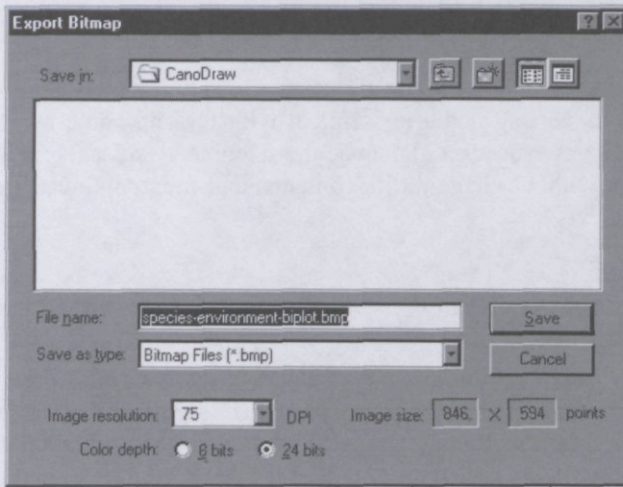


Figure 12-3 Dialog for exporting graph in bitmap format

12.1.7.2 Metafile G

CanoDraw supports the export of graphs in the metafile format. This format describes image contents in terms of graphical operations, not by specifying color of individual image dots. The older versions of Microsoft Windows® operating systems (16-bit versions) introduced the **Windows metafile format**, which was deficient in many respects, so it was replaced by so-called **Enhanced metafile format**, a more device-independent, more scalable vector format.

CanoDraw supports both these metafile formats, the older one under the name of *Placeable metafile*, which is an extension of the original Windows metafile format, proposed by the Aldus company for use with their Aldus Pagemaker™ desktop publishing software. The place-able metafile format is provided to support some legacy publishing software but use of the *Enhanced metafile* format is recommended under other circumstances.

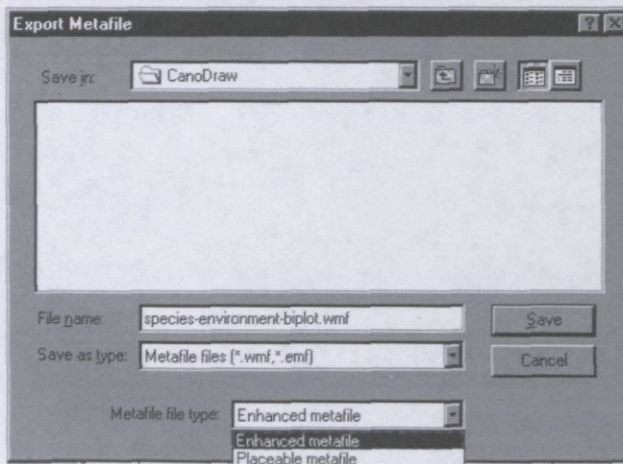


Figure 12-4 Dialog for exporting graph in metafile format

The Microsoft Word for Windows™ editor readily accepts enhanced metafile format, but other programs (like Adobe Illustrator® program) have problems with the proper import of labels depending on their alignment in respect to the labelled point. For Adobe Illustrator® or

Adobe Photoshop® format, export of the image in the Adobe Illustrator format is therefore recommended.

12.1.7.3 Adobe Illustrator **G**

CanoDraw exports graphs in the Adobe Illustrator™ version 3.0 format. This format specification is readily accepted by the newest versions of both Adobe Illustrator and Adobe Photoshop software. Colors specified in the CanoDraw program are transformed into CMYK color space and fonts used for the graph' labels are replaced with the similar PostScript fonts as specified in Table 12-1.

CanoDraw font name contains word	PostScript font
<i>Times or Garamond</i>	Times-Roman (Times-Italic, Times-Bold, and Times-BoldItalic)
<i>Courier</i>	Courier (Courier-Oblique, Courier-Bold, and Courier-BoldOblique)
<i>Symbol</i>	Symbol
All other fonts	Helvetica (Helvetica-Oblique, Helvetica-Bold, and Helvetica-BoldOblique)

Table 12-1 Transformation of fonts used by CanoDraw (TrueType™ fonts) into Adobe PostScript™ fonts. Typeface variants in parentheses correspond to italics, bold, or bold-italics font styles, respectively.

Additionally, the characters in the labels are transformed during export to Adobe Illustrator™ format files, so that only those from the ASCII character set are retained. The other ones are replaced by a question-mark symbol.

12.1.7.4 PNG **G**

PNG (Portable Network Graphics) format is a raster graphics format (like the Windows Bitmap format), conceptually most similar to GIF format. It is accepted by the recent versions of all mainstream graph editors and WWW browsers, so it can be used to create graphics for inclusion in WWW pages.

The dialog for exporting CanoDraw graphs in PNG format (not shown here) is similar to the dialog used for export in Windows Bitmap format, except the color depth cannot be specified and is always set to 24 bits.

12.1.8 Visual Attributes **G**

The two commands in this submenu allow you to store permanently application-wide settings influencing the look of graphs created with CanoDraw. As explained in section 11.8, CanoDraw makes the distinction between settings specific for individual projects and settings shared by all the projects. The latter ones can be changed using the dialogs invoked by the commands in the *View* menu (see sections 12.3.1 to 12.3.3). The current state of these settings is stored implicitly when you close the CanoDraw application in the Windows Registry and retrieved again when you start the program. But you can take their snapshot at any time, using the *Save* command in this submenu and replace the current settings with those stored in such snapshot (using the *Load* command).

12.1.8.1 Load

Retrieves the stored Visual Attributes settings from a file.

12.1.8.2 Save

Stores the settings, which are normally modified in the dialog shown by the *View / Visual Attributes* (see section 12.3.2) or *View / Diagram settings* (see section 12.3.1) commands. From the latter, only part of the settings is stored:

- * All the settings from the *Properties 2* page
- * From page *Properties 1* following settings are stored: *Rescale sample or species scores to optimality*, *Limit range of fitted values by extent ...*, *Plot also the extrapolated values ...*, *Apply additional smoothing of contour lines* and the two values in *SMOOTHER SETTINGS* section
- * Additionally, while legend plotting options are specific for individual projects, CanoDraw uses the latest settings of two legend options as defaults for newly created CanoDraw projects: the actual presence or absence of a legend in diagrams and its position on one of the page edges (left, right, top, or bottom).

Visual Attributes are stored in a proprietary file format, using the *cds* extension. This extension, unlike *cdw* or *cdg*, is **not** registered with the operating system and therefore does not have any specific icon in the Windows Explorer™.

12.1.9 Print **G**

This command displays the Print dialog (its look varies across the operating systems) and prints the active graph. In the Print dialog, you can change the destination printer and printing options for this particular print job.

12.1.10 Print Preview **G**

Allows you to preview the graph appearance on a printed page. Note however, that even the standard graph window content is very similar to the output you might expect with the current print settings.

12.1.11 Print Setup

The Print Setup command utility reaches far beyond influencing the appearance of a printed page. The output (printed) page metaphor stands behind the logic of scaling and measuring the size of each CanoDraw graph. Therefore, CanoDraw is not able to work appropriately without at least one printer being installed on the system.

When starting, CanoDraw takes the specification of the size of the default printer page to obtain default dimensions of the canvas space, available for the creation of graphs. CanoDraw changes the default output settings only in one respect – it changes the page orientation to **landscape** mode, where page width is greater than page height. This change is performed because the ordination diagrams very often have the lower-order axis longer than the higher-order one (e.g. the first axis is often longer than the second or third one).

If you have no window open in the CanoDraw workspace or the active window corresponds to a project document (not to a graph), the changes to the printer settings affect the CanoDraw defaults, which are then used when creating new graphs or re-opening existing ones. But the print settings are also maintained per graph (while the graph is open in CanoDraw), so you can have one graph fitted onto a page oriented in landscape mode, while another is adjusted to fit on the page with portrait orientation. You can change the output page orientation for a particular graph by selecting the graph window before using the *Print Setup* command. Note that print settings are not persistent. If you close a graph and reopen it again, the graph window adopts the current output page size and orientation. This is because the output page orientation is not really a property of the diagram: the aspect ratio (ratio of physical height to physical width of a graph) does not change with the output page size or orientation – only the physical size of the graph is adjusted.

12.1.12 Recently used files

The recently used graph and project files are listed at the bottom of *File* menu, just above the *Exit* command. If you select a file name from this list, the corresponding document is either opened in the CanoDraw workspace or – if it is open – its window is brought forward.

12.1.13 Exit

Closes CanoDraw program. If there are any modified graphs or projects that were not saved, you are asked whether to save them.

12.2 Edit

This menu provides general editing commands available for whole graphs, individual objects in the graphs, and also for the contents of Log views of the projects. Specialised commands for selecting and modifying objects within graphs are available from the *Object* menu (see section 12.6).

12.2.1 Undo

For changes of graph objects, CanoDraw tracks the last 32 changes to label positions, diagram contents (modified by adding new objects or deleting existing ones), label text, or visual attributes of any object within the graph. These changes can be undone in stepwise manner, starting from the most recent change. The changes, which are undone with this command, can eventually be re-done again using the *Redo* command (see 12.2.2). Keyboard shortcut for this command is *Ctrl+Z*.

Note that changes to labels' orientation (horizontal or vertical) cannot be undone here. Nevertheless, the orientation has just two states and can be easily flipped back by applying the *Make label horizontal / vertical* command again.

The *Undo* command is also available with the Log view of a project, but there only the last change can be undone and the meaning of the *Undo* command changes to "redo" after the last action was undone.

12.2.2 Redo **G**

Brings the graph contents state one step forward in the sequence of actions which were applied to this contents and were then undone. Keyboard shortcut for this command is *Ctrl+Y*.

12.2.3 Cut **P**

This command is available only if the active window represents a Log view of the CanoDraw project. It copies the selected text on the Clipboard and then removes it from the log. Keyboard shortcut for this command is *Ctrl+X*.

12.2.4 Copy

If the active window shows a graph, the whole graph is copied onto the Clipboard in two formats:

- * Windows bitmap format with a size corresponding to the selected output print page dimensions and resolution of 120 DPI. Color depth is compatible with the graphics mode used on the display. Together with the bitmap, a color palette is placed on the Clipboard, so a more precise color management can be performed in the program where the bitmap will be pasted
- * Enhanced metafile format

Copy command is also available with the Log view of currently active project, where it copies currently selected text onto the Windows Clipboard. Keyboard shortcut for this command is *Ctrl+C*.

12.2.5 Paste **P**

This command is available only for a Log view of a CanoDraw project, where it inserts text available on the Windows Clipboard. Keyboard shortcut for this command is *Ctrl+V*.

12.2.6 Delete

Available for graph windows, where it deletes all **selected** objects within the graph, and also for project Log windows, where the currently selected text is removed. Keyboard shortcut for this command is *Delete*.

12.2.7 Change text **G**

This command is available only if a single label object is selected in the graph window or in the Graph Contents (tree-like) view. A dialog is displayed where you can change the label text. This command is also available from the context sensitive pop-up menu invoked by right-clicking a **selected** label object.

12.2.8 Make label vertical / horizontal **G**

This command is available only if one or more label objects are selected in the active graph window. It flips the selected label(s) orientation between horizontal and vertical, choosing the most compatible alignment settings with respect to the existing alignment point. This command is able to flip the horizontal / vertical orientation setting correctly even if a mixture of horizontal and vertical labels is selected. The actual text of this menu item depends on whether multiple labels are selected (*Rotate selected labels*) or whether the single selected label currently has a horizontal or vertical orientation (*Make label vertical* or *Make label horizontal*).

This command is also available from the context sensitive pop-up menu invoked by right-clicking a **selected** label object (or one of several selected labels). Unlike the preceding command, the change of label orientation is not possible from the Graph Contents (tree-like) window.

12.2.9 Copy labels to Clipboard

This command copies all the object labels, contained in the currently active diagram onto the Clipboard in a text format. Labels of items representing separate categories (e.g. labels of species vs. labels of environmental variables, in a biplot diagram) are separated into individual paragraphs. Items within each paragraph are separated by a comma.

12.3 View

This menu contains commands for displaying dialogs where application-wide (project-independent) settings can be inspected and changed, as well as commands specific for the currently active window type (graph-related or project-related).

12.3.1 Diagram Settings

This command displays a tabbed dialog (property sheet), with five tabs (pages).

The options in the first two pages (*Properties 1* and *Properties 2*) affect the program behaviour during the creation of graphs and also some aspects of the graph appearance, which are project-independent (i.e. reflecting more the user preferences than the particular project properties). CanoDraw uses the actual values of these options at the time a graph is being created to decide about its contents. A later change of the options does not update the contents of already existing graphs. These graphs can be updated explicitly, after the options were changed, using the *Recreate graph* command available either from the *Create* menu (see section 12.5.6) or directly from the context sensitive pop-up menu.

The other three tabs specify default settings for the three families of regression models available in CanoDraw. If the value of **Offer approval of regression model settings ...** is **on** (checked) in the *Properties 1* page (described below), a dialog displaying these default values for the given type of regression model is shown immediately before the model is fitted, so that you can customise the model settings individually for each fitted model.

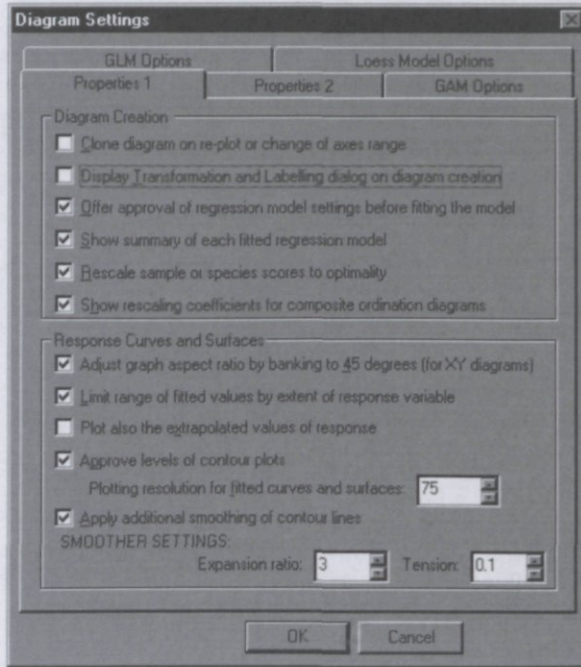


Figure 12-5 Properties 1 dialog page

Clone diagram on re-plot or change of axes range

After you change the application-wide settings (as described in sections 12.3.1 and 12.3.2) or the project-specific settings (see section 12.4.1), individual graphs can be re-created using the *Recreate graph* command (see section 12.5.6), so the recent settings are applied to their contents. Also, the range of diagram axes can be changed using the *Range of axes* command available from the context sensitive pop-up menu. In both cases, a new version of the existing graph contents is created and this option determines whether the original graph contents is retained and the new one is created as a separate entity (if this option is **checked**) or whether the original graph contents is replaced by the new one. Cloning of the graph contents is particularly useful when you want to experiment with the settings and compare results of different settings values.

Display Transformation and Labelling dialog on diagram creation

The *Labels and Transformations* dialog provides a final adjustment of the values displayed in a diagram and is illustrated in the following figure. The dialog is shown there in the context of creating a XY diagram, with moisture values plotted in the horizontal direction and samples diversity index plotted on the vertical axis. The bottom dialog area, used to adjust the values of an attribute acting as a response variable in XYZ diagrams, is disabled here. As you can see from the illustration, this diagram serves three different tasks:

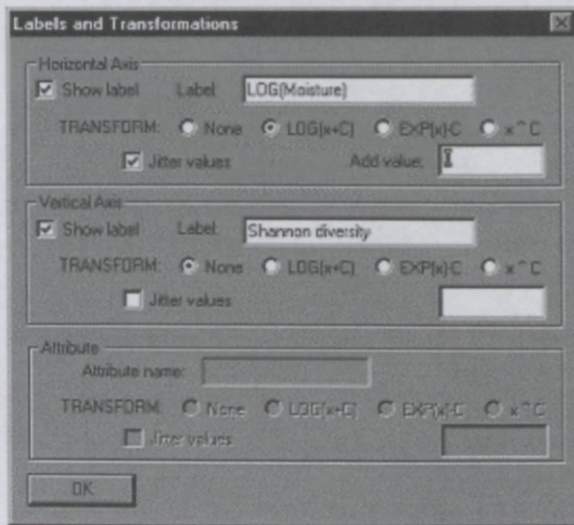


Figure 12-6 Labels and Transformations dialog box

- * Supplementary axis labels can be displayed. For XY(Z) diagrams, the labels' text defaults to the names of the plotted variables, and *Ordination axis X* is the default text for ordination diagrams, where **the axis number replaces X**.
- * Selective parametric transformation of plotted variable values can be selected. Logarithmic transformation, exponential transformation, and power transformation are available, with the option to adjust the additional parameter of the transformation functions. Note that in the dialog displayed above, the default axis label was automatically adjusted by CanoDraw, to provide information about the selected logarithmic transformation.
- * Values plotted along a particular axis can be **jittered**. Jittering means adding random noise to true values, with a sufficiently small extent so the distributional patterns are not washed-out, but large enough to identify overlap of multiple observations in the diagram. Such an overlap typically occurs when variables with a limited set of values are plotted, as illustrated in Figure 12-7.

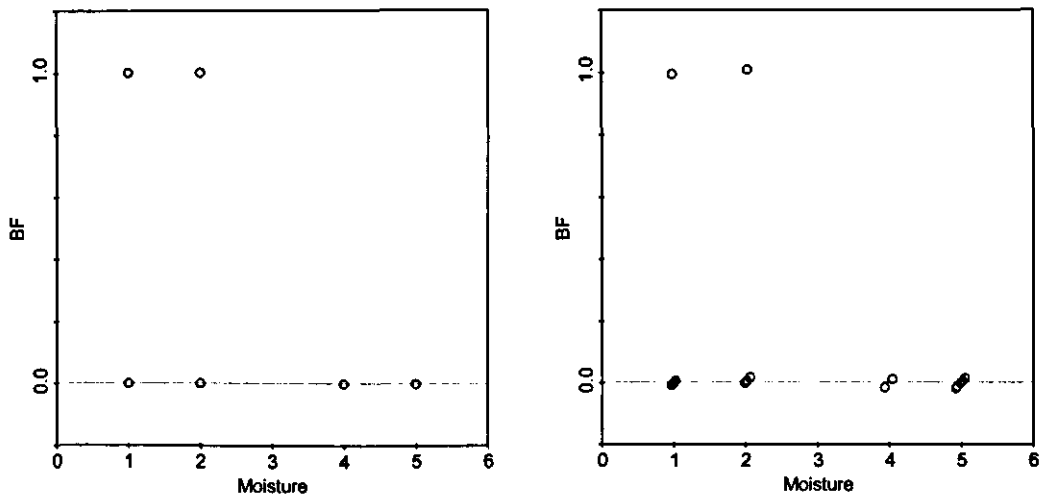


Figure 12-7 Effect of jittering on a XY graph where there is substantial overlap of points

The right-hand graph plots the same data as the one on the left side, but both the X and Y variables were jittered.

CanoDraw applies jittering by adding pseudo-random value obtained from a uniform distribution over the interval $(-Z, +Z)$ where Z is equal to 2% of the variable range. For example, for the variable *BF* plotted in Figure 12-7, with the range of values from 0 to 1, the added values are randomly drawn from an interval $(-0.02, +0.02)$, with an equal-selection probability throughout the whole range.

In an ordination diagram, the options for parametric transformations and for jittering of plotted values are disabled.

Offer approval of regression model settings before fitting the model

If this option is **enabled** (checked), CanoDraw displays a dialog with options for the regression model (generalized linear model – GLM, generalized additive model – GAM, or loess smoother model) immediately before fitting, so that the actual model parameters or the method used to select those parameters may be specified. If this option is not checked, the values specified in the last three pages of this dialog (see sections 12.3.1.3, 12.3.1.4, and 12.3.1.5) are implicitly used.

Note that this option value is ignored when fitting multiple species response curves using *Create / Attribute plots / Species response curves*, where the shared specification of regression model is set for all selected species, using one dialog.

Show summary of each fitted regression model

Each fitted model is summarised with a dialog illustrated in Figure 12-8. The actual dialog content varies with the regression model type.

Fitted Generalized Linear Model

Response variable:

Predictor(s):

Distribution: Link function:

Null model deviance: with residual DFs

Fitted model deviance: with residual DFs

Model significance: F = P = AIC =

Unimodal response curve:

Optimum: S.E.: Conf. interval:

Tolerance: S.E.: Max. value:

Regression coefficients

Model Term	B	s.e.	T
(Intercept)	1.09234	0.216751	5.03963
Samp.1	-0.78195	0.298081	-2.62328
(Samp.1)^2	-0.62254	0.314117	-1.98187

Figure 12-8 Summary dialog for a fitted generalized linear model

The names of response variable and predictor(s) are shown, as well as the most important parts of the model specification. Variability explained by the fitted model is compared with the variability explained by the null model and also additional important information is provided (here, for a GLM, the table of regression coefficients estimates is shown, and also information about unimodal response curve properties, specific to this second-order polynomial model).

All the regression summary dialogs have a *Copy* button. Click it to place the model summary in a text format onto the Windows Clipboard.

The *OK* button closes the dialog and lets CanoDraw proceed with fitting the next regression model (if there are multiple models involved) or with creation of the diagram. If you select the *Skip* button instead, this particular model is **not** included into the diagram being created. If this is the only regression model involved or all the models were “skipped”, diagram creation is cancelled. Otherwise, only a subset of potential diagram content is plotted.

Rescale sample or species scores to optimality

During the setup of a Canoco analysis, you have to select the option for scaling of ordination scores. There are two alternative scalings, providing better approximation – based on the ordination diagrams – to either inter-sample dissimilarities or to (dis-)similarities or correlations among the species. In the former case, you are focusing on inter-sample distances, in the latter case either on “inter-species distances” (the term used in unimodal ordination methods) or on “inter-species correlations” (for linear ordination methods). Selecting one of the two options results in ordination scores where one of the two entities (sample or species scores) are spread across the ordination space in stronger accordance with the inter-sample dissimilarities or interspecies dissimilarities / correlations. This does not mean, however, that we must stay with a sub-optimal scaling of species scores when we plot the results from an analysis, where the scaling was focused on inter-sample distances. With knowledge of the eigenvalues of the ordination axes, the scores with an optimal scaling can be easily calculated from the “suboptimal” ones.

When this option is checked, CanoDraw rescales the “suboptimally” scaled scores if plotted alone (in scatter plots, not when plotted in biplots, joint plots, or triplots).

Show rescaling coefficients for composite ordination diagrams

In an ordination biplot, the absolute scaling of the item scores often does not have any meaning when the length of vectors with respect to the point positions is concerned, or when two independent sets of vectors (like species arrows vs. arrows for environmental variables) are compared. But a proper rescaling of one such set with respect to the other one(s) may facilitate easier reading of the ordination diagram. CanoDraw uses a custom set of rules for rescaling sample scores (including also centroids of environmental variables), species scores, and scores of environmental variables in their mutual respects, to provide improved usability of the ordination biplots or triplots.

This option (if checked) allows you to fine-tune the default method for mutual rescaling of ordination scores. The suggested **rescaling coefficients** (constants by which the particular kind of scores is multiplied across all items and all ordination axes) are displayed together with the information about the range of raw (non-scaled) scores and you can change the coefficient values.

The mentioned dialog is illustrated in Figure 12-9. Note that in this particular case, both sample and species scores are suggested to be squeezed to fit into an unchanged range of explanatory (environmental) variables.

CanoDraw does not differentiate between standard environmental variables and supplementary (passively projected) variables.

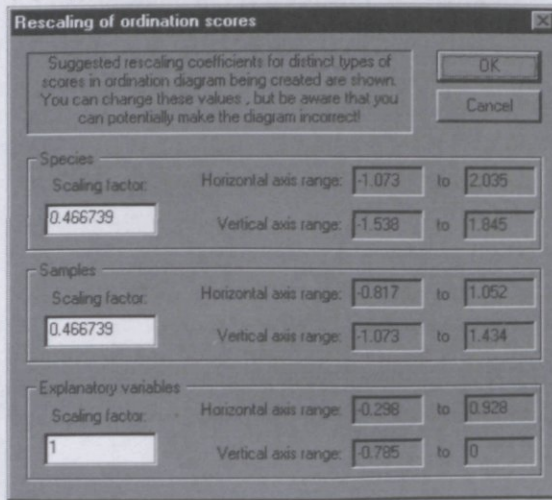


Figure 12-9 Rescaling of ordination scores dialog

Adjust graph aspect ratio by banking to 45 degrees (for XY diagrams)

Ordination diagrams use positions (point) and directions (vectors, presented as arrows) to summarise various aspects of the original data or of the tables derived from these data in a reduced number of dimensions (typically two dimensions, represented by horizontal and vertical diagram axes). Consequently, the physical scaling (the actual physical length on the output page or on the screen, corresponding to one unit in the ordination space) must be identical for the horizontal and vertical axes. The diagram is said to be **iso-scaled**.

On the other hand, if the variables used as coordinates for horizontal and vertical axis do not have any inherent relation, the scaling from the values of these variables to physical units of the

diagram is rather arbitrary and very often chosen so as to achieve unit aspect ratio. **Aspect ratio** is defined here as the ratio between the physical length of the line representing the vertical axis to the physical length of the line spanning along the horizontal axis ("diagram height divided by diagram width", said very approximately).^{*} The unit aspect ratio corresponds to a diagram with square shape.

But the shape of the diagram is not the purpose for creating it: the final goal is an efficient presentation of the information encoded in its contents. The patterns of relations between variables are often summarised using curves imposed over the plotted points, usually based on some kind of regression model. Even when we do not plot any curve, we tend to summarise the pattern seen in the scattered points by imaginarily superimposing such a curve. It was shown by Cleveland (1994) that the rate of change in response variable values (usually plotted on the vertical axis) is best judged by the viewer if the average absolute value of line segments inclination is equal to 45 degrees. The algorithm for calculating the optimum *aspect ratio* to be used to achieve this average 45-degree slope of curves is called **banking to 45 degrees**. If a smooth curve is plotted, it must be approximated by several line segments, but this is the way the smooth curves are usually presented.

If no curves are plotted in a diagram, the optimal aspect ratio is estimated based on an imaginary polyline, connecting plotted points in the order of their increasing X values. If there are multiple curves, the aspect ratio value is calculated separately for each curve and the recommended diagram aspect value is then calculated as an average of the aspect ratios of all the curves.

CanoDraw applies the "banking-to-45-degrees" algorithm only in diagrams which are not ordination diagrams (those created by the *Create / Attribute plots / XY(Z) Plot* and *Create / Attribute plots / Species response curves* commands). When XYZ plots (with an attribute - Z - variable present) are plotted, the banking algorithm is also inactive.

When banking to 45 degrees is active, CanoDraw calculates the suggested aspect ratio and displays it in a dialog box. This allows you to change it to any desired value. Note, however, that the changed aspect ratio will not result in an optimum slope for the plotted curves. If the plot contains any fitted curve(s), CanoDraw uses only the points on the fitted curves for banking to 45 degrees. When you select the *Cancel* button (or press *Esc* key) in the dialog for aspect ratio approval, CanoDraw temporarily sets the banking off and behaves according to the actual value of the "iso-scaling" option.

This option interacts with the option labelled *Iso-scaling* in the *XY Diagram Options* dialog, shown after the command *Create / Attribute plots / XY(Z) Plot*. If the *banking to 45 degrees* option is selected, the request for *iso-scaling* is ignored.

Plot also the extrapolated values

This option is used mainly in the contour plots, which are used by CanoDraw to present the dependency of one response variable on two predictors. A regression model (GLM, GAM, or Loess model) is fitted to describe such dependency and the fitted model is presented as isolines connecting points of the area spanned by the two predictors, having identical predicted values of the response variable. Often the resulting isolines (contours) cover large areas without any underlying observations, and there the values of response variable are extrapolated. To exclude the extrapolated areas from the presentation of fitted model, uncheck this option box. CanoDraw

^{*} Even the former definition is imprecise. The aspect ratio refers to the ratio of height to width of a *data rectangle*, defined as the smallest rectangle enclosing all the data points.

calculates a polygon enclosing the coordinates of the available data points and if this option is unchecked (**off**), only the isolines within such polygon are plotted.

Approve levels of contour plots

For the contour plots (see also the preceding option description), CanoDraw suggests the plotted contour levels using the same algorithm as the one used for determining tickmark positions along the diagram axes. If this option is **on** (the box is checked), CanoDraw displays the suggested levels in a dialog box and you can change them.

Plotting resolution for fitted curves and surfaces

To estimate the shape of the response surface in XYZ diagrams (with two predictor axes and one additional response variable), CanoDraw imposes a rectangular grid (with uniform distances between the grid nodes) over the two-dimensional space of predictors. For each grid node, the predicted (fitted) value of the response variable is estimated and this value then represents the height of the response surface at that particular point. The value you specify in this option field (which must be between 5 and 100) corresponds to the number of points on each side of the grid. Therefore, if you specify the value of 20 here, the fitted value of the regression model is predicted at 400 points, distributed regularly across the plane of predictors' values.

The same value is also used to regulate the smoothness of displayed curves in XY diagrams, which represent the fitted regression model with one predictor (the variable plotted on the horizontal axis). Each such fitted curve is approximated by a sequence of connected straight line segments, with their endpoints representing the fitted response values for particular predictor values. Predictor values are, again, distributed regularly across the whole range and the value in this option field determines their number.

Apply additional smoothing of contour lines

If a contour plot is created, the smoothness of the determined contours (isolines) can be optionally increased by applying a two-dimensional B-spline smoother. This setting enables or disables the application of that smoother to contour lines. If this option is **off** (unchecked), the values in the following two fields (under the *SMOOTHER SETTINGS* heading) are not used at all.

Smoother Settings / Expansion ratio

The increased smoothness of the contour lines is achieved by a closer approximation of the expected smooth paths using a B-spline smoother. The improved approximation is added to the existing contours by inserting further points between the vertices of the original polylines. The expansion ratio value determines the increase in the number of used points. For example, if the value is 2, the number of polyline vertices is doubled during the additional smoothing. This parameter takes only integer values between 2 and 5 (inclusively).

Smoother Settings / Tension

This parameter determines the smoothness of the approximated B-spline lines. The lower the value, the smoother the resulting contours are. Values of this parameter should be between 0.01 and 10.0.

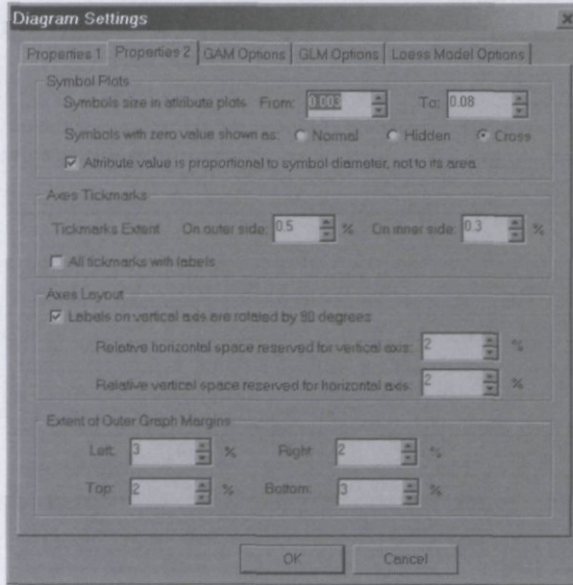


Figure 12-10 Properties 2 dialog page

This page collects additional project-specific options concerning the appearance of the diagrams created by CanoDraw.

Symbol size in attribute plots

Symbols plots represent one particular type of attribute plots provided by CanoDraw for Windows. In all attribute plots, the distribution of values of a response variable is shown in relation to two predictor (“independent”) variables, plotted along the horizontal and vertical diagram axes. In **symbol plots**, the values of the response variable are coded by the size of each symbol. The range of values of the response variable is linearly projected onto a range of either symbols diameter or of symbols area values (see the *Attribute value is proportional to symbol diameter*, described below, for explanation). Minimum and maximum values of the response variable then correspond to the minimum and maximum diameters of the symbols, which are specified here. The two symbol size values are quantified using the virtual coordinates (0-1) scaling, explained in section 11.6.

Symbols with zero values shown as

This option governs the presence and type of special treatment of the zero values in symbol attribute plots. The **Normal** option corresponds to treating the response value 0.0 in the same way as the other values. This choice is always used for symbol attribute plots where negative (less than 0.0) values are present. The choice labelled **Hidden** leads to zero values being plotted as empty symbols (existing, but not visible symbols) which are not labelled. The last possibility is labelled **Cross** and the zero values are plotted again unlabelled, but this time the actual data points are presented by cross symbols, each with radius of 0.0035 in 0-1 scaling units.

Attribute value is proportional to symbol diameter

If this option is **on** (checked), the radius (and diameter) of each symbol in a symbol attribute plot is proportional to the value of the response variable in the corresponding sample. If this option is **off** (unchecked), the radius (diameter) of each symbol is proportional to the square root of the response variable value. This, in consequence, leads to the response variable value being proportional to symbol area, not to its effective diameter.

Tickmarks Extent

Specifies the relative extent of axis tickmarks in both the inward and outward directions from the axis lines. The extent is measured in the virtual coordinate (0-1) units, post-multiplied by 100 (to be on a percentage-like scale). The same values are used for both the horizontal and vertical axis. An inward (**inner side**) direction value is treated in a special way, if it is larger or equal to 10.0. In that case, the ticks are replaced by lines spanning the whole length of the rectangle defined by the diagram axes. In this way, the inner parts of the tickmarks are changed into a reference grid.

All tickmarks with labels

CanoDraw determines the division of axes into identically sized steps, which are multiples of either $1.0 \cdot 10^x$ or $5.0 \cdot 10^x$, with x corresponding to the order of the plotted values. The resulting size of “steps” can, therefore, be 0.01 in case of $x=-2$, 0.5 for other diagram with $x=-1$, 10.0 for another with $x=1$, etc. CanoDraw plots the tickmarks at each of the steps to show their position on the plotted axis. By default, only the positions of the minimum and maximum step values are labelled. If this option is **on** (checked), the axis value is shown at each tickmark position.

Labels on vertical axis are rotated by 90 degrees

If this option is **on** (checked; this is the default setting), the labels of tickmarks on the vertical axis do not run from left to right, but rather from bottom to top, with each character rotated the same way.

Relative horizontal space reserved for vertical axis

This is the relative space in virtual (0-1) coordinate units (see section 11.6) multiplied by 100, reserved for the vertical axis line and tickmarks. CanoDraw uses this value (together with the height of font used for labelling the axes and left and right values of outer margins, see below) to shift the position of the rectangle where the actual diagram contents are plotted to the right.

Relative vertical space reserved for horizontal axis

This is the relative space in virtual (0-1) coordinate units (see section 11.6) multiplied by 100, reserved for the horizontal axis line and tickmarks. CanoDraw uses this value (together with the height of font used for labelling the axes and with the top and bottom values of outer margins, see below) to shift the position of the rectangle, where the actual diagram contents is plotted, in an upward direction.

Extent of Outer Graph Margins

Outer graph margins represent additional space around the graph, inserted there to visually separate the graph contents from the output page edges and to change the graph adjustment in respect to the page outline. The values are in virtual coordinate (0-1) units (see section 11.6), further multiplied by 100 to bring the scale to a percentage-like scale. For example, if you specify a value of 50 for the **left** field (50 is the maximum value for all four fields), the diagram is placed into the right half of the output page.

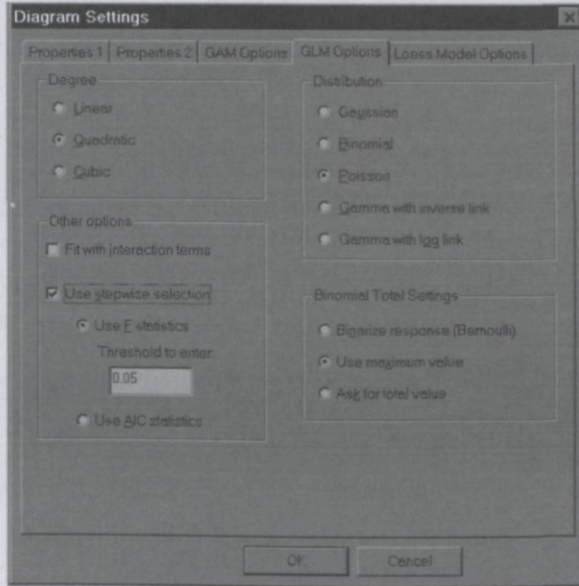


Figure 12-11 GLM Options dialog page

This page specifies the default options for generalized linear models (GLM) fitted by CanoDraw. Generalized linear models (McCullagh & Nelder 1989, Chambers & Hastie 1992) are an extension of classical linear regression model, which allows one to specify distributional properties of the stochastic component, and also the transformation function, which transforms the scale of predictor variables onto the scale of the response variable (the *link function*).

CanoDraw uses GLMs primarily as a tool for simplifying the visual presentation of patterns in the data, so the available options are simplified compared with a full-fledged statistical software. In such packages (like the S system, see Chambers & Hastie, 1992), the user has a greater freedom in combining the type of distribution of the stochastic component with the link function, in selecting the predictor variables and the form in which they enter the model. Also the number of tools for exploring the fitted models is substantially larger than is provided in CanoDraw (see section 13.6.4 for more details on the available regression diagnostic plots).

Distribution

Specifies the type of conditional distribution for the response variable*. Rough hints for selecting the appropriate distribution type are provided in Table 12-2.

* The distribution of actual response variable values around the value predicted for the particular values of the predictor variables

Type of response variable	Suggested distribution	Implied link function
Counts of animals or plant units with no strictly set upper bound; subjectively estimated percentage cover values on the scale from 0 to 100, but not approaching too often these limits	Poisson	log
Count of organisms or events out of an a priori given fixed number (like the number of plants surviving in a plot from a determined total number); presence or absence of an organism or events in a plot	Binomial	logit
Measures related to the weight or dimensions (plant biomass values, animal size, plant height, leaf area)	Gamma with log link	log
Ratio of two measurements with the same units; other dimension-less coefficients (e.g. various competition indices)	Gamma with inverse link	inverse
Measurements where assumptions of classical linear model (fitted using least squares) or classical ANOVA are fulfilled	Gaussian	identity

Table 12-2 Recommended choices for the *Distribution* field in GLM Options or GAM Options property pages

Degree

The systematic part of a GLM with a single predictor variable can be expressed in the following way:

$$g(EY) = \beta_0 + \beta_1 X$$

where g is the link function and EY are the expected values of the response variable given the values of the predictor variable X . Here, the value of the *linear predictor* (the right side of the above equation) depends linearly on the predictor values and this dependency is quantified using the single regression coefficient β_1 , estimated in the fitted regression model by value of b_1 . If you want your predictor to be represented in the fitted model in this form, select the **Linear** value for this option. With two predictors ($X1$ and $X2$, say) the linear form can be expressed as:

$$g(EY) = \beta_0 + \beta_1 X1 + \beta_2 X2$$

Of course, we can express the effect of predictor(s) upon values of response variable in more complex, non-monotonous form. For example, the classical unimodal model of species response to environmental gradients (see Ter Braak and Prentice, 1988) can under certain conditions also be expressed as second-order polynomial. This form then leads to the following expression for GLM with a single predictor:

$$g(EY) = \beta_0 + \beta_1 X + \beta_2 X^2$$

which is the classical Gaussian response curve if $\beta_2 < 0$ and $g()$ is $\log()$ (see p. 59-60),

and a GLM with two predictors, both in quadratic form, can be expressed in the following way:

$$g(EY) = \beta_0 + \beta_1 X1 + \beta_2 X1^2 + \beta_3 X2 + \beta_4 X2^2$$

Both preceding equations correspond to the **Quadratic** choice in this option and in the case of two predictors, it is assumed that the **Fit with interaction terms** option is **off** (unchecked). If that option is **on** (checked), the model with the added interaction term can be written as:

$$g(EY) = \beta_0 + \beta_1 X1 + \beta_2 X1^2 + \beta_3 X2 + \beta_4 X2^2 + \beta_5 X1X2$$

Similarly, the last choice of **Cubic** form of predictor variable(s) represents a third-order polynomial and here is shown only in its most complex form, assuming two predictor variables and including the interaction terms (represented by terms with coefficients β_7 , β_8 , and β_9):

$$g(EY) = \beta_0 + \beta_1X_1 + \beta_2X_1^2 + \beta_3X_1^3 + \beta_4X_2 + \beta_5X_2^2 + \beta_6X_2^3 + \beta_7X_1X_2 + \beta_8X_1^2X_2 + \beta_9X_1X_2^2$$

Binomial Total Settings

If the *Distribution* option value is specified as **Binomial**, each observation in the response variable must be described by two values: the number of events (successes, surviving organisms, etc) and the total, representing the maximum possible number of such events for a particular sample. The variable selected as response for the regression model fitted by CanoDraw is expected to represent the former type of values, and therefore CanoDraw needs to acquire values of the **Total** variable. In many cases, the value of the *Total* is identical for all the samples.

Binarise response (Bernoulli) is an option appropriate for the extreme (but frequently occurring, for biological data) case of a binomial distribution, also called the Bernoulli distribution, where the total value is fixed to be 1 (there is always one “trial”, so the outcome can be just success vs. failure, presence vs. absence). If the values of the response variable are not all equal to 1 or 0, they are binarised (all nonzero values are replaced with value 1).

Use maximum value – a general form of binomial distribution is assumed but with the constant value of the total, and an additional assumption is that the highest observed value corresponds to a situation where all the “trials” were successful, i.e. to the constant value of Total parameter.

Ask for total value – this is the most flexible choice for Total parameter specification. When this choice is active, a dialog allowing you to specify the Total parameter is shown at the time the model is fitted, similar to the diagram in Figure 12-12.

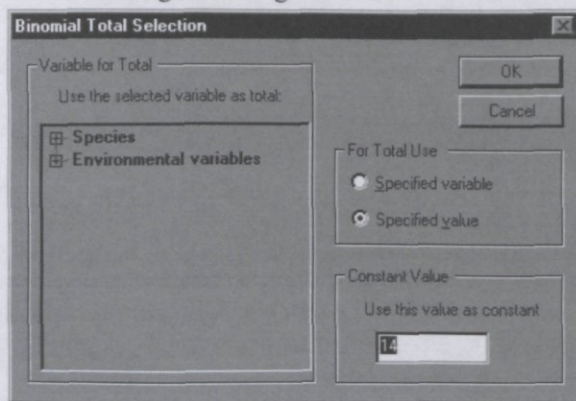


Figure 12-12 Binomial Total Selection dialog

Total parameter can be specified here using either an existing variable (the choice **Specified variable** enables the list on the left side) or by a constant value. The default value for the constant is again set to the maximum value of the response, but it can be changed (increased). The offered variables which may represent the binomial total are checked for their appropriateness, concerning the type and availability of all observations needed to support values of the actual response variable.

Fit with interaction terms

This option is used only if either Quadratic or Cubic forms of predictors are used and the regression model has two predictors. See the **Degree** option description above for additional explanation.

Use stepwise selection

CanoDraw allows you to select the complexity of the fitted regression model using either a stepwise selection procedure or by comparing candidate regression models using their parsimony. The two possible approaches are described below. This checkbox allows you to enable their use. If its value is **off**, the model specification is determined solely by the number of predictors (one or two) and by the choice made for the **Degree** parameter, as described above.

Use F statistics

This is the classical way of stepwise model selection. CanoDraw starts from the simplest regression model, called the **null model**, where no predictors are used and so the expected response value is assumed to be a constant. The quality of this model is then compared with the next more complicated model, which is a model with linear terms for the predictor (or both predictors). Comparison is based on a test using an F-like statistic, comparing the scaled residual deviance with the scaled residual deviance of the original, simpler model (see chapter 6.2.4 of Chambers & Hastie 1992, for a more detailed account of this method). The assumption of a F distribution for the calculated statistic under null hypothesis relies on the assumption of a χ^2 distribution of residual deviances, but the robustness of the F statistic-based test is higher than for direct tests on deviance change. The more complex model is accepted if the Type I error estimate of the test (the probability that the calculated F statistics originates from the F distribution) is smaller than the **Threshold** value specified in this property page.

If the linear model is accepted and the selection for the **Degree** option is *Quadratic* or *Cubic*, more complex models are tried. The set of tried models depends not only on the value of the **Degree** option, but also on the setting of the **Fit with interaction terms** option (only if it is **on**, the models with interaction terms are tried).

CanoDraw does not proceed with stepwise selection of individual terms. For example, when selecting the model with two predictors, CanoDraw compares the null model (simplified notations follow) $Y = \beta_0$ with model $Y = \beta_0 + \beta_1X_1 + \beta_2X_2$, but does not compare the null model with models containing only one of the two predictors. Similarly, if the linear model is selected, CanoDraw compares it with the quadratic model $Y = \beta_0 + \beta_1X_1 + \beta_2X_1^2 + \beta_3X_2 + \beta_4X_2^2$, but not with the models standing in complexity between these two. On the other hand, if interaction terms usage is **on**, CanoDraw tries both the models without and with interaction terms.

There is one additional peculiarity of stepwise selection in CanoDraw, increasing its utility in modelling species responses to environmental gradients (represented either by environmental variables or by ordination axes). If a regression model with a linear form of predictors is rejected in favour of the null model, CanoDraw continues with a comparison of the null model with a model where the predictor(s) are used in the second-order polynomial form (without interaction terms). This helps in the situation where the relation between the predictor(s) and response variable has a strongly parabolic form: the linear model is then rarely judged as significantly better than the null model and, therefore, the well-fitting quadratic model would be never tried. Note that this feature is applied only when the linear form of the GLM fails. If a quadratic form is deemed no better than the linear one, the cubic form is never tried.

Use AIC statistics

Akaike Information Criterion (AIC) provides a synthetic statistics for judging the *parsimony* of particular regression model on a scale comparable across different models as long as they have an identical set of data points, identical explanatory variable and identical assumptions of the distributional properties. AIC value is based on residual deviance of the fitted model – the lower the value, the better is the model able to predict response variable values. But the deviance is also penalised by the model complexity – number of model parameters (number of regression

coefficients in the case of GLMs) defining that particular model. The actual formula for the AIC used in CanoDraw is:

$$AIC = deviance + 2 * \varphi * p$$

where p is the number of model parameters (e.g. $p = 3$ for the linear form of the model with two predictors) and φ is the Pearson statistic - based *scale* estimate (see McCullagh & Nelder, 1989, p. 328).

When AIC is used for model selection, such a selection is, in fact, not truly *stepwise*, because the complete set of candidate models (limited by the values of the **Degree** and **Fit with interaction terms** options) is compared and the model with lowest AIC statistics is chosen.

In the model selection approach based on AIC, much of the discussion about problems with multiple comparisons performed on the same data-set, which leads some people to use Bonferroni corrections, loses its appeal.

12.3.1.4 GAM Options

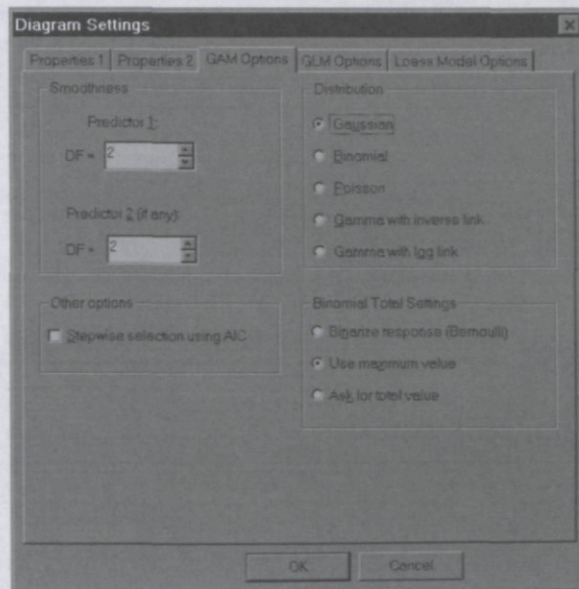


Figure 12-13 GAM Options property page

Generalized additive models (GAM) are a natural extension of generalized linear models (GLM), where predictor(s) effects upon the response variable are not expressed using a linear combination of the predictor values (and, eventually, of the second and third powers of those values), but using a *smooth* semi-parametric *term*, based on one or other kind of smoothing model. In GAMs, we do not control the exact shape of the curves corresponding to the smooth terms, but we rather control their complexity using a parameter that can be expressed on the scale comparable with the degrees of freedom (equivalent to number of parameters – regression coefficients in GLM). The two widely applied types of smoothers used in the smooth terms are the loess smoother and the smoothing spline model. Additional details can be found in Hastie & Tibshirani (1990) and Chambers & Hastie (1992).

The general formulation of the systematic part of a generalized additive model with two predictors (X1 and X2) can be written as follows:

$$g(EY) = \beta_0 + f_1(X1) + f_2(X2)$$

where EY is the expected value of the response variable Y , g is the **link function** (see section 12.3.1.3) and f_1 and f_2 are the semi-parametric smooth functions describing the effects of individual predictors and are called **smooth terms**. We can demonstrate the relationship between GAMs and GLMs by specifying smooth terms in the form $f_i = \beta_i X_i$ or even $f_i = \beta_i X_i + \beta_{i+1} X_i^2$.

CanoDraw supports only the cubic spline smoothers for the specification of the f functions and does not allow fitting of “mixed” GAMs (i.e. where one predictor is represented by a smooth term while the other is represented by a linear term, comparable to GLM).

Distribution

The offer of assumptions concerning the conditional distribution of the response variable values is identical with the one used for generalized linear models (GLM) and you can find its description in preceding section 12.3.1.3.

Smoothness

The two fields specify the (maximum) complexity of the smooth terms for the first and second predictor variable. The minimum value is 1.0, the maximum value is 6.0. Note that despite the units of these complexity parameters being comparable to the number of degrees of freedom, you can still specify fractional values (e.g. value of 2.4). The greater flexibility available with GAMs is achieved at the expense of the more difficult task of finding the proper model specification. Model selection using AIC statistics is therefore recommended.

Binomial Total Settings

Specification of the Total parameter in the situation where the response variable is assumed to have a binomial distribution is identical with generalized linear models and is described in the preceding section 12.3.1.3.

Stepwise selection using AIC

AIC statistics is calculated for the candidate additive models in the same way as described for GLMs in section 12.3.1.3, except that the parameter p is based on the number of degrees of freedom of the smooth term(s). The value(s) specified in the *Smoothness* field(s) of this property page represent(s) the upper limit of the parameter complexity for the smooth term(s). CanoDraw starts with a null model (with no predictor) and in the case of a model with single predictor continues with using smooth term with 1 degree of freedom, increasing this complexity parameter by one until the specified *Smoothness* value is reached or exceeded. In models with two predictor variables, the two sequences of smooth term complexity values are combined to define the set of evaluated candidate models. For example, if you have a model with two predictors and in the *Smoothness* section you specify value of 2.4 for *Predictor 1* and value 2.0 for *Predictor 2*, the complexity values of compared models (value for first predictor goes first) are: (null model), (1 1), (1 2), (2 1), (2 2), (2.4 1), and (2.4 2).

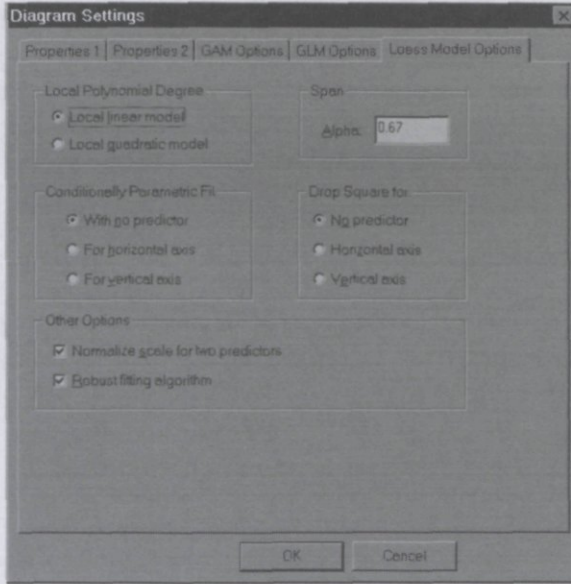


Figure 12-14 Loess Model Options dialog page

Loess (sometimes also called *lowess*) method is a locally weighted regression smoother which summarises the relation of one response variable to one or more predictor variables using a smooth curve or surface. The vertical position of such curve (or surface, for two predictors) is estimated at a particular point (for particular value(s) of predictor(s)) by fitting a weighted linear regression model to a subset of points. The subset membership is determined by the *span* parameter, which quantifies the fraction of closest neighbour points used to fit the model. Additionally, the data points used to fit the model have different weights, depending on their distance from the estimation point: the closer they are, the higher is their weight. Obviously, as the curve or surface is estimated over the whole span of the values of the predictor variable(s), the set of points used to estimate the regression model parameters changes, as do the weights of the included points. Traditionally, two kinds of local regression models can be fitted: either a linear model or a second-degree polynomial model.

Robust loess method applies to individual observations at each estimation point not only the *neighbourhood weights* (depending on their distance from the actual estimation point), but also so-called *robustness weights*, with their values determined iteratively at each estimation point. First, a standard loess model is fitted as described in the preceding paragraph. The robustness weights are then calculated, based on the distance of the data points from the fitted curve or surface. The further they are (i.e. the larger their residuals are), the lower weight they get. The loess model fitting is repeated, this time combining (by multiplication) the neighbourhood weights with the newly estimated robustness weight value, and a new set of robustness weights is estimated. This process is repeated until the robustness weight values converge.

Additional details about fitting a loess model, as well as about other options described below, can be found in Cleveland (1994) and Chambers & Hastie (1992), Chapter 8. Technical details about the algorithm used in the CanoDraw program can be found in Cleveland & Grosse (1991).

Span

The span value determines the set of observations, closest to the actual estimation point in the space (one- or two-dimensional) of predictors, that is used to fit the weighted regression

model. Note that even with value of span equal to 1.0, the loess model is not identical with fitting a simple regression line or second-order polynomial, because neighbourhood weights (and, optionally, robustness weights) are still applied. Therefore, a span larger than 1.0 can be also used (value up to 10.0 is allowed by CanoDraw), with the loess model gradually converging to a standard regression model.

Local Polynomial Degree

Here you can choose the complexity of the locally fitted regression model. If you select **Local linear model**, the fitted model has the form of either $EY = \beta_0 + \beta_1 * X$ (if you have just one predictor) or $EY = \beta_0 + \beta_1 * X1 + \beta_2 * X2$ (with two predictors). If you select **Local quadratic model**, the corresponding regression model is either $EY = \beta_0 + \beta_1 * X + \beta_2 * X^2$ with one predictor or $EY = \beta_0 + \beta_1 * X1 + \beta_2 * X1^2 + \beta_3 * X2 + \beta_4 * X2^2 + \beta_5 * X1 * X2$ in the case of two predictors. The model specification can be additionally modified using the *Conditional parametric fit* and *Drop square* options, described below.

Conditionally Parametric Fit

The fitted loess model with two predictors can be conditionally parametric in one of its two predictor variables. If we select, for example, **For horizontal axis** and the *Local Polynomial Degree* option has value **Local linear model**, the fitted loess model has a linear parametric dependency on the first predictor for any given value of the second predictor. The actual parameterisation (values of the regression coefficients) changes with the values of the second predictor variable. In this way, a loess model conditionally parametric in one of its predictor is a “mixed” model standing between the standard linear regression model and the standard loess model. This option has no effect in loess models with only one predictor.

Drop Square for

This option allows you to decrease the complexity of a local quadratic regression model used to estimate the loess surface. For example, if we select **Horizontal axis**, the fitted model will be $EY = \beta_0 + \beta_1 * X1 + \beta_3 * X2 + \beta_4 * X2^2 + \beta_5 * X1 * X2$ (compare with the above description of *Local Polynomial Degree* option). This option has no effect for loess models with just one predictor or using a local linear model.

Normalize scale for two predictors

The selection of the closest neighbouring points, used to estimate the loess model, can be substantially influenced by differences in the scale of the predictor values if you have two predictor variables. If this option is **on** (checked), the values of the two predictor variables are standardised by dividing the original values by the standard deviation of the particular predictor. The algorithm does not use the traditional standard deviation estimates, but rather 10% trimmed estimates, where 5% of the most extreme value at both ends of the range are ignored.

Uncheck this box if your two predictors are on the same scale (e.g. spatial coordinates of sampling points or sample scores in ordination space).

Robust fitting algorithm

If this option is **on** (checked), the robust loess model is fitted, as described above.

12.3.2 Visual Attributes

CanoDraw constructs graphs from the ordination results, original data, fitted regression models, and imported or otherwise constructed additional variables, using the settings specified in the *View* and *Project* menus. The actual graphs are composed from **graph objects** which are labels, symbols, arrows, lines, polylines, pie-symbols, or bars (see section 11.5). Each graph object has a particular set of attributes which determine its drawing color, fill color and pattern, line width and style, symbol or label size, typeface for labels, type for symbols, etc. You can

modify all these **visual attributes** for a single object or for any selected group of graph objects in CanoDraw (see section 12.3.4). But you can also change the default settings for the graphs to be created in future, using the *Visual Attributes Settings* dialog, illustrated in Figure 12-15.

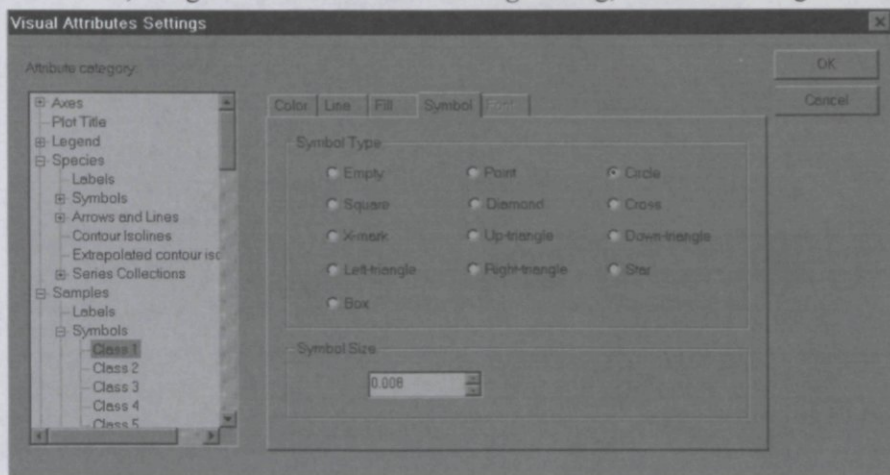


Figure 12-15 Visual Attributes Settings dialog

The left side of the dialog lists, in a hierarchical arrangement, the categories of graph objects recognised by CanoDraw. The right side displays five tabbed pages, usually some of them are disabled – you can work only with the pages, which are relevant for the type of graph objects selected in the left-hand list. The content of individual pages is explained in more detail in section 12.3.4.

Figure 12-15 demonstrates that you can, for example, set different attributes for samples being classified into different classes. They can be represented by symbols with different type, color, and size. The default attributes for classified items (e.g. symbols or arrows for samples, species, environmental variables) are visually different only for the first few classes, so you must modify the settings if you have a larger number of classes. The settings specified for *Class 1* are also used if the items are **not** classified.

The visual attributes you specify here apply to all projects and are usually maintained by CanoDraw in the system registry, separately for each user account. Additionally, you can store the application-wide settings affecting the appearance of your graphs in a file with *.cds* extension (see section 12.1.8). These settings include, among others, all the information specified in this dialog.

12.3.3 Workspace Settings

This command displays the dialog shown in Figure 12-16, which allows you to modify options regulating the behaviour of CanoDraw when working with projects and logging information about your work on a particular project.

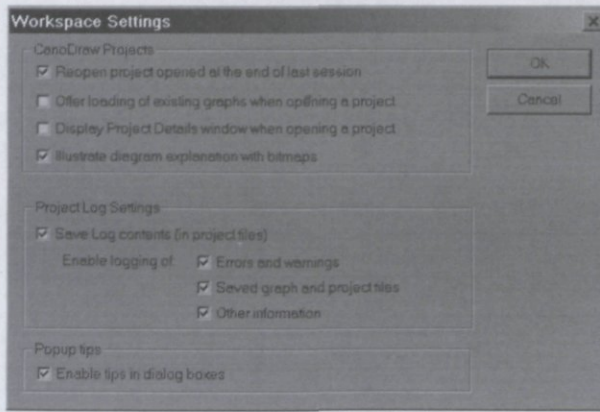


Figure 12-16 Workspace Settings dialog

Re-open project opened at the end of last session

If you close the CanoDraw application at the time when exactly one CanoDraw project is open (number of windows with CanoDraw graphs does not matter), CanoDraw records the file in which the project is stored and if this option is **on** (checked), attempts to re-open it at the start of the next session. CanoDraw tries to open it only if no other project is requested to be opened (i.e. you neither double-clicked an existing CanoDraw project icon in a Windows Explorer window, nor you clicked the *CanoDraw* button within the Canoco for Windows application).

Offer loading of existing graphs when opening a project

If you are opening an existing CanoDraw project (.cdw file) and this option is **on** (checked), CanoDraw offers you a list of graphs, which were created in this project and saved to disk. The list does not contain graph documents, which are not available for opening or which are found to be already open in CanoDraw workspace. The dialog is illustrated in another place, in section 12.4.10, in Figure 12-50. Each of the CanoDraw graph documents listed in this dialog box is preceded by a check-box. You should place a checkmark there for each graph you want to open. You can also delete reference to a particular graph (so that it is not offered for opening the next time) by selecting it and clicking the **Delete** button. Note that this does not delete the actual file, only the CanoDraw project "loses track" of this graph file. The graphs with checked boxes are opened in CanoDraw workspace after the project was opened.

Display Project Details window when opening a project

The Project Details window shows a hierarchically arranged list of variables available for a particular project (see section 11.3). If this option is **on** (checked), the window is displayed automatically when a CanoDraw project is opened.

Illustrate diagram explanation with bitmaps

If you select the **Describe contents** command from the context-sensitive pop-up menu of a CanoDraw ordination diagram, the Graph Description dialog box is displayed, providing hints on interpreting diagram contents (see section 13.5). If this option is **off** (unchecked), only a textual description is shown. If it is **on** (checked), the description is supplemented with schematic illustrations.

Save Log contents in project files

CanoDraw displays in its workspace Log windows for each opened project (see section 11.3). This option determines whether each log is created newly each time the project is opened or whether the log contents are made persistent across the separate sessions, by storing it in the .cdw file.

Enable logging of

This group of options decides what kind of information is stored in the project log:

Errors and warnings – CanoDraw records here information which may point to potential problems with the data, unreliable fitted regression models, missing information, etc.

Saved graph and project files – CanoDraw records the names of files in which new graphs or project files were saved, including a short summary of their contents

Other information – any other type of information which can be stored in the logs.

Enable tips in dialog boxes

The tips pop-up window is shown if the mouse pointer rests for a while over a dialog field and contains a short description of the meaning of that particular field. If you want to suppress the display of those hints, uncheck this option.

12.3.4 Properties Sheet **G**

This command displays a floating window, which shows the actual visual attributes of the graph objects currently selected in the active graph and allows you to change them. Visual attributes settings are divided into five property pages and usually only a subset of those pages is available, depending on the type of the currently selected graph object(s). The individual pages are described in the following text. The shortcut key for displaying this window is **F5**.

This window does not need to be closed if you want to continue work with the other windows in CanoDraw. Before you change the active window or select a different set of graph objects in the currently active graph, you must click the **Apply** button to commit any changes made to the visual attributes.

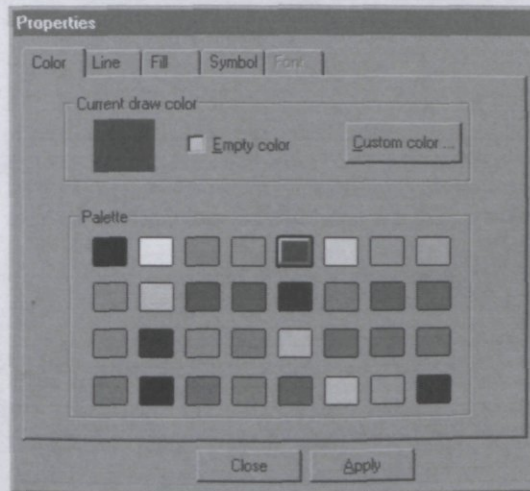


Figure 12-17 Color property page

This page displays the outline (drawing) color of the currently active graph object(s). You can select a different drawing color either from the palette of 32 colors (16 standard Windows colors, followed by another 16 – CanoDraw specific – colors) or you can define a new one, using the **Custom color** button. It displays the standard Windows' Color dialog box, where you can specify a new color either by selecting from a wide palette, pointing to a particular color hue in a color matrix, or entering the numeric values for R-G-B (or H-S-B) model color components.

If you want to make the color empty (no lines drawn), check the **Empty color** box.

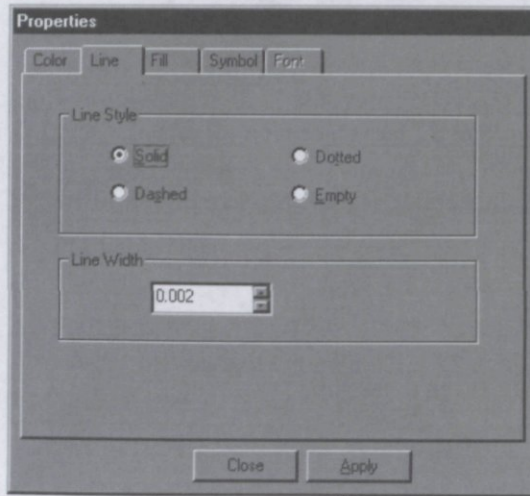


Figure 12-18 Line property page

On this page, you can change the line drawing style (solid, dotted, dashed, or empty line) and the line width, which is specified in the virtual coordinate (0-1) units (see section 11.6).

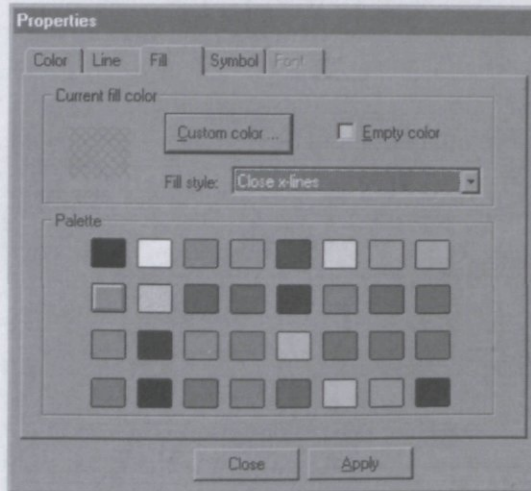


Figure 12-19 Fill property page

This property page allows you to specify fill style and fill color. The options for selecting the fill color are comparable with the **Color** property page described above. The *Fill Style* is only here, so there are two alternative ways of specifying empty fill style: you can either check the **Empty color** box or select *Empty* value from the **Fill style** list.

The **Fill** property page is also available for text labels, and in this case the fill style defaults to *Empty* value. But if you select a different fill style (preferably *Solid*), a background rectangle, underlying the displayed label, is drawn.

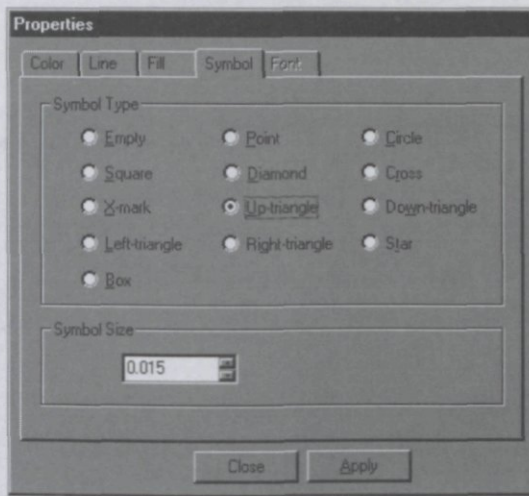


Figure 12-20 Symbol property page

This page specifies the symbol size and type for the selected symbol (or group of symbols). The **Symbol Size** is specified in the virtual (0-1) scaling units (see section 11.6).

This property page is also available for the pie-symbols, but in that case you should keep the symbol type selection of *Circle* and modify only the symbol size (pie-chart radius). If you modify the symbol type, your action is after-corrected and you are informed about that.

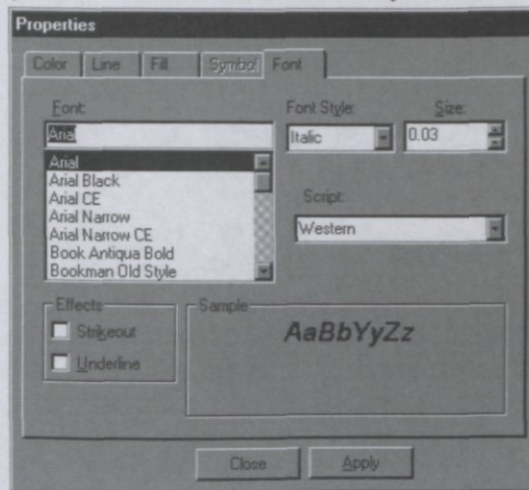


Figure 12-21 Font property page

This page is available only for text label objects. CanoDraw provides only the TrueType™ fonts. You can select here the typeface (font) name, the **Font Style** (*Regular*, *Bold*, *Italic*, and *BoldItalic*). Additional font effects (**Strikeout** and **Underline**) can be selected using separate check boxes. Another field, named **Script**, allows you to specify support of special characters beyond the standard U.S. ANSI character set. Font **Size** is specified in the virtual coordinate (0-1) units (see section 11.6). Field named **Sample** shows the approximate look of the label formatted with the currently selected font options.

12.3.5 Tree view **G**

Displays the Graph Contents window (see section 11.3).

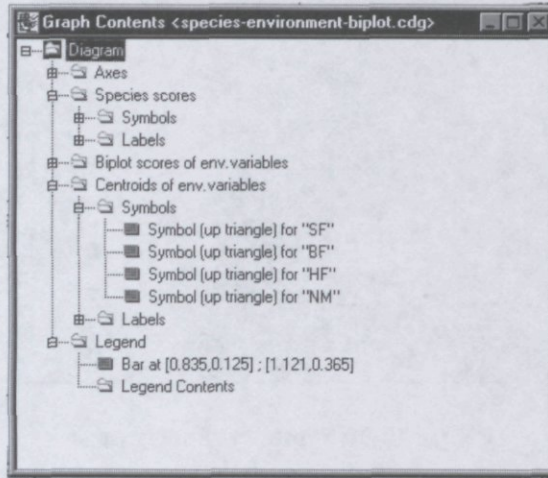


Figure 12-22 Graph Contents window

You can use this view of the graph objects hierarchy, in particular CanoDraw graph, to select one or more graph objects and modify their visual attributes.

12.3.6 Zoom **G**

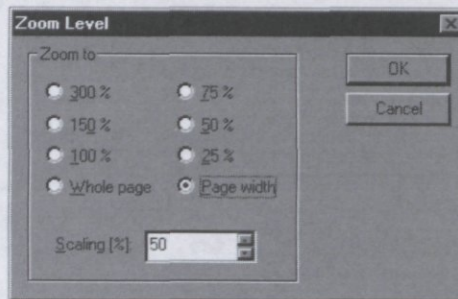


Figure 12-23 Zoom Level dialog

Allows you to specify precisely the zoom ratio for the currently active graph window. In addition to the choices in the upper part of the dialog box, which are also available from the drop-down list in the CanoDraw main toolbar, this dialog box provides also the **Scaling** field, where you can enter the requested numeric value of the zoom level.

12.3.7 Project Details **P**

Displays the Project Details window (see section 11.3) for the currently active project.

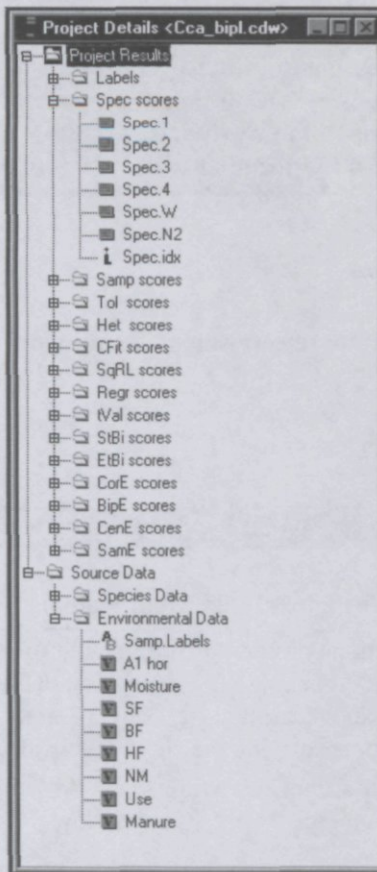


Figure 12-24 Project Details window

This window lists the variables available in a CanoDraw project and belonging to one of three broad classes:

Project Results group corresponds to the contents of the Canoco™ solution (.sol) file and provides primarily the scores of samples and species (and, optionally, of environmental variables and / or of supplementary variables) on the first four ordination axes. Also the other statistics available for species, samples, and environmental and supplementary variables in the solution file are provided here. The contents of this group are further structured into sections, which – except the *Labels* section – correspond to the score sections in the original solution file.

Source Data group collects the Canoco source data variables – values of individual species, environmental variables, supplementary variables, and covariables. These are available under sub-classes corresponding to individual kinds of data files (*Species Data*, *Environmental Data* in Figure 12-24).

Imported Variables group lists the variables which were imported from the Clipboard (see section 12.4.8.1), from a Canoco data file (see section 12.4.8.2), created as PRC scores (see section 12.4.8.3), or stored during the creation of a residuals plot (see section 13.6 for further details). The variables in this group are further collected into one or more subgroups, depending on which kind of items they refer to (*For samples*, *For species*, etc.).

You can display a window with a summary of any variable by clicking the variable name with the right mouse button. A floating dialog appears, showing a statistical summary of that variable. You can use the *Copy* button to place a copy of the variable values (together with corresponding item indices) onto the Windows Clipboard. You can paste the values from there either into a spreadsheet document or into another CanoDraw project.

12.3.8 Bars

This submenu allows you to select which control-bars are visible in the CanoDraw workspace.

12.3.8.1 Main Toolbar

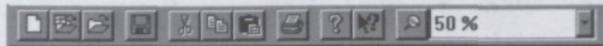


Figure 12-25 Main toolbar

The main toolbar collect buttons providing short-cuts to the commonly used operations like creating a new CanoDraw project, opening an existing CanoDraw project or CanoDraw graph, saving the currently active document, printing it, copying a CanoDraw graph to the Windows Clipboard, getting on-line help, or specifying the zoom level value.

You can also drag the toolbar into CanoDraw workspace so it will become a floating window, or dock it on another edge of the CanoDraw workspace window.

12.3.8.2 Graph Tools

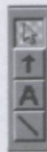


Figure 12-26 Graph Tools toolbar

Graph Tools toolbar enables you to select a tool for adding new objects of specific type (arrow, label, or line) to a CanoDraw graph.

To create an **arrow**, you should select the arrow tool (second from the top in Figure 12-26) and then click on the point where the arrow starts (the place of arrow base). As you reposition the mouse pointer, the outline of the currently implied arrow object is previewed. After you click with the left mouse button the second time, the tip of the arrow object is defined and the new arrow object is drawn.

To create a **line**, you should proceed in a similar way, selecting the line drawing tool this time.

To create a **label**, you should select the label-defining tool. You then click just once to define the point on which the new label is centred. The label object is defined with a default text and default font properties, which can be changed later on.

12.3.8.3 Status bar

The status bar of the workspace window is useful primarily when you are browsing through the application menu commands. The status bar area gives you a short help summarising the potential effect of the currently selected menu item.

12.4 Project

Commands in this menu allow you to manipulate the currently active CanoDraw project. The commands in this menu are also enabled if the active window represents a CanoDraw graph related to a project, currently open in the CanoDraw workspace.

12.4.1 Settings

This command invokes a dialog containing four tabbed (property) pages, collecting options specific for individual CanoDraw projects. The actual settings for these options are stored in the *.cdw* files.

12.4.1.1 Contents page

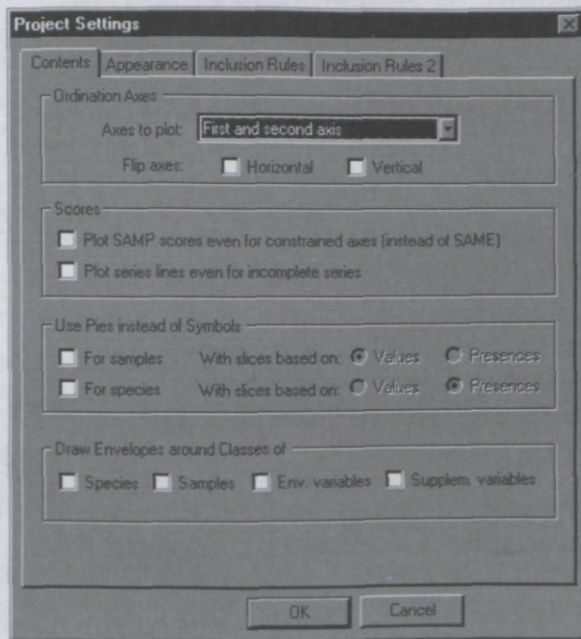


Figure 12-27 Contents page

This page collects project-specific options which affect the diagram contents.

Axes to plot

This option determines the pair of ordination axes used in the ordination diagrams. The axis with lower number is always plotted in the horizontal direction. The pair of axes selected here is used in all the ordination diagrams (or attribute plots based on an ordination diagram), except

those created with the *Create / Simple Ordination Plot* where the axes combination is specified directly (see section 12.5.1).

Flip axes

Changes the sign of scores on the **Horizontal** and / or **Vertical** axis if selected (checked). The interpretation of ordination diagrams (using either the biplot rule or distance-based interpretation) is invariant to such mirroring operations, if performed on all types of items plotted in the same diagram. This feature is useful in the situation where the same type of analysis is performed on a group of related data sets (say measurements from different areas and / or from different years). The analyses results may then show comparable patterns, except that sometimes the patterns are mirrored with respect to the horizontal or the vertical direction: an explanatory variable pointing in most diagrams to the upper left corner points to the upper right corner in one or few others, and similarly the arrows (or symbols) for species may be swapped in the horizontal direction. You can flip the axis orientation back for such exceptional cases. Note that the biplot rule and distance rule, which are used to interpret the contents of ordination diagrams, are invariant to flipping orientation of axes.

Plot SAMP scores even for constrained axes

If canonical (constrained) axes are plotted in ordination diagrams, samples are by default represented on such axes by *sample scores which are linear combinations of environmental variables* (also called **SamE** scores) – see section 6.3.7. This is because in the constrained ordination analyses, the primary task is to explain patterns in the primary ("species") data by the values of explanatory ("environmental") variables, using a constrained multivariate regression model. The *SamE* scores represent the fitted values from this model and reflect differences among the samples in terms of the values of the explanatory variables. The other kind of sample scores, derived from the species (response variables) scores (also called **Samp** scores), is therefore less important for the primary task of constrained ordination. It is used, however, by default for unconstrained (non-canonical) ordination axes in a constrained ordination and for all axes in unconstrained ordinations (like PCA or DCA).

When this option is **on** (checked), the unconstrained sample scores are used even for the canonical ordination axes.

Plot series lines even for incomplete series

The series collections represent the description of one or more independent sequences of samples (or species or explanatory variables) which can be displayed in an ordination diagram or XY diagram, where objects of particular type are plotted.

If some series items are missing from the plot (due to a limited range of diagram axes or due to item suppression based on various selection / suppression methods available in CanoDraw), the line(s) representing that particular series are not plotted: connection of non-contiguous items by a series line would provide false feeling of contiguity. You can override this behaviour by setting this option **on** (checking this box).

Use Pies instead of Symbols: For samples / For species

If this option is **on** (checked) for samples and / or for species, their symbols are replaced by pie-symbols. Two additional conditions must be met, however. First, there must exist a classification for the complementary type of items (e.g. if you want to plot pie-symbols for samples, a classification for species must be available, and vice versa) and the classification must be active (see section 12.4.3 for more information about classifications). The second condition concerns only pie-symbols for species: the species must in the diagram be represented by symbols (points), not by arrows (vectors).

Use Pies ...: With slices based on

This option is available separately for samples and species. Each pie-symbol represents one sample [or species] and is divided into slices (wedges) corresponding to classes of species [samples] which are present in the sample [in which the species occurs]. The angle taken by particular wedge can be calculated in two different ways corresponding to the two choices for this option:

Presences – the relative size (angle) of an individual wedge in a pie-slice representing a sample [a species] is proportional to the number of species belonging to the particular species class and occurring in this sample [is proportional to the number of samples belonging to the particular sample class, in which this species occurs].

Values – in this case, the wedge angle is based on the importance of individual occurrences, expressed by the actual values in the species data. For example, if your primary (species) data represent biomass values of individual species and you want to plot pie-symbols for individual samples (with species being classified into distinct classes), then the relative size taken by a pie-wedge (representing one species classes) in one sample corresponds to the relative fraction (percentage) of the total sample biomass, corresponding to the biomass of species from that particular class.

Draw Envelopes around Classes of

Classified items (samples, species, environmental variables, and supplementary variables) can be plotted using symbols of different type and color (or else by arrows of different color and line style), if they belong to different classes. Additionally, you can visually assemble symbols belonging to items from the same class by drawing a polygon, enclosing all of them and having its vertices on the "outermost" members of the class' group. These polygons are drawn if the corresponding checkbox is on.

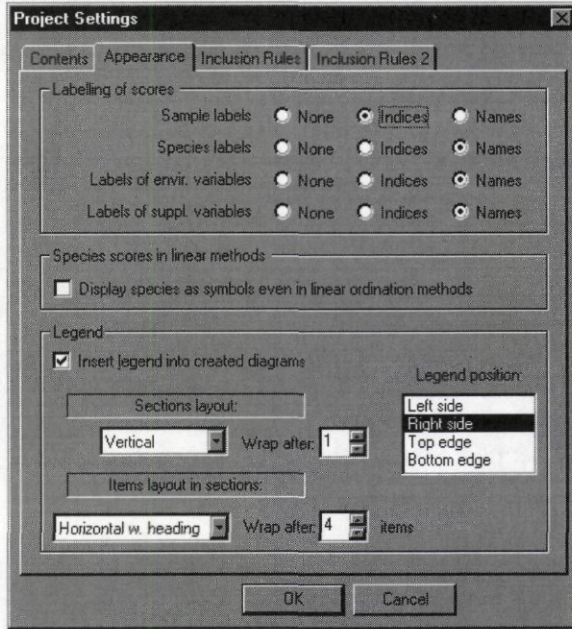


Figure 12-28 Appearance page

This page collects the project-specific options which affect the way the diagram contents is presented to the user.

Labelling of scores

The settings in this group apply only to **ordination** diagrams. These options, with three choices (**None**, **Indices**, and **Names**) are available for samples, species, environmental variables, and supplementary variables. If **None** is selected, only bare symbols (or arrows) are plotted, without any label. Alternatively, the indices (starting with 1 for the first item) as used in the original Canoco source data files are used as item labels (**Indices** choice). The choice **Names** causes original labels (with up to 8 character positions), available from the Canoco source data files (or from the Canoco SOL file), to be used in the ordination diagrams.

Display species as symbols even in linear ordination methods

When displaying results from linear ordination methods (PCA or RDA), the response variables ("species" in Canoco terminology) should be plotted as vectors (arrows) in the ordination diagrams. They are a natural presentation of the response variables in such analyses, representing directions of predicted steepest increase in values of a particular variable or (with a different presentation of the same concept) the values of regression coefficients of a multiple regression using the corresponding species as the response variable and the sample scores on the displayed axes as predictors.

Plotting the species scores from the linear analyses as symbols is not a recommendable practice, because it misguides the viewer into judging correlation among species based on the distance between such points. Nevertheless, there might be reasons for plotting only the tips of the (imaginary) species arrows (e.g. to lower the clutter of arrowlines) and this is acceptable, provided the user is well aware of the correct ways of interpreting the resulting plots. This option allows you to enable the plotting of species by symbols in linear methods.

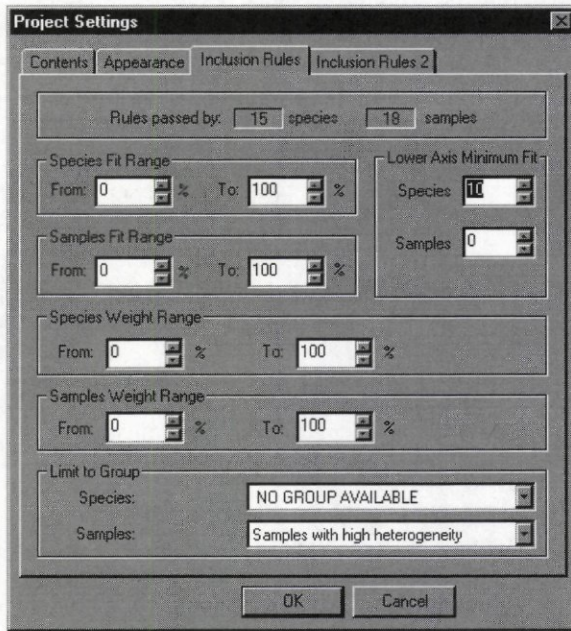


Figure 12-29 Inclusion Rules page

This page controls the presence of species and samples in the diagrams produced by CanoDraw. Their presence is regulated either by a specification of selected sample / species statistics (see below) or by selecting a group of species or samples and so limiting the set of plotted items just to the group members. Multiple restrictions implied by this property page and by the additional explicit removal, available from the *Project / Suppress* submenu (see section 12.4.6), are combined together, so any item failing in one or more of the restrictions is not plotted.

Rules passed by ...

These two fields provide estimates of the expected number of samples and species passing the currently specified settings on this page.

Species Fit Range

You specify here the lower and upper limit of the interval of values of species fit into the ordination space, into which a species must fall to be plotted in an ordination diagram.

Fit of species into ordination space is based on *CFit* statistic calculated by the Canoco for Windows program (see section 6.3.11.2). Unlike the actual Canoco output, the values of the fit as used in CanoDraw are multiplied by 100.0, so they roughly represent the "percentage of variance in the values of the particular species, explained by the given ordination subspace". To characterise this statistic more precisely, we can imagine that we would take the values of particular species in all the samples and use them as a response variable in a linear regression (in the case of linear ordination method; weighted regression on relativised species values would be used in unimodal methods), with the scores of the samples on the horizontal and vertical axis being used as two predictor variables. The coefficient of determination (R^2 - relative amount of explained variance) for such a regression then corresponds to the fit value for the particular species.

Two things should be noted here. First, the values specified in this dialog change their effect on the selection of subset of plotted species with the change of plotted axes. That is, the same

range of fit values will select different species if you plot second vs. first axis or third vs. first axis, for example. Second, the used statistic is **not** the cumulative fit of the species up to the highest plotted axis – it is rather the direct contribution of the two plotted axes to the explanation of the species abundances. If you plot, for example, the third vs. first axis, the values used to decide which species to plot correspond to cumulative fit for the first three axes, with the effect of the second axis subtracted.

This option provides both an upper and a lower limit, so either species with low or high fit values can be selected. Additionally, species with intermediate fit values can be selected by using limits like 15 and 85, for example.

Samples Fit Range

This option is similar to the preceding one, but provides a restriction on samples, not on species. The statistic used here is based on the **SqRL** ("squared residual length per sample") statistics provided by Canoco (see section 6.3.11.2), but further post-processed by CanoDraw to give it comparable scaling units as for the species fit values. The resulting statistics then estimate the percentage of variance in the values of all species (primary variables) in the particular sample, explained by the plotted ordination plane. Therefore, a sample with value 100 would have the values of all the species predicted precisely from the fit using the two ordination axes (for linear methods; only the relative proportions of individual species would be predicted precisely for unimodal ordination methods).

Lower Axis Minimum Fit

The lower axis minimum fit values select the plotted samples or species based on their fit just on the horizontal ordination axis (the axis with the lower order). This is an important option in the situation where you plot, for example, an ordination diagram with the first two ordination axes, but where the first axis is canonical (constrained by the explanatory variable(s)), while the second is not. The standard "Fit Range" option does not allow you to differentiate between the contribution of the horizontal and the vertical axis to the fit of species and / or samples.

Species Weight Range

You can specify here the range of weights the species must have in the ordination analysis to be displayed. **This option is available only in projects based on an unimodal (weighted averaging) ordination (CA, DCA, CCA, DCCA).** The original weight values (imported by CanoDraw from the Canoco solution file) are rescaled so that the largest value becomes 100.0. Therefore, the weights used by CanoDraw represent the percentage of the weight of the species (or sample) with the largest impact on the analysis results.

Samples Weight Range

This option is similar to the one described in the preceding paragraph, except that the sample weights are used here. It is available only for the projects resulting from unimodal ordinations.

Limit to Group

If you defined one or more groups for your samples or species (see section 12.4.4) you can specify here one of them and only the samples (or species) from that group will be plotted (if they pass through the other inclusion rules).

If you need a selection rule combining several groups (like "plot all samples where Spec05 occurs with quantity at least 10.0 and where the total number of species is more than 7"), you can combine existing rules (using logical complement of one group or logical overlap or unification of two groups) in the group manager (see section 12.4.4).

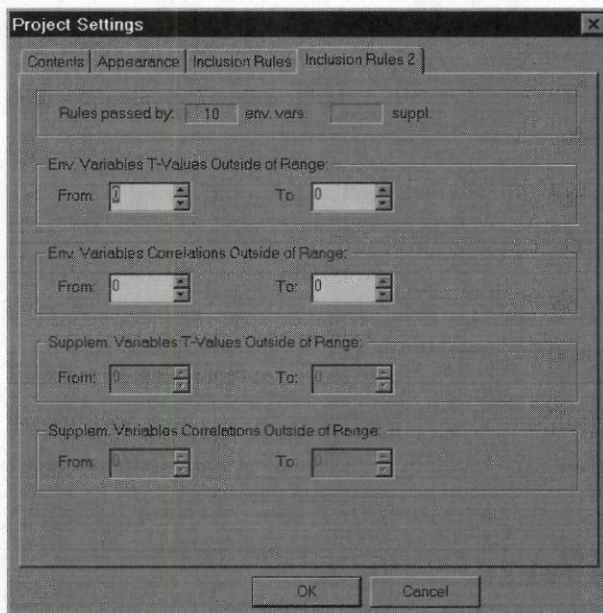


Figure 12-30 Inclusion Rules 2 page

This page provides control over the set of environmental or supplementary variables which are plotted in ordination diagrams. For both types of explanatory variables, you can specify an *exclusion range* for T-values and for the correlations with the displayed ordination axes. The **exclusion range** is a range of values of a particular statistic, which leads to the exclusion of the corresponding variable from the set of plotted variables. This kind of specification is used because the statistics used as inclusion rules for environmental and supplementary variables have the property that the values largest in absolute magnitude represent the most important variables. For example, a strong relation between an environmental variable and an ordination axis is signified by a value approaching -1.0 or $+1.0$, while a value near 0 implies a very weak or non-existent relation between the variable and the ordination axis.

Rules passed by

These two fields estimate how many environmental variables and supplementary variables pass through the inclusion rules corresponding to the actual thresholds set in the individual fields below them.

Env. Variables T-Values Outside of Range

T-values used here correspond to the T statistics of linear regression coefficients from a multiple regression model where sample scores on a particular ordination axis represent the response ("dependent") variable, while the environmental variables are used as predictors. An environmental variable passes this inclusion rule, if the value of its T statistics is outside the specified exclusion range at least for one of the two ordination axes plotted in the ordination diagram. Therefore, the set of environmental variables passing this rule usually changes with the change of displayed ordination axes. CanoDraw uses the T-values provided by Canoco in the solution file (see section 6.3.6).

The T values are for regression coefficients from a multiple regression model, in which all the environmental variables are used at the same time. Therefore, these values are sensitive to the inclusion of other, correlated environmental variables, and so a selection based on these statistics should be carefully reviewed. Generally, redundant explanatory variables should be first excluded from the analysis (e.g. using the forward selection of environmental variables). Also note that if a group of dummy variables is used to code values of a factor, the T-value statistic for one of such dummy variables is not calculated at all and its value is set to 0. Therefore the corresponding factor level will be excluded from the plot if any non-empty exclusion range is specified here, unless this dummy variable is forced into the ordination plot (see *Project / Enforce* menu).

Env. Variables Correlations Outside of Range

The exclusion range to be specified here refers to the correlations between environmental variables and individual ordination axes, termed *Inter-set correlations* in the Canoco™ program. These correlations are (weighted) linear correlation coefficients between sample scores on the particular ordination axis (derived from the species scores) and values of the particular environmental variable. CanoDraw uses the correlation values provided by Canoco in the solution file (see section 6.3.8). As for the above criterion, CanoDraw chooses between the two candidate correlation values for each variable (correlation either with the horizontal or with the vertical axis), using the one with the larger absolute value.

Supplem. Variables T-Values Outside of Range

This inclusion rule limits the set of plotted supplementary variables. See the preceding paragraph describing the same criterion, used with environmental variables, for an explanation.

Supplem. Variables Correlations Outside of Range

This inclusion rule limits the set of plotted supplementary variables, based on their correlation with the ordination axes. See the description in the above paragraph explaining the use of correlations for environmental variables.

12.4.2 Nominal variables

The nominal explanatory variables describe the state of an object or sample using a limited set of values. The nominal variables can be also called **factors** and their possible 'states' (values) are called **factor levels**. In the simplest case of a nominal variable with just two states, 0 and 1 can be used to identify, respectively, absence or presence of a feature or event. Qualitative explanatory variables with more than two possible states (values) must in the Canoco™ program be replaced by a group of **dummy** variables, with 0 and 1 values. There is one such variable for each level ("class") of a nominal variable and a value of 1 indicates that the particular observation belongs to this level (state). Consequently, there is always just one variable in such a group with value 1 and the remaining ones have zero value.

The two commands in this submenu enable you to specify which of the explanatory variables (either environmental variables or supplementary variables) can be treated as levels of nominal variable(s). This is necessary because the levels of a nominal variable are often displayed in ordination diagrams using symbols representing the centroids of sample scores belonging to the particular class (factor level), not by the biplot scores (which are plotted as arrows). These centroids are available as **CenE** scores in the Canoco™ solution file. Additional information can be found in the section 6.3.10.

12.4.3 Classify

The individual samples, species, and – if available –environmental variables or supplementary variables can be classified into several distinct classes. If one kind of items (e.g. the samples) is classified, every such item must be assigned to exactly one class. Such a set of assignments is called **classification**. Each classification may use up to 64 different classes. CanoDraw supports, for each kind of items, an unlimited number of classifications, but only one of them (or, alternatively, none of them) may be active at a time for a particular kind of items.

The active classification is reflected in the diagrams created by CanoDraw by symbols of different type (circles, squares, crosses, etc.), color, or even size. CanoDraw presets the distinguishing symbol attributes for the first 16 sample and species classes (and for the first 8 classes of environmental and supplementary variables), but you can modify those settings and specify distinctive attributes for additional classes in the dialog shown by the *View / Visual Attributes* command (see section 12.3.2).

Classification of items can also be reflected in envelopes drawn around symbols belonging to the same class and in the pie symbol plots (see section 12.4.1.1 in both cases).

When you select one of the commands (named after the kind of classified items) from the *Classify* submenu, a dialog box similar to the one in Figure 12-31 is shown.

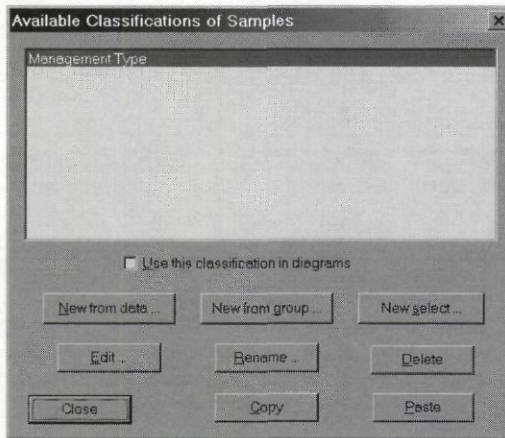


Figure 12-31 Available Classifications dialog

Not all the buttons are enabled for all the types of items. *New from data* is available only for a classification of samples, *New from group* is available only for a classification of samples or species. The buttons in the second row (*Edit*, *Rename*, and *Delete*) and the checkbox are available only if at least one classification is defined and selected in the list in the upper part of the dialog.

12.4.3.1 New select

This command (button) is always enabled and allows you to define a new classification from scratch, defining new classes and their members manually. After you select this command, a dialog similar to the one in Figure 12-32 is shown.

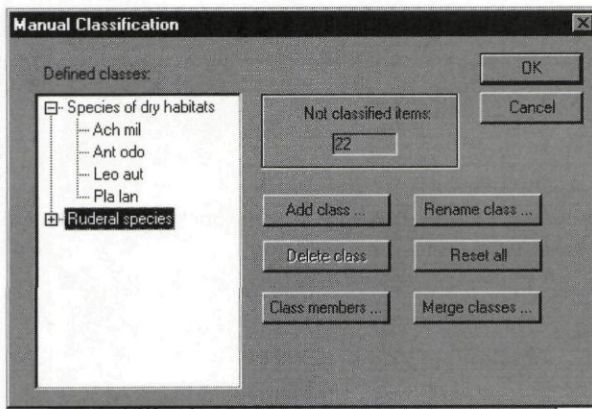


Figure 12-32 Manual Classification dialog

The hierarchical list on the left side shows the current state of classification. Individual classes are listed at the top level (these are represented by the labels *Species of dry habitats* and *Ruderal species* in Figure 12-32) and the nested items correspond to members of the particular class. The class members are shown in the list only if the total number of items to be classified is not larger than 2000. Otherwise, only the class names are displayed. This does not have any adverse impact on the dialog functionality, however, because the commands available here work only with the whole classes.

The non-editable field *Not classified items* shows how many items are not a member of any of the currently defined classes. This value must be zero (i.e. all items must be classified) before the *OK* button is enabled.

Add class

If you click this button, a dialog similar to the one in Figure 12-33 is displayed.

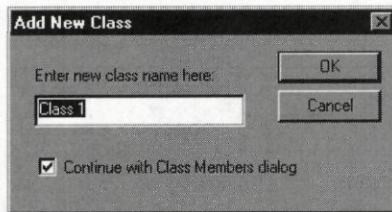


Figure 12-33 Add New Class dialog

This dialog suggests default class name (consisting of the word *Class* and a sequential number) but you can change it. The checkbox below the edit field, which is checked by default, specifies whether another dialog – used to specify members of this new class – should appear after this dialog is closed with the *OK* button.

Delete class

The class selected in the left-hand list is deleted (after confirmation).

Class members

This command shows a dialog, which is illustrated in Figure 12-34, for the class currently selected in the left-hand list of classes. It is also shown when you add a new class with the *Continue with Class Members dialog* option being checked in the *Add New Class* dialog.

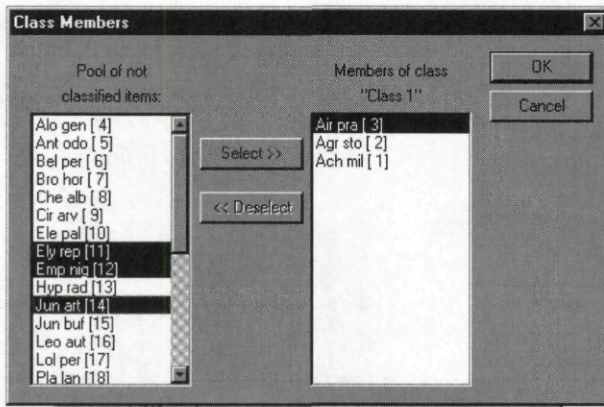


Figure 12-34 Class Members dialog

The left-hand list (*Pool of not classified items*) shows the names of items, which can be assigned to the current class, because they are not yet members of any existing class. The right-hand list (*Members of class "XXX"*) shows the items, which are already members of the current class. In both lists, the names are supplemented by the item indices in the square brackets. To move items from the left to the right list or in the other direction, select the items in the source list and click the *Select >>* or (*<< Deselect*) button. Items are listed in both boxes in alphabetical order.

Rename class

Displays a simple dialog box allowing you to change the name of currently selected class.

Reset all

Completely resets the classification, removing all defined classes and moving all the currently classified items into the pool of non-classified items.

Merge classes

This button displays the dialog illustrated in Figure 12-35.

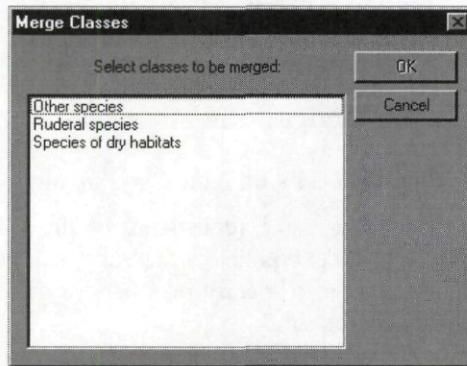


Figure 12-35 Merge Classes dialog

The currently defined classes are listed here and to merge classes, you should select two or more items and click the *OK* button.

12.4.3.2 New from data

This command (button) displays the *Classify From Data* dialog (illustrated in Figure 12-36), which allows you to define classes of samples based on values of one or more variables.

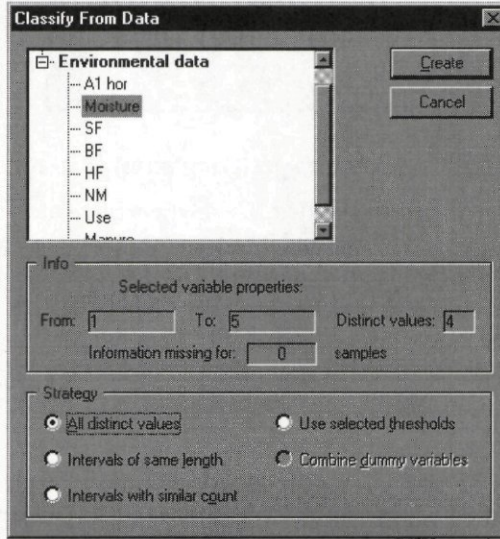


Figure 12-36 Classify From Data dialog

The box in the dialog's upper left corner shows the variables available as a source of information for the definition of a new classification. To define a new classification, select either one (quantitative) variable or a group of several dummy (0/1) variables and then choose the classification strategy in the bottom part of the dialog. If you select a single variable in the box, the range of values and the number of distinct values are shown in the middle part of the dialog. In that case, the first four choices are enabled in the *Strategy* field. Alternatively, you can select two or more variables representing, if combined, individual levels of a factor (nominal variable). If more than one variable is selected, CanoDraw checks that values of each of them are either 0 or 1 and displays an error message if other values are found. Nevertheless, CanoDraw does **not** enforce a crisp classification in the sense that exactly one of the selected 0/1 variables has a value of 1 for each sample.

After you selected the variable(s) and selected the classification strategy, click the *Create* button to proceed with the classification. The possible classification strategies are discussed in the following paragraphs.

All distinct values

This classification strategy places samples with different values of the selected variable into separate classes. If there are more than 63 distinct values, CanoDraw collects all the values largest than the upper threshold for the 62nd class into one composite class.

After the classification is created, it is displayed in a dialog similar to the one in Figure 12-32, where you can inspect and fine-tune the classification.

Intervals of same length

CanoDraw divides the range between the minimum and maximum values of the selected variable into specified number of intervals of the same length. You are asked to set the number of intervals with the dialog illustrated in Figure 12-37. There should be at least two intervals and the maximum number of intervals is 63 or the number of distinct variable' values, whichever is lower.

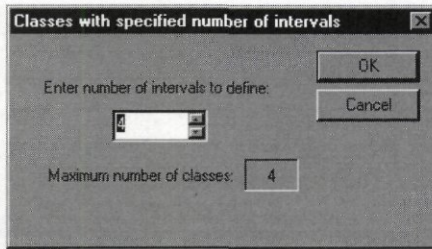


Figure 12-37 Dialog for specifying number of intervals

After the number of intervals is specified, CanoDraw calculates the threshold (boundary) values and displays them in a dialog box similar to the one shown in Figure 12-38. You can use this dialog to modify the actual boundary values and to check how many members the individual classes would have. Use of this dialog is explained in more detail in the section *Use selected thresholds* below. After you close this dialog with the *OK* button, the classification is created and displayed in a dialog similar to the one shown in Figure 12-32. You can make additional adjustments to the classification, e.g. change names of individual classes.

Intervals with similar count

This classification strategy is similar to the one described in the preceding section (*Intervals of same length*), but this time the boundary values (thresholds) between the intervals are found so as to create classes with as closely similar number of members as possible. A completely identical size of individual classes is usually not achievable even when the number of samples is a multiple of the requested number of intervals, due to the presence of identical values in the variable used for classification.

Use selected thresholds

This is the most flexible form of classification based on the values of a single quantitative variable. The dialog illustrated in Figure 12-38 is shown. This dialog displays supporting information about the selected variable, as well as the actual **boundary values** (the limits, dividing the range of values of the selected classification variable into intervals corresponding to individual classes).

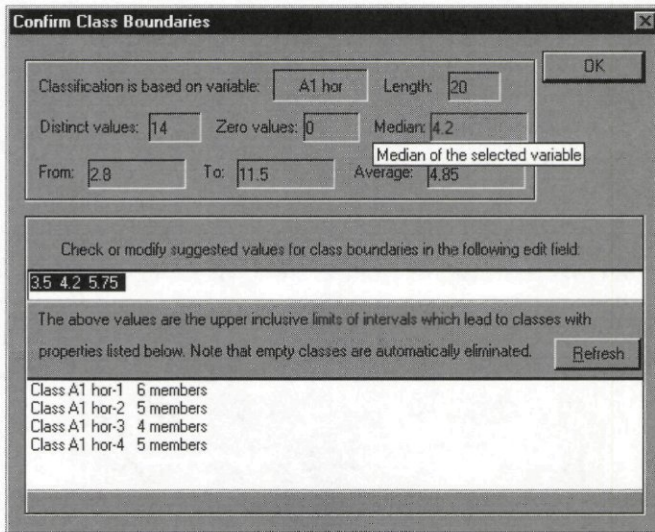


Figure 12-38 Dialog for specifying class thresholds (boundaries)

The area in the upper part of the dialog provides statistics summarising the variable selected as the classifying variable. Name of the variable, total number of values (entries), number of distinct values, and number of zero values are shown together with the minimum and maximum values, median, and arithmetical average.

The central part of the dialog is an editing field where the required boundary (threshold) values must be specified **in ascending order**. When you enter this dialog using the *Use selected thresholds* strategy, this field is pre-filled with three values representing lower quartile (0.25 quantile), median (0.5 quantile), and upper quartile (0.75 quantile) of the variable. If the lower quartile is not smaller than the median, it is not displayed. Similarly, if the upper quartile is not larger than the median, it is also skipped. If you entered this dialog by selecting the *Intervals of same length* or *Intervals with similar count* strategies, this field is initialised with boundary values (thresholds) calculated from those strategies.

The boundary values represent the upper, **inclusive** limits of the intervals. For example, if you use for classification a variable with only two values, 0 and 1, for example, you can specify 0 as the (only) boundary value, and all samples with value 0 fall into the first class, while those with value 1 (larger than 0) are in the second class.

You can check how many samples belong to individual classes, based on the currently specified boundary values, by inspecting the list area in the lower part of the dialog box. Note, however, that if you change the boundary values in the edit field, you must first click the *Refresh* button to update the list contents.

Combine dummy variables

This classification strategy is the only strategy enabled when multiple variables are selected and it is always disabled if only one variable is selected. CanoDraw allows you to select multiple variables if each of them has only 0 and 1 values. After you select a group of such variables, confirm the strategy and click the *Create* button. CanoDraw then creates separate classes for each selected variable and one or two additional classes: the class named *OTHER* collects the samples which have, for all the selected 0/1 variables, only zero values, while the class named *MissingEntries* collects samples with no available information for the selected variables. CanoDraw places the remaining samples into the class corresponding to the first of the selected variables which has value 1 for the particular sample. If you selected real dummy variables

representing collectively the levels of a factor, there is only one "1" value for each sample, but CanoDraw does not check whether the selected variables fulfil this condition.

After the classification is created using any type of classification strategy, CanoDraw scans through the new classes and removes the empty ones.

12.4.3.3 New from group

You can create a new classification from a group of samples or of species. A group of samples or species often represents a subset of items having a particular property. It can be defined manually (by explicit selection) or using a rule (see section 12.4.4 for additional details). CanoDraw displays a list of existing groups, as shown in Figure 12-39.

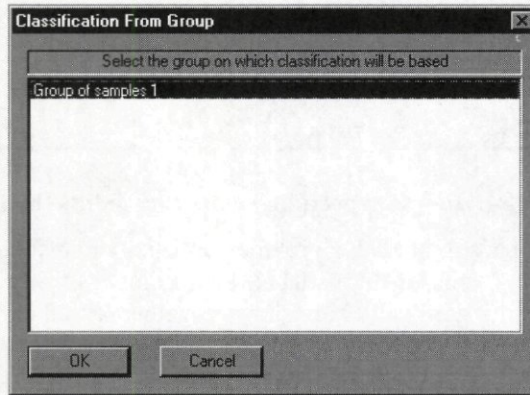


Figure 12-39 Classification From Group dialog

After you select one of the groups and click the *OK* button, a new classification with two classes is created. One class contains items (samples or species) which are members of the selected group, while the other class contains the remaining ones. The ability to create a class from a group is particularly useful because the range of possible criteria for creating groups is wider than for the classes (including ordination diagnostics, scores on ordination axes, summary properties of primary data, etc).

12.4.3.4 Edit

If you click on this button, a dialog similar to the one shown in Figure 12-32 is displayed, allowing you to add new or delete existing classes, change membership of items in the classes, or change names of classes.

12.4.3.5 Rename

Use this button to rename the currently selected classification.

12.4.3.6 Delete

Deletes the classification currently selected in the list.

12.4.3.7 Use this classification in diagrams

This checkbox indicates that the items will be displayed with different symbols (as long as the symbol types, colours, and/or sizes assigned to individual classes differ) in the diagrams created in this project. Note that either none or exactly one of the existing classifications may be active at any time. Therefore, if you have more than one classification defined and you select a classification different from the active one and check this box again, the original classification ceases to be active.

12.4.3.8 Copy

Copies definition of the currently selected classification to the Clipboard. The definition can be then pasted into another CanoDraw project, assuming that the identity of classified objects (samples, species, etc) is comparable between the source and destination project. If you, for example, copy a definition of species classification and then you paste it into another CanoDraw project, it is assumed that the species with index, say, 20 in the two projects refers to the same biological species.

12.4.3.9 Paste

Creates a new classification by importing it from the Clipboard. It is assumed that the classification was placed on the Clipboard using the *Copy* button in this dialog box, but when a different CanoDraw project was active. Because only the indices of samples, species, or explanatory variables are copied, they must refer to identical entities in both the source and destination project.

12.4.3.10 Close

Closes the dialog managing existing classifications.

12.4.4 Define Groups of

Groups allow you to specify subsets of samples or species and work with them. Groups can be used to limit the set of plotted items (see section 12.4.1.3), to define a classification of items (see section 12.4.3.3), and to visualise the importance of species groups in individual samples (see description of the *XY(Z) Plot* command, section 12.5.5.3).

Commands in this submenu allow you to manage existing groups of species or samples and create new ones. Both commands display primarily a group manager dialog (illustrated in Figure 12-40), which can be used to define new groups, combine existing ones, change their definition, or delete them.

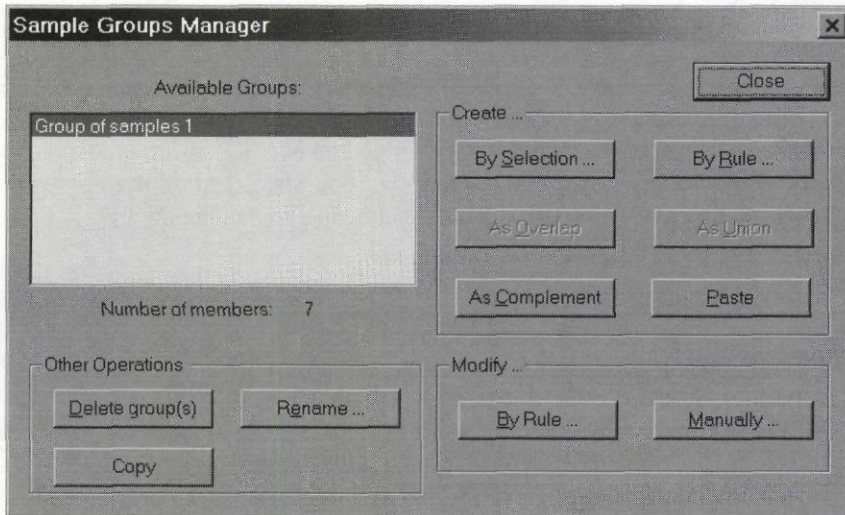


Figure 12-40 Group Manager dialog

The list in the upper left part of the dialog (titled *Available Groups*) shows the currently defined groups and you can select one or more groups there. For the currently selected group, CanoDraw displays information about its size immediately under the listbox. The other commands, accessible via the dialog buttons, are described in the following sections.

12.4.4.1 Create Groups

Commands in this area can be used to create a new group either from scratch or based on the existing one(s).

By Selection

This command creates a new, empty group, displays a list of available items (species or samples), and allows you to select which items should become group members.

By Rule

This command displays a dialog where you can specify a rule, defining item membership in the new group. The content of this dialog is described below, in the section named *By Rule* in *Modify Groups* (see 12.4.4.2).

As Overlap

This command is enabled only if two different groups are selected in the listbox. It creates a new group with members defined as items, which are members of **both** selected source groups.

As Union

This command is enabled only if two different groups are selected in the listbox. It creates a new group and all the items, which are members of at least one of the two selected groups, become members of this new group.

As Complement

Creates a new group as a complement of the currently selected group. This means that only the items, which were not members of the original group, become members of the new one.

Paste

Create a new group by importing its definition from the Clipboard. It is assumed that the group definition was placed on the Clipboard using the *Copy* button from this dialog box, but

when a different CanoDraw project was active. Because only the indices of samples, species, or explanatory variables are copied, they must refer to identical entities in both the source and destination project.

12.4.4.2 Modify Groups

The two commands allow you to modify the definition of an existing group.

By Rule

For groups with membership based on a selection rule, this command allows you to define or change this rule. For groups specified by direct selection, their original definition is discarded and the group is redefined from scratch.

In both cases, a dialog illustrated in Figure 12-41 is shown.

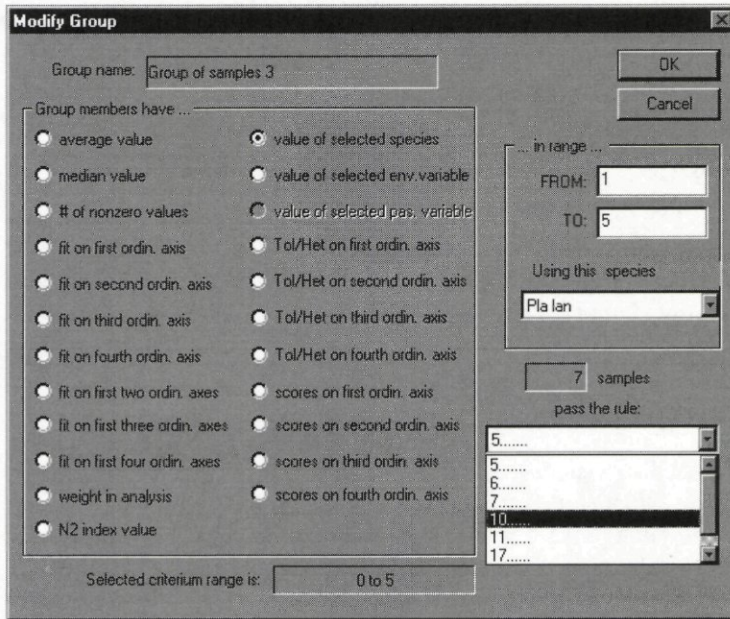


Figure 12-41 Dialog for group definition by a rule

The dialog shows the group name in the top left corner. The largest area with radio-buttons lists the possible criteria (statistics) on which a selection rule may be based. These include average values (or medians) from individual samples or for individual species, number of non-zero values (i.e. absolute frequency of species), fits of species or samples (per individual axes or cumulatively expressed), values of a selected species or environmental / supplementary variable (only for groups of samples), scores on ordination axes and *Tol* or *Het* values from the ordination output (solution file). Note that many of these options may be disabled (non-accessible) either because they are not appropriate for the particular kind of items or because the particular statistic is not available (incompatible type of analysis, missing source data, etc.).

After you selected one of the options (type of inclusion rule), the range of values is updated in the field at the dialog box bottom. Additionally, the fields in the area named *In range* (below the *OK* and *Cancel* buttons) are initialised to specify the full range of values, covering therefore all the items. These two fields (*FROM* and *TO*) are used to limit the range of values, which implies membership of each sample or each species in the group.

The rules named *value of selected species*, *value of selected env. variable* or *value of selected pas. variable* require, if selected, specification of the variable to use. This variable can be selected in the *Using this XXX* roll-down list in the middle right part of the dialog box.

The area in the lower right part of the dialog previews the group size and contents (list of items fulfilling the currently specified rule settings).

Manually

Allows you to explicitly specify which items are members of the modified group. If the group was originally defined using a rule, you are requested to confirm the change to the explicitly (manually) selected group.

12.4.4.3 Other Operations

Remaining commands for the manipulation of the group(s) are placed in this area.

Delete Group(s)

Removes one or more groups currently selected in the list.

Rename

Shows a dialog where you can change the name of the currently selected group.

Copy

Copies the group definition (list of its members) to the Clipboard. This definition can be then pasted into another CanoDraw project, assuming that the identity of group members is comparable (for identical indices) between the source and destination project.

12.4.5 Define Series of

Series can be used to visualise a spatial, temporal, or any other logical sequence of items (species, samples, or explanatory variable). A single series is presented in a CanoDraw diagram with a line connecting individual points in the order they have in the series. Individual series are grouped into series collection. A series collection for samples, for example, can represent repeated yearly measurements performed on permanent plots. Each permanent plot is then represented by one series and samples taken at that particular plot are ordered in the series by the year of sampling. Each item (e.g. sample) can be a member of multiple series in a series collection. Therefore, if your samples were collected in field in a spatial arrangement resembling a rectangular grid with, say, 5 rows and 6 columns, you can visualise the spatial contiguity of the samples in a CanoDraw diagram by defining (and displaying) 11 series. First 5 series (each with 6 samples) represent the rows, while the other 6 series (each with 5 samples) represent the columns of the grid. The defined series collections do not need to include all the available items.

When plotting a series in a diagram, CanoDraw does not draw a particular series if some of its members are not shown. This is because otherwise non-sequential series member could follow immediately in the displayed series sequence. You can overcome this behaviour in the dialog displayed by the *Project / Project Settings* command, in the page *Contents*, by checking the option *Plot series lines even for incomplete series* (see section 12.4.1.1).

When you select any item from this submenu, a dialog (illustrated in Figure 12-42) is displayed.

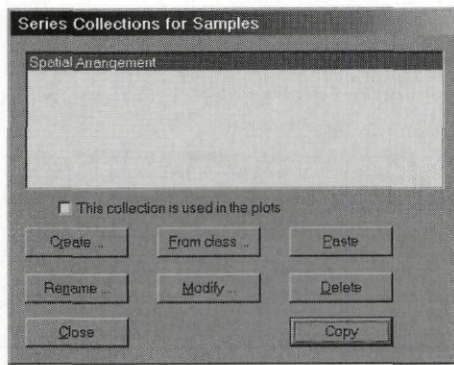


Figure 12-42 Series Collections dialog

The list in the upper part of this dialog shows the currently defined series collections for the particular type of items (species, samples, etc.). The checkbox *This collection is used in the plots* indicates whether the currently selected series collection is plotted within ordination diagrams created by CanoDraw. Only one series collection can be visualised at any time and CanoDraw switches the "activity" flag for series collections as needed. In the bottom half of this Series Collections manager dialog are the buttons allowing you to work with the existing series collections and to create new ones.

Create

Creates a new collection of series. A new dialog box is displayed (see Figure 12-43), where you can create new series and select the items they contain and the item order.

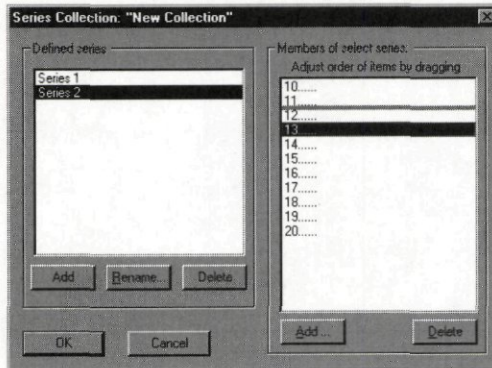


Figure 12-43 Dialog for editing series collection

The left half of this dialog shows the existing series within the series collection and allows you to add a new series (the *Add* button), delete the currently selected series (the *Delete* button), and change the name of currently selected series (*Rename* button). The right half of the dialog box contains the list of the items belonging to the series, which is currently selected in the left-hand list.

You can change the order of items by selecting the item you want to move and dragging it to the desired position. The dragging is achieved by pressing the left mouse button over the selected item and moving the pointer while keeping the button pressed. You cannot relocate list items beyond the last item, but the same effect can be achieved by dragging the last item upwards.

Use the *Add* button to add new items to the currently selected series. A currently selected item can be deleted either using the *Delete* button or by pressing the *Delete* key.

From class

Creates a new series collection based on the existing classification of items. CanoDraw displays first the dialog where you must select the classification to use for the definition of the new series collection (see Figure 12-44).

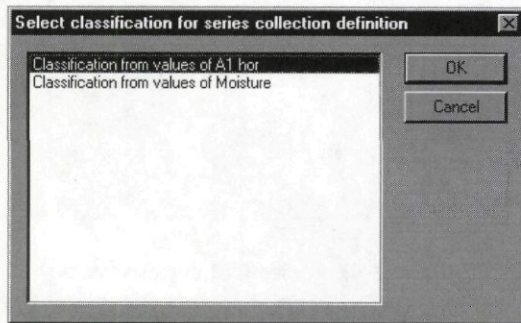


Figure 12-44 Dialog for selecting classification

After you have selected one of the classifications and clicked the *OK* button, CanoDraw creates a new series collection where each series corresponds to one of the classes and the class members are arranged in the series in their original order in the primary data.

Paste

Creates a new series collection by importing its definition from the Clipboard. It is assumed that this definition was placed on the Clipboard using the *Copy* button in this dialog box, but when a different CanoDraw project was active. Because only the indices of samples, species, or explanatory variables are copied, they must refer to identical entities in both the source and destination project.

Modify

Changes the definition of the currently selected series collection. The dialog box is similar to the one illustrated in Figure 12-43.

Close

Closes the series collections manager

Rename

Displays a simple dialog box where you can change the name of the currently selected series.

Delete

Deletes – after confirmation – the currently selected series collection.

Copy

Copies the definition of currently selected series collection to the Clipboard. The definition can be pasted into other CanoDraw project, assuming that the identity of classified objects (samples, species, etc) is comparable between the source and destination project. If you, for example, copy a definition of a sample series collection and then you paste it into another CanoDraw project, it is assumed that the sample with index, say, 5 is identical in the two projects.

12.4.6 Suppress

Use the commands in this submenu to select items of particular type (species, samples, environmental variables, and supplementary variables) which are not displayed in the diagrams

created by CanoDraw. Additionally, the suppressed items are also ignored when fitting regression models (GLM, GAM, and Loess) in CanoDraw.

12.4.7 Enforce

The enforced items are unconditionally included in the diagrams created by CanoDraw, even if they do not fulfil the other inclusion rules (fit into ordination space, item weight in the analysis, etc).

12.4.8 Import variables

CanoDraw provides three ways to import additional information about your samples and, eventually, species and explanatory variables, into an existing project. The imported variables are stored in the *Imported* folder, which is accessible from several places. The primary point of use is the dialog box for creating XY(Z) diagrams (see section 12.5.5.3), but you can also use the imported variables to define new classifications of samples.

12.4.8.1 From Clipboard

CanoDraw is able to import data from the Windows Clipboard, assuming the information is stored there in a standard format, resulting from copying a data table from a spreadsheet program. The table should have a rectangular form (with fixed number of columns and rows), with individual entries (columns) in each row separated by the TAB character, and the individual rows separated by the newline (NL) character. The first row should contain the names of individual variables (columns). The table being imported may contain indices for individual entries, which should be 1-based (i.e. the logically first entry should have value of 1). If the index column is present, it must be the first column in the table. Its presence provides greater flexibility in matching the values being imported with individual items available in the CanoDraw project – you can import information for a subset or a superset of the items present in the project.

When you select this command (which is enabled only if a suitable data format is detected on the Clipboard), a dialog is displayed, similar to the one shown in Figure 12-45.

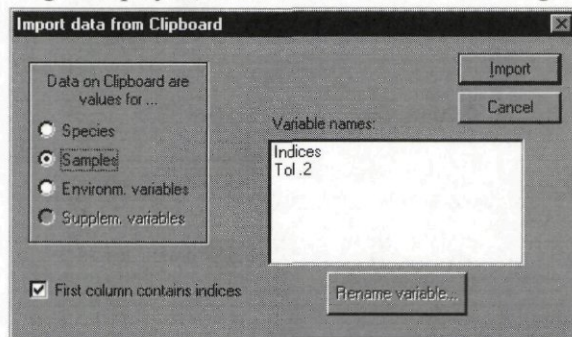


Figure 12-45 Import data from Clipboard dialog

You can import not only the information available for individual samples, but also information concerning individual species or environmental (or supplementary) variables. The type of items must be selected in the options field, in the upper left corner of the dialog.

CanoDraw disables the choices which are not compatible with the type of items present in the analysis (for example in Figure 12-45, the analysis does not contain supplementary variables, so the corresponding type option is disabled). Also, if you specify in this dialog that there is no index included in the imported table, then the number of rows (after accounting for the first row containing the variable names) must exactly match the number of items in the CanoDraw project.

In the dialog displayed in Figure 12-45, the option box *First column contains indices* is checked, so several item types are enabled, irrespective of the number of items present on the Clipboard.

In the list-box named *Variable names*, CanoDraw displays the column names parsed from the first row of the table present on the Clipboard. You can select one of those names and use the *Rename variable* button to change the name under which the variable will be imported into the CanoDraw project.

If the first column contained item indices, it will not be visible in the list of imported variables, but CanoDraw imports it and maintains its relation to the variables imported together with it.

You cannot influence which variables will be imported. All the data columns present on the Clipboard are imported. Note, however, that the typical usage context for this command involves preceding the selection of columns in a spreadsheet application (like Microsoft Excel®), so you can make the selection of imported variables there.

12.4.8.2 From Canoco file

CanoDraw enables you to import additional variables from data files in Canoco format. Note that this command is usually not needed for the data files which were used in the Canoco analysis on which the CanoDraw project is based. It is rather more useful if you have any additional information about the samples, which was not used in the original analysis.

After you have selected this command, CanoDraw displays dialog for selecting the file containing data in Canoco compatible format. After that, the file is parsed and on success a dialog similar to that in Figure 12-46 is shown.

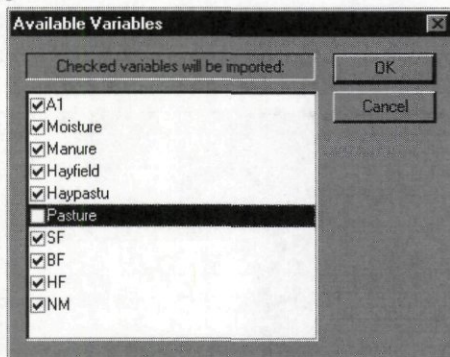


Figure 12-46 Dialog for importing selected variables from Canoco data file

All the variables found in the data file are listed, each with a pre-checked box. You should uncheck the boxes for all variables you do not want to import.

12.4.8.3 Setup PRC scores

Canoco for Windows version 4.5 provides only a limited support for creating principal response curve (PRC) diagrams, which were described in section 8.3.11 of this manual and in the paper of Van den Brink & Ter Braak (1999). To calculate the scores to be plotted in PRC diagram, you must combine information from the Canoco solution file with information provided in the Canoco output file (TAU value and the standard deviations of the environmental variables).

To setup PRC scores, you must have a CanoDraw project with appropriate options: it must be a redundancy analysis with covariables (partial RDA) where time (coded as a series of dummy variables for individual time points) is used as covariable(s) and the interactions between the time indicators and treatment variables are used as the explanatory variables ("environmental variables" in Canoco terminology). CanoDraw does not even enable this command if the project is not based on a constrained analysis with covariables.

Additionally, after you have executed the redundancy analysis in Canoco, you must save the analysis log (present in the Log View window) into a text file (presumably with the *.log* extension). This log contains the information, which is ultimately needed to calculate the PRC scores.

When you execute this menu command (*Setup PRC scores*), a dialog similar to the one illustrated in Figure 12-47 is displayed.

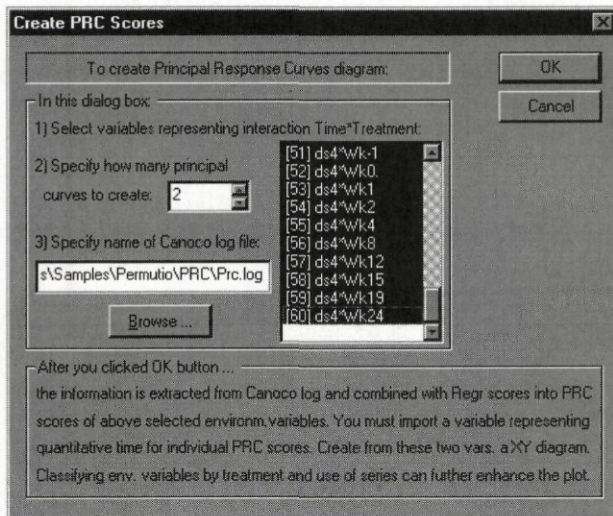


Figure 12-47 Create PRC Scores dialog

You must select in the list (placed in the central part of the dialog) the environmental variables representing the interactions between the dummy variables corresponding to individual time points and the dummy variables representing the treatments. Then you should specify how many sets of PRC scores are to be created. The first principal response curve corresponds to the first axis of RDA and provides the most important patterns in the community response to the treatments, but the higher order PRCs can also provide interesting information. Scores for the first PRC are stored among the imported variables under the name *PRC1*, scores for second PRC as *PRC2*, etc. Additionally, you should specify in this dialog box the name of the file with the stored Canoco log of the redundancy analysis, on which this project is based.

After you press the *OK* button, CanoDraw parses the log file and reports any encountered problems. It calculates the scores and stores them in the *PRCi* variables. The use of these scores is illustrated by the example *PRC_SIM* (section 14.8).

12.4.8.4 Delete

This command displays a simple manager of imported variables, where you can select one or more imported variables and remove them from the CanoDraw project using the *Delete* button.

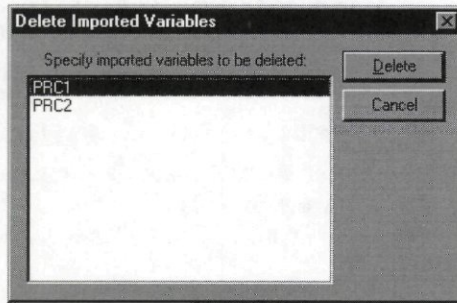


Figure 12-48 Delete Imported Variables dialog

12.4.9 Export Statistics

This command displays a dialog similar to the one shown in Figure 12-49 and containing a list of species data statistics available for export. These statistics are further described below. Originally, all the items in the list are pre-selected (their check-boxes are checked), so you must uncheck the boxes for the statistics you do not want to export. After you have made your selection, press the *OK* button to proceed. CanoDraw informs you about the successful accomplishment of the command or reports an error. The statistics are placed on the Clipboard in a TAB-separated text format, so they can be easily pasted into any spreadsheet application.

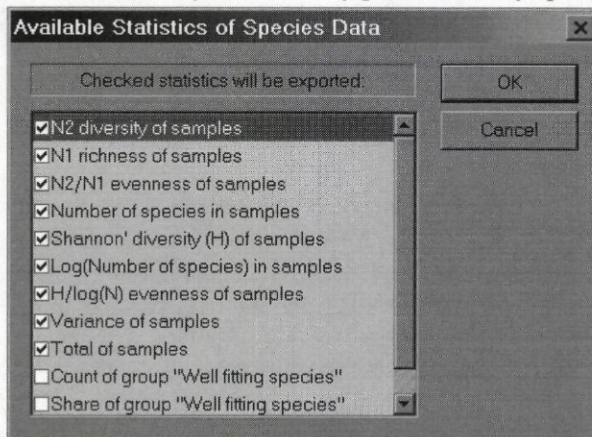


Figure 12-49 Export Statistics dialog

CanoDraw provides several species data statistics. For their description, you are advised to further study section 6.5 of Legendre & Legendre (1998). These sample statistics are always

calculated from the original, non-transformed species data, using all the species that are available there (ignoring any selection rules used in CanoDraw graphs). For samples with peculiar properties (usually with none or one species occurring), some of the statistics described below cannot be calculated and are replaced by some reasonably chosen *degenerate* values. In such case, CanoDraw places information about this failure into the project log window. The conditions leading to the use of degenerate values are noted (using text in *italics*) at the end of the section describing particular statistics.

N2 diversity of samples

The N_2 diversity statistics, introduced by Hill (1973), is equal to the inverse value of the Simpson concentration index (see Legendre & Legendre 1998, p. 242) and can be also interpreted as the number of "effective species occurrences" (see section 6.3.5).

A degenerate value of 0 is used if the sum of species values in a sample is equal to zero, i.e. for a sample without any occurring species.

N1 richness of samples

This is another measure from the system of diversity numbers, introduced by Hill (1973). It can be related to Shannon entropy statistics (H), discussed below, using the relation $N_1 = e^H$. Most often, this statistics is used as a measure of species **richness**. The other component of diversity – the **evenness** can be then calculated as a ratio N_2 / N_1 .

A degenerate value of 0 is used if the sum of species values in a sample is equal to zero, i.e. for a sample without any occurring species.

N2/N1 evenness of samples

The ratio of Hill's N_2 diversity number to N_1 richness number can be used as a measure of a compositional evenness of particular sample.

A degenerate value of 1 is used whenever the N_1 value for a sample is not greater than 0.0.

Number of species in samples

The number of species present in a sample is probably the simplest and most often used measure of sample richness. Formally, it can be labelled as Hill's N_0 coefficient. The number of species in a sample is always larger or equal to N_1 , as it represents its upper bound (maximum value), achieved with maximum evenness of species composition.

Shannon' diversity (H) of samples

This is the familiar entropy measure. It is calculated by first taking the sum of species values in a sample, referred to as SUM and then calculating the H statistics from the relative proportions of the abundances of individual species, $p_j = Y_j / \text{SUM}$, where Y_j is abundance of the j-th species in the sample. H is then calculated as: $H = -\sum p_j \log(p_j)$ where $\log(x)$ is the natural (base e) logarithm.

A degenerate value of 0 is used if the sum of species values in a sample is equal to zero, i.e. for a sample without any occurring species.

Log(Number of species) in samples

The logarithm of number of species occurring in a sample has a similar relation to Shannon' entropy statistics H as the (non-transformed) number of species (N_0) has to Hill's N_1 coefficient. It represents the maximum achievable value of the H statistics for a given number of occurring species.

A degenerate value of 0 is used if there is no species occurring in the sample. Consequently, samples with none or one species end up with an identical value of this statistics

H/log(N) evenness of samples

This ratio represents a widely used statistic for compositional evenness, with values between zero and one.

A degenerate value of 0 is used if the number of species in a sample is less than two, i.e. for samples without any or with just one occurring species.

Variance of samples

Calculates the variance of the species values in a sample, around the sample average. Note that the statistical **sample** variance is calculated (dividing the sum of squares by the number of species in the data decreased by one) and also that the zero values (for species not occurring in the sample) are included in the calculation.

There are no degenerate values generated, but the calculation would fail for species data with just one species.

Total of samples

Calculates the sum of species values in each sample.

Count of group "XXX"

This statistics is (together with the following one) offered for each group of species defined in the current CanoDraw project. *Count of group* statistic calculates how many species from the specified group occur in a particular sample.

Share of group "XXX"

For a particular group of species, this statistic calculates the percentage of the total sum of the abundances belonging to species from that particular group. The values are on the scale from 0 to 1, with a value of 0.5 implying, for example, that half of the sum of values in that sample represents the species which are members of that particular species group.

A degenerate value of 0 is used for samples with a non-positive total of the species abundances.

12.4.10 Manage graphs

This command displays a dialog listing the CanoDraw graphs created from the active CanoDraw project, saved to a permanent file, and currently not opened in the application workspace. An example of this dialog is shown in Figure 12-50.

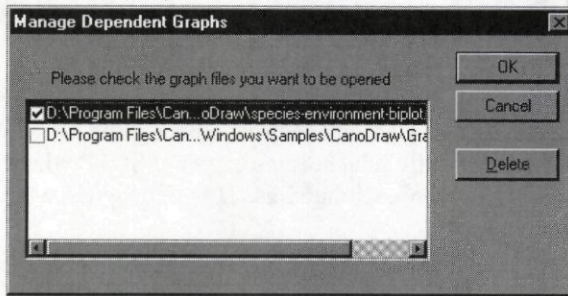


Figure 12-50 Manage Dependent Graphs dialog

This dialog can be also used to remove reference to the created graph files: select the desired entries in the list and click the *Delete* button. Note that only the reference will be removed, not the actual files. More importantly, you can use the checkboxes in the front of the file names to specify which of the referred graphs should be opened. After you click the *OK* button, the graphs stored in the checked files will be opened.

12.5 Create

All the commands used to create graphs in CanoDraw are placed in this menu. Execution of any of these commands results in the creation of a new graph window, with its initial name consisting of the word *Graph*, followed by a sequential number. These numbers are unique for the current CanoDraw session (i.e. during the time between opening and closing the CanoDraw application), but are restarted with a new session. When you save a newly created graph to file, CanoDraw suggests to use this initial label, supplemented with the *.cdg* extension, as the file name. You can change it to any other name, however.

The diagrams are initially represented by the *Graph* view, but you can also display an additional Graph Contents view (using the command *View / Tree view*). See section 11.3 for additional information.

This reference guide does not provide any systematic explanation of the rules for interpreting the contents of ordination diagrams. These rules can be found in a variety of publications, including the section 3.5 of this manual, p. 39), Ter Braak (1994) for ordination diagrams resulting from linear ordination methods, or Ter Braak & Verdonschot (1995) for ordination diagrams resulting from ordinations based on the unimodal species response model. Note, however, that CanoDraw for Windows provides an utility for suggesting an interpretation of any *ordination* diagram – see section 13.5 for more details.

12.5.1 Simple Ordination Plot

This command provides the easiest way to create a traditional ordination diagram, but only the most frequently used types are available from here.

CanoDraw displays a dialog box illustrated in Figure 12-51.

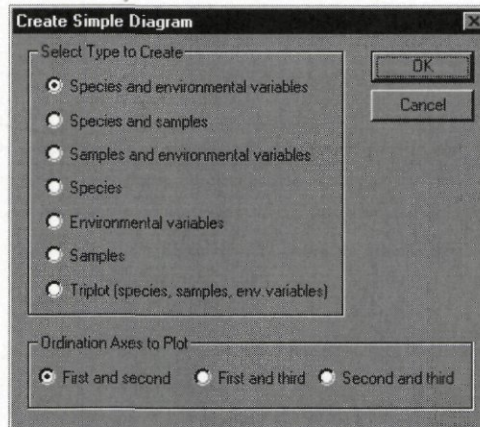


Figure 12-51 Dialog for creating simple ordination diagram

You must choose the type of the ordination diagram to be created and which ordination axes to plot (this option is located in the lower part of the dialog).

CanoDraw normally uses the ordination axes specified in the *Project Settings* dialog, on its *Contents* page (see section 12.4.1.1). Only here can you temporarily override this setting and specify the plotted pair of ordination axes directly. The Simple Ordination Plot command supports plotting only the first three ordination axes, however.

12.5.2 Scatter Plots

Commands in this submenu create simple diagrams containing only one kind of items (species, samples, environmental variables, or supplementary variables) at a time. The items are represented by symbols. The **symbols** are always used for plotting samples, and environmental or supplementary variables if they are specified as nominal variables. Species are plotted as symbols only in projects based on a weighted-averaging ordination method (CA, DCA, CCA, and DCCA). In linear ordination methods, the tips of the vectors (arrows) representing the species can be exceptionally displayed as symbols (check section 12.4.1.2 for more details). This type of presenting species in PCA or RDA ordination space is generally not recommended, however. **Arrows** are used for species in results from linear ordination methods and for (semi-) quantitative environmental or supplementary variables.

If symbols are plotted, the items can be represented by symbols of varying appearance (different symbol types, colors, and sizes). This differentiation of symbol types is governed by the currently active classification of items. Only one (or none) classification may be active at any time for each of the item types. To use a classification, you must first create it, and then you can specify the active classification from the *Available Classifications* dialog (see section 12.4.3.7). The symbol graphical attributes can be specified per class, using the dialog displayed by the *Visual Attributes* command in the *View* menu (see section 12.3.2). Specification of symbol attributes is illustrated by Figure 12-15).

As can be seen from Figure 12-15, the visual attributes of arrows can be also differentiated among the individual classes of species and / or explanatory variables.

If the plotted items are represented by symbols, there are three additional types of visual information, which can be encoded into the plots: envelopes, series, and pie symbols.

The **envelopes** represent the borderline of the area, in which all the symbols belonging to a particular class lay. The envelopes are drawn as the smallest convex polygons using the outermost symbols of the group as their vertices. The envelopes are displayed when the option for drawing envelopes around the items of particular kind is checked (see section 12.4.1.1). Each envelope is drawn with a solid coloured line, using the drawing color of symbols of the particular class.

You can also display **series** of items using lines connecting series members in their specific order. Each item can be a member of one or several series or it may not belong to any of the defined series (see section 12.4.5 for a more detailed discussion). Series have a separate subgroup of attributes in the *Visual Attributes* dialog box, but you can diversify series lines only for the first 16 series in a series collection. Seventeenth and following series take the attributes of the first series.

Pie symbols may replace the standard CanoDraw symbols, but only for samples or species. For each sample (or species) pie symbol, a traditional pie-chart circle is displayed which shows the distribution of species (or samples) quantities or presences across the individual classes of species (or samples). Therefore, the pie symbols are applied only if a classification of complementary items is defined and active. For example, to replace the standard sample symbols with pie symbols, you must have an active classification of species. Each pie symbol will then be divided into slices, indicating how large part of the total abundance or how large

fraction of the occurring species belongs to individual species class. Similarly, pie symbols for species report, for a classification of samples, how large a fraction of the sum of that species abundances over the whole data set occurred in samples from individual sample classes or what was the fraction of species occurrences in samples belonging to individual sample classes. The use of pie symbols is governed (beside the availability of an active classification of complementary items) by the checkboxes in the *Contents* page of the dialog invoked by the *Project / Project Settings* command (the section *Use Pies instead of Symbols*). Whether the species occurrences or abundances are used is determined by the attached option named *With slices based on* with two choices (*Values* or *Presences*). If you want to change visual attributes of the whole pie symbols (size and outline properties) or of the wedges corresponding to individual classes, you should open the *Visual Attributes Settings* dialog (see section 12.3.2) and go to the *Pie Wedges* group. This group has two subgroups, one for samples and the other for species. In both subgroups, the first item is named *Size and outline* and allows you to specify the line color and style as well as pie symbol size. **Note that you cannot change the symbol type from Circle. If you do, CanoDraw displays a warning dialog box and reverts back to a circle.** The other items in the two subgroups provide the possibility to change the properties of the fill style for wedges of up to 64 classes, supported in a CanoDraw classification. You should note that the class entries, for example in the *Sample pies* subgroup, refer to classes from a classification of species.

The symbols and/or the arrows may be labelled with the 8-character labels taken from the Canoco project output (the "SOL" file), or by their indices, or they can be plotted without any labels. The options for selecting among these three possibilities for each type of items are available in the *Appearance* page of the *Project Settings* dialog (see section 12.4.1.2, *Labelling of scores* group). These settings are used when plotting symbols or arrows in ordination diagrams. The settings are **not** applied to graphs created using the commands in the *Create / Attribute plots* submenu, where the labelling method can be specified within their respective setup dialogs (see sections 12.5.5.1 and 12.5.5.3).

12.5.2.1 Species

The scatter diagram of species can display either the positions of individual species in the plotted ordination plane (for results from weighted averaging ordination methods) or directions of the fitted steepest increase of values of individual species, indicated by arrows (for results from the linear ordination methods).

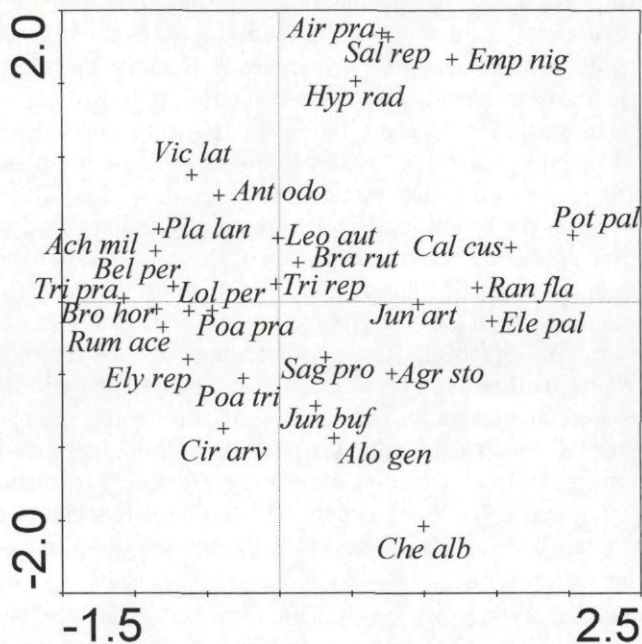


Figure 12-52 Scatter of species symbols

To limit the set of species plotted in such scatter-plot, you can use the rules conditioning species appearance by their fit in a regression model using ordination axes as predictors or by the weights they had in the analysis (for weighted averaging ordinations only) – see section 12.4.1.3 for additional information. In the same dialog, you can also specify a single group of species and then only the species belonging to this group will be plotted (see section 12.4.4). You can also enforce the plotting of the species not qualified to appear in the diagram, based on those rules (see section 12.4.7) or explicitly exclude particular species from the diagrams (see section 12.4.6).

12.5.2.2 Samples

Samples are always shown in ordination diagrams as a scatter of symbols (or pie-symbols, if specified), as illustrated in the sample diagram in Figure 12-53.

In constrained ordination methods (also called direct gradient analysis), two types of sample scores are available: the scores based on species scores (*Samp* scores) and the scores defined as a linear combination of environmental variables (*SamE* scores, also called *fitted sample scores*). For further explanation, see section 6.3.3 of this manual (p. 156), or Legendre & Legendre (1998), section 11.1.1 (p. 584). CanoDraw for Windows displays for constrained ordination methods (RDA and CCA) usually the *SamE* scores, if they are available (i.e. if canonical [=constrained] axes are used and not for the supplementary samples). To change this behaviour, use the option *Plot SAMP scores even...* in the *Contents* property page of Project Settings dialog (see section 12.4.1.1).

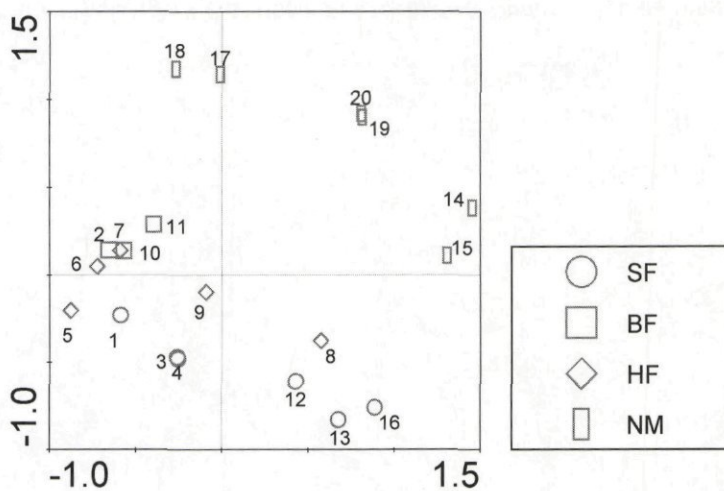


Figure 12-53 Scatter of sample symbols, with symbol type coded by the management type

To limit the set of samples which are drawn in scatter plots, you can use the rules conditioning samples appearance by their fit in a regression model using ordination axes as predictors or by their weights in the ordination analysis (for weighted averaging ordinations only) – see section 12.4.1.3 for additional information. At the same dialog page, you can also specify a single group of samples and then only the samples belonging to the selected group will be plotted (see section 12.4.4). You can also enforce the plotting of the samples not qualified to appear in the diagram based on the above listed rules (see section 12.4.7) or explicitly exclude particular samples from the diagrams (see section 12.4.6).

12.5.2.3 Environmental variables

12.5.2.4 Supplementary variables

The scatter plots for those two types of explanatory variables are described jointly, because there is no difference in their treatment, they only have a different default visual appearance of their symbols and arrows.

Unlike species and samples, a scatter plot of explanatory variables often contains two different visual presentations at the same time (as illustrated by the diagram in Figure 12-54). This is due to the difference between so-called **nominal** explanatory variables and the (semi-)quantitative variables. Nominal variables usually possess only two distinct values, typically 0 and 1. A group of nominal variables can be used to represent a single factor, with each factor level represented by a single nominal variable. In such case, the nominal variables are best represented by the *centroids of environmental variables (CenE)*, which are calculated by Canoco also for the supplementary variables, if they are present in the analysis. These centroids represent (weighted) averages of scores of the samples which had the value of 1 for that particular nominal variable (i.e. they possessed that particular factor level / belonged to that particular class).

The other explanatory variables, which cannot be interpreted in the way just described, are supposed to be quantitative ones, or at least to be on an ordinal scale. When you create a new CanocoDraw project (by importing an existing Canoco project), all explanatory variables are treated as (semi-)quantitative ones. To designate a variable as nominal, you should use one of

the two commands in the *Nominal variables* submenu in the *Project* submenu (see section 12.4.2).

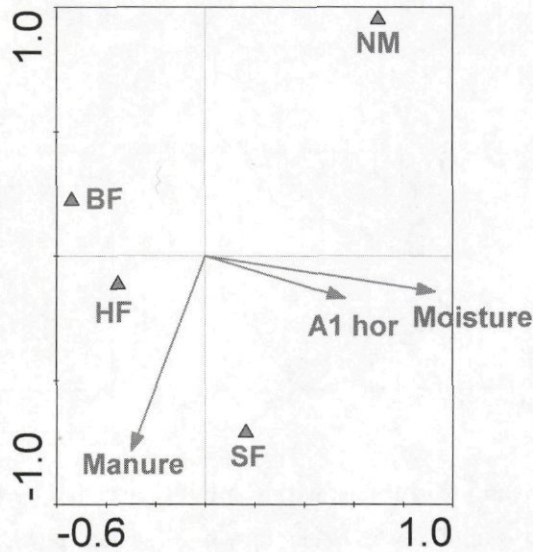


Figure 12-54 Scatter of explanatory variables represented by arrows (quantitative variables) and symbols (dummy variables)

To limit the set of explanatory variables which are drawn in scatter plots, you can use the rules conditioning their appearance by their values of T statistics calculated by Canoco or by their correlations with the ordination axes appearing in the diagram – see section 12.4.1.4 for additional information. You can also enforce plotting of the variables not qualified to appear in the diagram based on the above listed rules (see section 12.4.7) or explicitly exclude particular variables from the diagrams (see section 12.4.6).

12.5.3 Biplots and Joint Plots

While the scatter plots of particular kinds of items provide the simplest summaries of the ordination results (displaying similarity among the samples, correlations among the explanatory variables, etc.) the most interesting information is provided by diagrams where two or more kinds of entities are plotted together: biplots, joint plots, and triplots (see section 12.5.4). These diagrams summarise in a few dimensions the original primary data table ("species data"), relations between variability in community composition and explanatory variables, or similar types of information. References to publications discussing the use of biplots (and joint plots) in detail are given in the introductory remarks of section 12.5.

When specifying the contents of ordination diagrams with multiple kinds of items, you should proceed as described in the preceding section about *Scatter Plots* submenu (section 12.5.2). But here the contents of individual scatters are superposed in the same ordination diagram. The scaling of different kinds of items is not always on a comparable common scale and therefore CanoDraw provides an algorithm for choosing the respective re-scaling of the ordination scores, depending on the type of ordination analysis and the type of items combined in an ordination diagram. See the description of the option named *Show rescaling coefficients for composite ordination diagrams* in section 12.3.1.1 for further explanation.

If you combined several kinds of items in the same ordination diagram, particularly if optional features are included (like series lines, envelopes, pie symbols), the diagrams become quickly complicated even for small data-sets. In such a situation, it becomes useful to supplement the ordination diagram with a **legend**, displaying a graphical summary of the visual attributes used to plot different kinds of items (see the Figure 12-55 below), or even of classes of such items. Reference material concerning the legends in CanoDraw diagrams are available in the section 12.4.1.2. Additional description of the concepts behind the legends in CanoDraw diagrams and procedures needed for their efficient use are described, respectively, in sections 13.4 and 13.3.

12.5.3.1 Species and env. Variables

This command creates an ordination biplot diagram containing species and environmental variables. Such a biplot is illustrated in Figure 12-55.

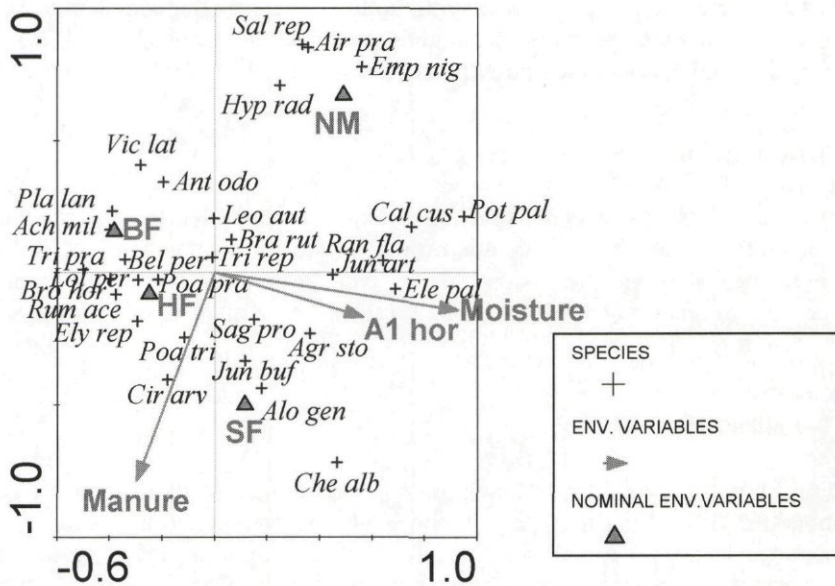


Figure 12-55 Biplot with species and environmental variables, based on a Canonical Correspondence Analysis (CCA)

12.5.3.2 Species and samples

This command creates an ordination diagram displaying, at the same time, species and samples. This diagram can be called a biplot or a joint plot, based on the interpretation rule appropriate for the ordination method used and the data properties. Anyway, interpretation of this diagram can lead (at least in theory) to a summary of the primary ("species") data table. The problem concerning the two types of sample scores that may be used in ordination diagrams is explained in the section about scatter plots of samples (12.5.2.2).

12.5.3.3 Samples and env. variables

The command creates a diagram displaying both the samples and environmental variables. The problem concerning the two types of sample scores that may be used in ordination diagrams is explained in the section about scatter plots of samples (12.5.2.2).

12.5.3.4 Species and suppl. variables

This command creates a diagram combining the presentation of species scores (shown either as points or as arrows) with supplementary variables, which can be presented as centroids or as arrows or a mixture of those two types (see section 12.5.2.3).

12.5.3.5 Samples and suppl. variables

The command creates a diagram displaying both the samples and supplementary variables. The problem concerning the two types of sample scores that may be used in ordination diagrams is explained in the section about the scatter plots of samples (12.5.2.2).

12.5.3.6 Environm. and suppl. variables

Displays a diagram where both types of explanatory variables (environmental variables, used as predictors in the fitted ordination model; and supplementary variables, *post-hoc* projected into the resulting ordination space) can be present. Both environmental and supplementary environmental variables can be presented by a mixture of centroid symbols and arrows (see section 12.5.2.3).

12.5.3.7 T-values biplot

T-values biplot is a special type of diagram approximating a table of T value statistics, each one corresponding to a simple regression model with one predictor (explanatory variable) and one response (species). You can find the species with an important response to the particular explanatory variables using the interpretation rules for the T-values biplot (see section 6.3.12, starting on p. 179 of this manual, and Ter Braak & Looman 1994). Beside drawing the T-values biplot scores for species and explanatory variables, CanoDraw offers two additional options, which can be specified in a dialog that appears after selecting this command (see Figure 12-56).

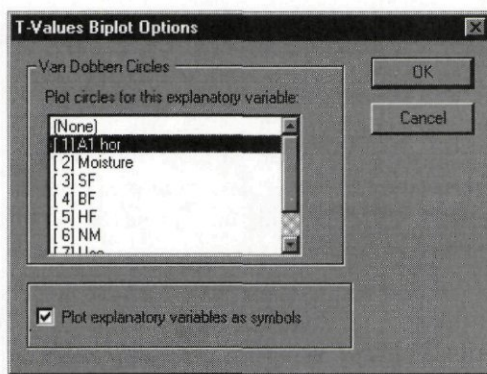


Figure 12-56 T-values Biplot Options dialog

First, CanoDraw allows you to plot so-called Van Dobben circles for a selected variable. This is a pair of circles, touching each other at the origin of coordinate system. Species that are represented by arrows which end within one of these two circles are predicted to have their T-value statistic larger than 2.0. This indicates a low probability of Type I Error when testing the null hypothesis of the regression coefficient (describing the relation of a species to the selected explanatory variable) being equal to zero.

Second, the explanatory variables can be shown as symbols, not as arrows. This facilitates an alternative method of interpreting the T-values biplot - perpendicularly projecting the explanatory variables onto species arrows (see the above quoted references for more details).

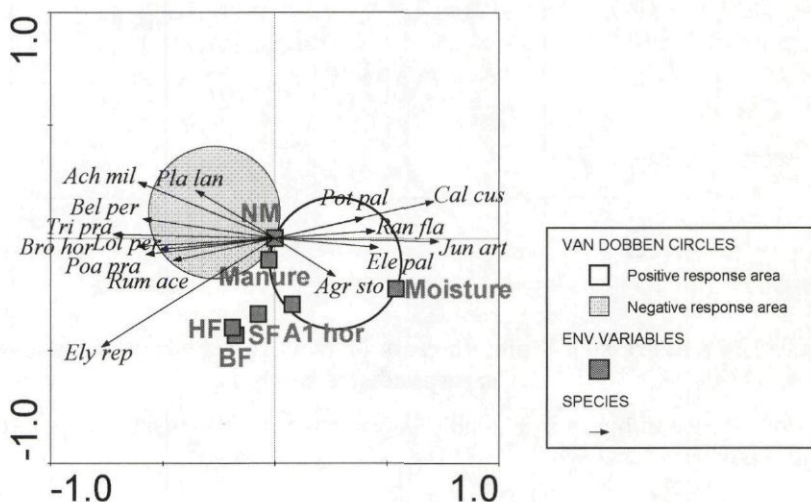


Figure 12-57 T-values biplot diagram

The visual attributes of the Van Dobben circles are pre-set to a hollow red circle for the circle enclosing species with a preference for higher values of the explanatory variable, and to a hollow blue circle for those with a preference for lower values of the explanatory variables. The circle properties can be manipulated post-hoc, e.g. by changing their fill style, as illustrated in Figure 12-57.

12.5.3.8 Regression biplot

A regression biplot (illustrated in Figure 12-58) can be used to approximate individual regression coefficients of a multiple regression model, where one of the species (represented by points in our example) is used as the response variable and all the environmental variables are used as predictors (explanatory variables). The relative extents and signs of regression coefficients can be deduced by the perpendicular projections of that species symbol onto individual arrows of environmental variables.

It should be noted that the imaginary regression model referred by the preceding description differs between weighted averaging and linear ordination methods. In both cases, the environmental variables were standardised to zero mean and unit variance. For linear ordination methods (PCA, RDA), the species are usually centred and, optionally, also standardised to unit variances. In weighted averaging methods (CA, DCA, CCA), the relative abundances (proportions of individual species within each sample) are used in the response variables and the weighted linear regression is calculated, using the standard sample weights to weight individual observations.

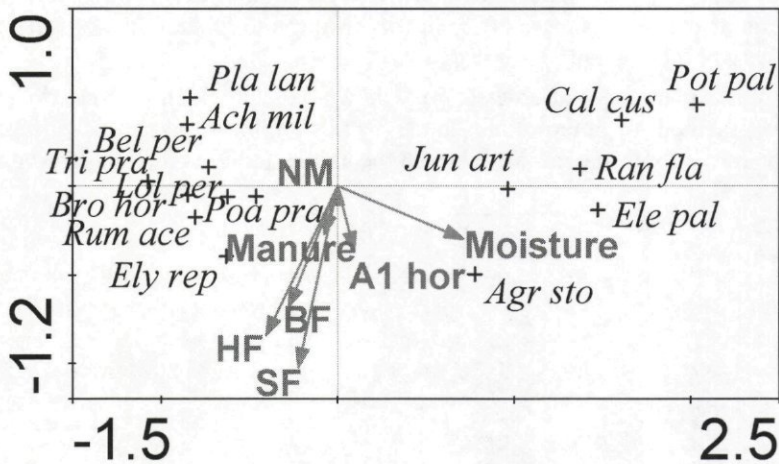


Figure 12-58 Regression Biplot diagram from a project based on Canonical Correspondence Analysis.

Creation of regression biplot is available only for the environmental variables, not for the supplementary variables.

12.5.3.9 T-values biplot for suppl. variables

This command creates a diagram comparable with the T-value biplot for environmental variables, described in section 12.5.3.7, except that supplementary variables are plotted instead of environmental variables.

12.5.4 Triplots

The triplots submenu collects commands for creating ordination plots containing three or four different kinds of items at the same time. The most often used type, a triplot with samples,

species, and environmental variables (some represented by arrows, some by centroid symbols), is illustrated in Figure 12-59.

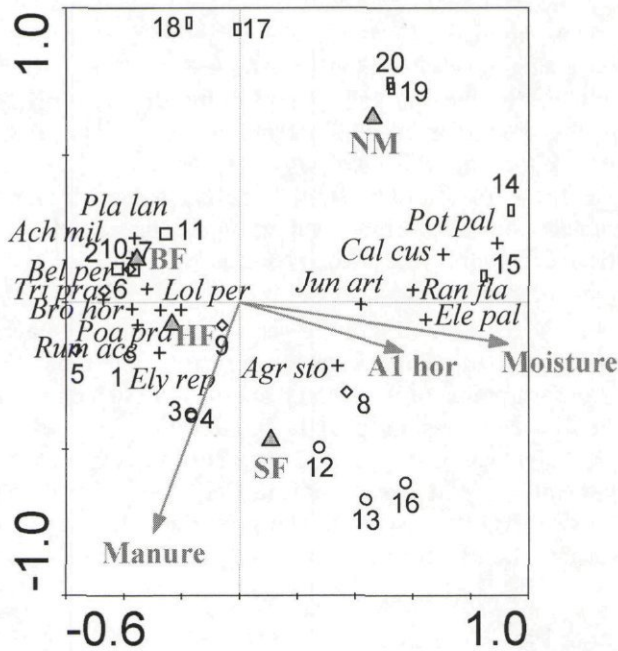


Figure 12-59 Triplot diagram

A triplot can be viewed as a superposition of ordination scatter-diagrams for individual kinds of items, except the scores in individual sets can be stretched or shrunk before being combined with other kinds of items, to facilitate easy interpretation of the diagram. See introductory comments of *Biplots and Joint plots* section (12.5.3) for additional information.

12.5.4.1 with Environmental variables

Creates a triplot with species, samples, and environmental variables.

12.5.4.2 with Supplementary variables

Creates a triplot with species, samples, and supplementary variables.

12.5.4.3 with Env. and suppl. variables

Creates an ordination diagram with four kinds of items: species, samples, environmental variables, and supplementary variables.

12.5.5 Attribute Plots

This submenu collects the commands that create diagrams, which cannot be called ordination diagrams, in the strict meaning of the term. Nevertheless, you will see later that the boundary between attribute plots and standard ordination diagrams is very blurred and many

ordination diagrams can be also plotted using the procedures for creating attribute plots. There are two principal categories of diagrams created with commands in *Attribute Plots* submenu.

The first category of diagrams starts from a simple scatter-plot of sample⁺ scores and enhances the information provided by the positions of sample points using an additional **attribute**, representing a single variable. In the simplest variation of this diagram are the relativised values of the attribute variable encoded by the size of symbols representing individual samples in the ordination plane. Depending on the type of plotted attribute, this diagram can be created either by the *Data Attribute Plot* or by the *Results Attribute Plot* commands. Alternative forms of this diagram can be created by fitting a regression model, describing the dependency of the attribute values upon the sample positions on the two ordination axes. Instead of plotting the actually observed values of the attribute, the fitted regression surface is shown, using a contour plot.

General XY scatter diagrams represent the second category of diagrams where any two selected variables can define the horizontal and vertical axes. Additionally, a regression model can be fitted to this scatter of points or you can create diagram, which is a generalisation of the attribute plots described in the preceding paragraph. Here the points of the XY diagram have their size varying with the values of the third (Z) variable, so we speak about a XYZ diagram here. Alternatively, we can again fit a regression model, where the X and Y variables act as predictors and the Z variable is the response. The procedures leading to the creation of such diagrams are described in section 12.5.5.3.

12.5.5.1 Data Attribute Plot

This and the following (12.5.5.2) commands share most of the contents of their setup dialogs, but also the meaning of the attribute plots, which they create. These two commands differ only by the kind of variables, which are offered as attributes for the diagrams. Therefore, their shared functionality is described in substantial detail here, but applies also to the commands described in section 12.5.5.2.

An example of the *Attribute Plot Options* dialog is shown in Figure 12-60.

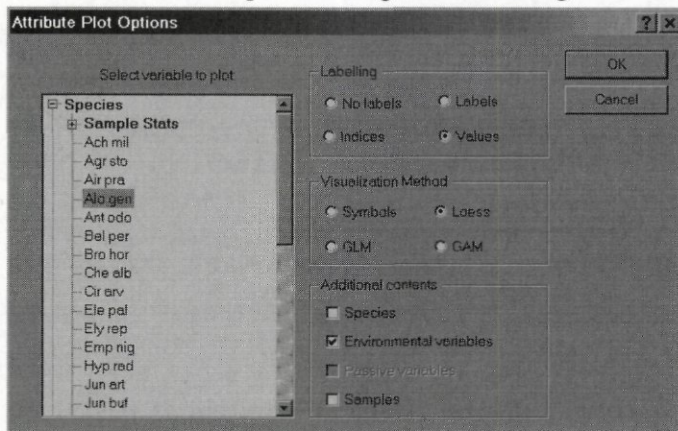


Figure 12-60 Data Attribute Plot Options dialog

The dialog has a list of variables, which can be used as an attribute in the diagram. To the right of this list are three groups of options which you can use to adjust the labelling of symbols

⁺ or, exceptionally, scores of species or explanatory variables

(if any are plotted), type of the fitted regression model, and additional items to be shown in the diagram.

For data attribute plots, the list contains primarily the individual species, environmental and/or supplementary variables (if present), and covariables, retrieved by CanoDraw from the Canoco data files, which were used in the analysis. In addition, the list of available species is preceded by a sub-list named *Sample Stats* which contains some summary characteristics of the samples, based on the species composition recorded in the species data. These characteristics were already described in section 12.4.9.

Where the attribute variable is visualised by a varying size of symbols (the **symbol attribute plots**, see Figure 12-61 for an example), the options in the *Labelling* group enable you to specify the method of labelling those symbols. Note that the labelling options do not apply to the items included in the attribute plot with the choices in the *Additional contents* options group, described below. Instead, the same settings, which are used for the standard ordination plots (see section 12.4.1.2), are used here.

The group of options labelled *Visualization Method* determines whether a symbol attribute plot or a contours attribute plot will be created. In the latter case, CanoDraw offers a selection of three different kinds of regression models, discussed elsewhere (see section 13.6).

The *Additional contents* options allow you to enhance the diagram interpretation by including additional kinds of items in the ordination plots. If you are creating contour-based attribute plots, any kind of items available in the active CanoDraw project may be added. If you specify symbol attribute plots (by selecting *Symbols* option in the *Visualization Method* group), you cannot add the plotting of corresponding symbols.

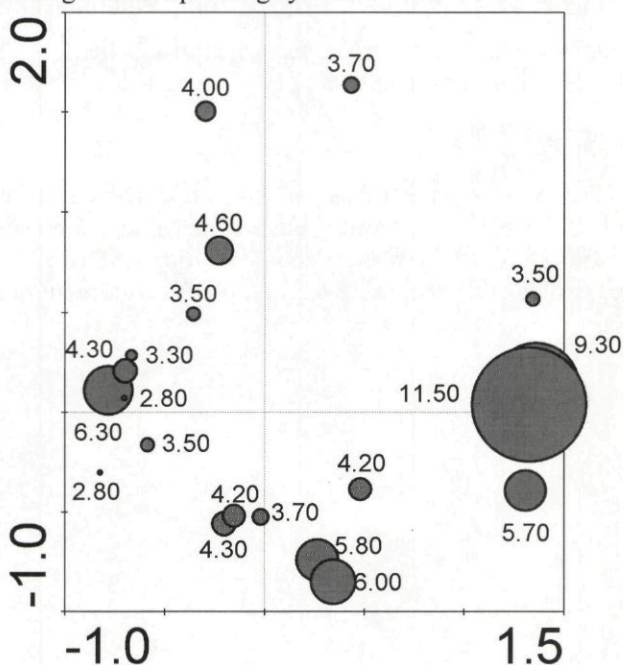


Figure 12-61 Symbol Attribute Plot diagram

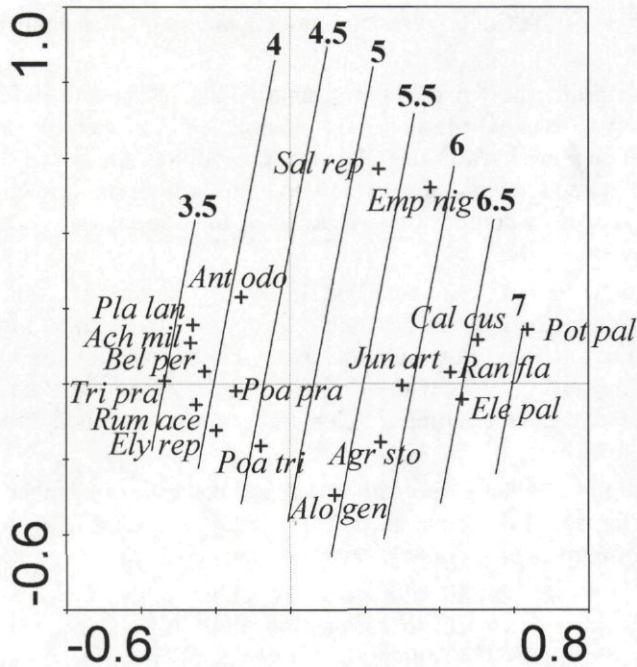


Figure 12-62 Contour (3-D) Attribute Plot diagram

Contour-based attribute plots always have their contours labelled.

12.5.5.2 Results Attribute Plot

The dialog shown by this command lists the scores and statistics provided by Canoco for samples, species, and (if present in the project) environmental and supplementary variables. All the variables listed were retrieved from the Canoco solution (.SOL) file and mostly have the names used there and explained in the Table 6.22 on p. 133 with more details in the following parts of the section 6.3).

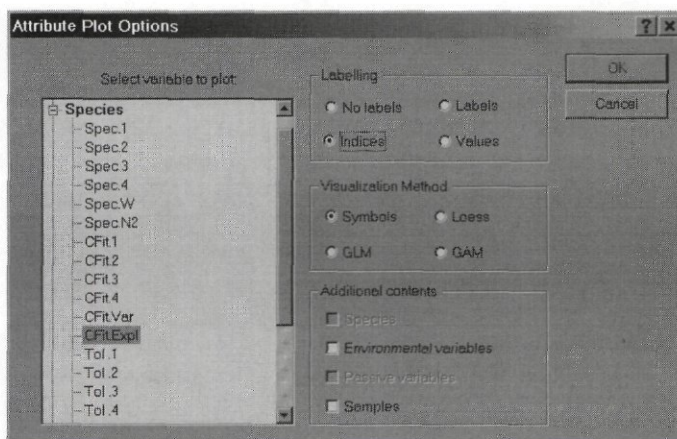


Figure 12-63 Results Attribute Plot dialog

The meaning of the options on the right side of this dialog box (see Figure 12-63) is already explained in the preceding section 12.5.5.1.

12.5.5.3 XY(Z) Plot

The dialog shown by this command (Figure 12-64) provides the most flexible tool for creating diagrams in CanoDraw for Windows. You can use it to create plots encoding information on the patterns involving up to three variables (like the XYZ diagram in Figure 12-65) or to create a XY scatter diagrams with fitted regression models (Figure 12-66).

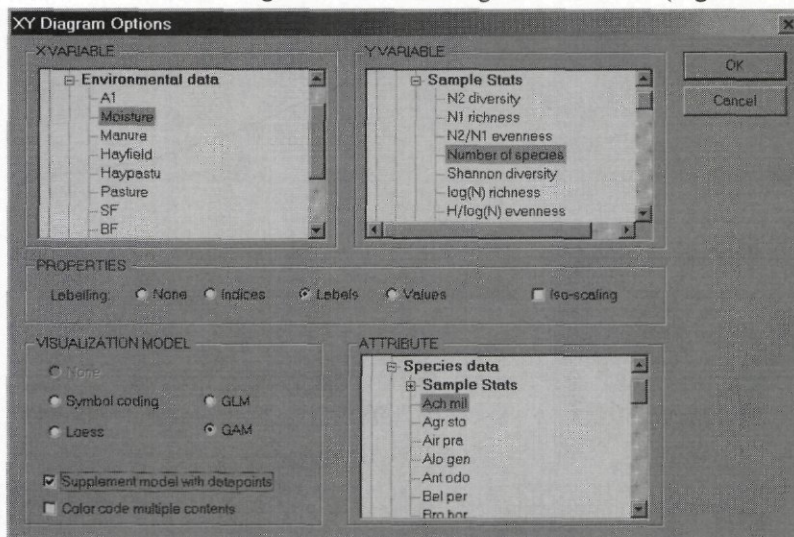


Figure 12-64 XY Diagram Options dialog box

The type of the graph, which will be created, depends on the selections made in the lists **X VARIABLE**, **Y VARIABLE**, and **ATTRIBUTE** and on the options chosen in the **VISUALIZATION MODEL** area. In any case, there must be one variable selected in the **X VARIABLE** listbox and its type determines which variables are shown in the remaining two boxes. Therefore, it is best to start your selection there. The possible choices for the other two list boxes together with options selected in the dialog box are summarised in the following table:

Y VARIABLE	ATTRIBUTE	VISUALIZATION MODEL	Resulting diagram
single variable	no variable ⁽¹⁾	None	XY scatter plot
single variable	no variable ⁽¹⁾	Symbol coding	<i>NOT ALLOWED</i>
single variable	no variable ⁽¹⁾	GLM, GAM, Loess	fitted model $Y \sim f(X)$, like Figure 12-66 ⁽²⁾
multiple variables	<i>NOT ALLOWED</i>	None	superposed XY scatters with shared X values ⁽³⁾
multiple variables	<i>NOT ALLOWED</i>	Symbol coding	<i>NOT ALLOWED</i>
multiple variables	<i>NOT ALLOWED</i>	GLM, GAM, Loess	multiple fitted models $Y_j \sim f_j(X)$ ⁽²⁾ ⁽³⁾
single variable	single variable	None	<i>NOT ALLOWED</i>
single variable	single variable	Symbol coding	symbol attribute plot like Figure 12-65
single variable	single variable	GLM, GAM, Loess	contour attribute plot for model $Z \sim f(X, Y)$

Comments: (1) *no variable* means that either no item is selected in the *ATTRIBUTE* list or that the first item named *No attribute used* is selected; (2) the original data points used for fitting the regression model can be displayed by checking the option *Supplement model with datapoints*; (3) when multiple response variables are plotted either directly or in the form of fitted regression models, the identity of regression curves and/or of the data points can be differentiated using color coding: to do so, check the *Color code multiple contents* option. If this option is applied to plotted symbols, it is active only if the items serving as data points are not classified.

The four choices for the *Labelling* option are applied to symbols used in the scatter diagrams. The multiple response curves of regression models based on XY scatters are always labelled, unless the *Labelling* setting is *None*. The contour lines are always labelled.

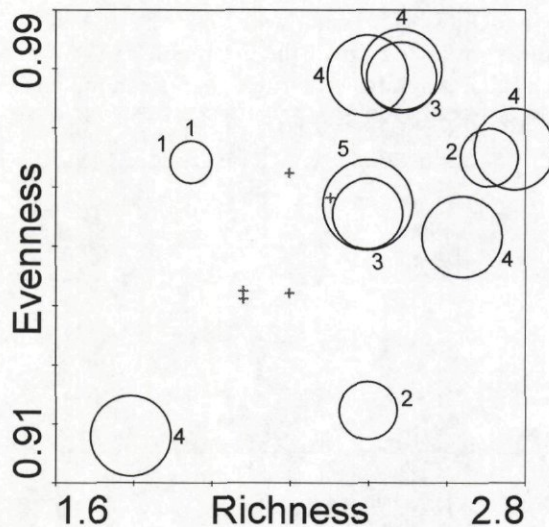


Figure 12-65 XYZ Diagram with Z values coded by symbol sizes

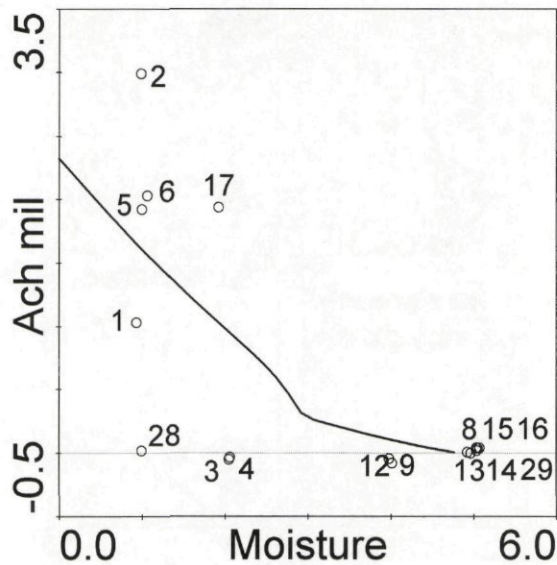


Figure 12-66 XY scatter diagram with fitted loess model

XY(Z) diagrams also support the option for maintaining identical scale for horizontal and vertical axis. This option can be activated with the *Iso-scaling* checkbox and results in the same physical distance on a printed graph corresponding to the same extent of the units of the variables plotted along the X and Y axes. Iso-scaling is used in the standard ordination diagrams and this option allows you to plot the ordination scores without distorting the reported relations between the displayed items. Note that the iso-scaling option is ignored if **banking to 45 degrees** is active (see section 12.3.1.1 for more details about banking).

12.5.5.4 Species response curves

This command provides a shortcut for fitting multiple regression models characterising the change of values of multiple species (response variables) along an ordination axis or with the values of an environmental variable.

When you select this command, the dialog shown in Figure 12-67 is displayed.

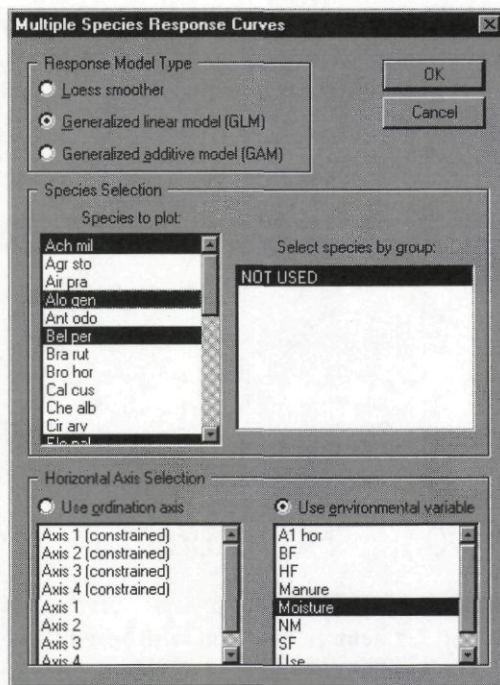


Figure 12-67 Dialog for fitting multiple species response curves

When using this dialog, you first specify the family of regression models to use (loess, generalized linear models, or generalized additive models) and then you select the species (*Species to plot*). The species can be selected from the list directly, using left mouse clicks, *Ctrl*-clicks (for multiple non-contiguous selection), or *Shift*-click (for multiple contiguous selection). Alternatively, the list on the right side of the species list shows the currently defined groups of species and if you select one of them, all the species which are members of that group are selected. The two lists at the bottom of this dialog provide you with two alternative selections for the horizontal axis (the predictor). You can use either the positions of samples on an ordination axis (if available, both constrained and unconstrained sample scores are offered) or the values of environmental variables.

After you close the dialog with the *OK* button, CanoDraw displays a dialog where you specify settings of the particular type of regression models. **These settings are then shared for all the response models being fitted at the same time.** If you, for example, ask for a second-order polynomial form of the predictor in GLM, this specification is then used for all the selected species. Note, however, that if you selected a stepwise selection of the regression model, this selection is performed independently for each response variable (species). If the option *Show summary of each fitted regression model*, accessible in the dialog invoked by the command *View / Diagram settings* (see section 12.3.1.1), is **on**, CanoDraw displays a summary of each fitted model (preceded by a summary of the stepwise model selection, if performed). Each summary dialog can be closed not only by *OK* button, but alternatively by a *Skip* button. In that case, the particular fitted model is not included in the resulting diagram. This is useful, for example, if the stepwise selection concluded for a species that no alternative model is better than the null model (i.e. the model stating no change of species values with the predictor).

CanoDraw does not allow the plotting of original data points here and it automatically sets the color-coding of multiple curves.

Note that if you change the CanoDraw options and then select re-creation of the graph created with the *Species response curves* command, CanoDraw offers the selection of regression

model separately for each of the previously selected species. This allows you to fine-tune the model settings for individual species.

12.5.6 Recreate graph **G**

This command re-builds the currently active graph, applying the currently active settings, as specified in the options available in the first two pages of the dialog invoked by the *View / Diagram Settings* command (see section 12.3.1), in the dialog displayed by the *View / Visual Attributes* command (see section 12.3.2), and in all the pages of the dialog invoked by the *Project / Project Settings* command (see section 12.4.1). The options selected in the specific dialog of the graph-creating command or in the following dialogs for the specification of regression model cannot be changed during graph re-creation (except for regression model settings for the diagram created by the *Species response curves* commands).

12.5.7 Range of axes **G**

This command allows you to set explicitly the range of values covered by the axes of the active graph or to return to the implicit settings for the axes range. When you select this command, CanoDraw displays a dialog similar to the one illustrated in Figure 12-68.

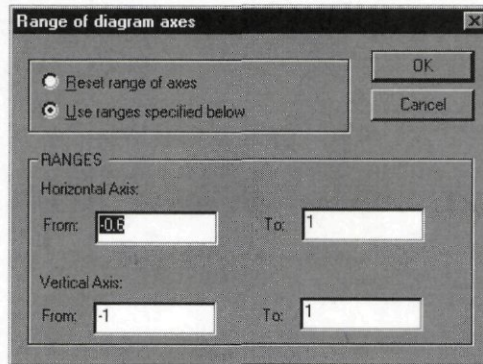


Figure 12-68 Range of diagram axes dialog

Initially, for a newly created diagram, the first option in the diagram top (*Reset range of axes*) is selected. Clicking the *OK* button will then have the same effect as executing the *Recreate graph* command. If you select the *Use ranges specified below* option instead, the edit fields in the *RANGES* group are enabled. They are initially set with the current range of the horizontal and vertical axis, but you can change them as needed. After you click *OK*, the current graph is re-created and the range of axes is fixed on the values you specified here, with eventually small adjustments. If you select this command again for a graph with an already changed range of axes, the dialog box appears in the state illustrated in Figure 12-68, with the current axis range settings being displayed. To return to the default axes range, as determined by CanoDraw, change back to the *Reset range of axes* option and click the *OK* button.

12.5.8 Lock legend **G**

This command closes the previously unlocked graph legend for editing and the legend starts to behave as a single entity when moved around the graph or when deleted (and undeleted). Additional information can be found in the next section.

12.5.9 Unlock legend **G**

CanoDraw maintains the legend area, if present in a graph, as a separate opaque entity. The only manipulation available for a legend is the change of its position or its deletion. The frame enclosing the legend can be selected and moved around the graph, and its contents moves together with it. If you select the frame and press *Delete* key or select the *Delete* command from the *Edit* menu, both the frame and the legend contents are deleted.

Nevertheless, this behaviour of the legend area does not allow you to modify its contents. To change size and other properties of labels and of other graphical objects in the legend area, you must first **unlock** the legend. After unlocking, the legend starts to behave as a loose group of graphical objects. Additionally, all the objects contained in the legend can be moved, independently of the others. This allows you to fine-tune the positioning of individual legend parts. After you have made the required changes of legend contents, you may again lock it, using the command described in the preceding section.

After you have unlocked the legend, you can select the individual objects in the legend area to change their properties. Note, however, that the frame enclosing the legend area is "on top" of the other items, so you can initially select and change only this frame. Therefore, you must first either move the frame outside its original position or disable it. You can disable the legend frame by locking it down. Note that this is not the same operation as locking the whole legend (see section 12.6.3 for a more detailed description of locking individual graph objects).

12.6 Object **G**

This menu appears in the CanoDraw menu bar only if the active window represents a graph. Many of the commands in this menu can be also found in the context sensitive pop-up menu, which is displayed when you click the graph outside of any selected graph objects, using the right mouse button.

12.6.1 Select Suchlike

This command can be executed if **exactly one** graph object is selected. Keyboard shortcut for this command is *Ctrl+H*. It selects in the active graph all the graph objects, which fulfil the following three conditions:

- * They are of the same type as the selected object (see Table 11-2 in section 11.5 for a list of graph object types).
- * They are not locked (see section 12.6.3 below).
- * They have identical settings for the selected attributes listed in Table 12-3.

Graph object type	Attributes
labels	draw color, background fill color and fill style, font name, size, and style (label orientation is ignored)
lines	draw color and line style
arrows	draw color and line style, fill color and fill style
polylines	draw color and line style
symbols	symbol type, draw color and outline style, fill color and fill style
pie-symbols	draw color and outline style
rectangles (bars)	draw color and line style, fill color and fill style

Table 12-3 Attributes for individual graph object types, checked for identity in *Select Suchlike* command

12.6.2 Select Similar

This command can be executed if **exactly one** graph object is selected. It selects in the active graph all the graph objects, which fulfil the following two conditions:

- * They are of the same type as the selected object (see Table 11-2 in section 11.5 for a list of graph object types).
- * They are not locked (see section 12.6.3 below).

This type of selection works on a quite broad scale: if you select one text label and execute this command, all labels in the graph are selected. Keyboard shortcut for this command is *Ctrl+I*.

12.6.3 Lock selected

This command temporarily disables access to currently selected objects. Keyboard shortcut for this command is *Ctrl+L*. The locked objects cannot be selected by any selection method (see section 13.1 for additional information about selecting graph objects) and, therefore, cannot be deleted or visually changed. Locking is a solution for the problem with overlapping graph objects during selections. If an object overlaps another one(s), these cannot be easily selected, unless the topmost object is first selected and then locked.

Locking does not unlock the objects that were locked earlier. The list of locked objects is extended with each new execution of the lock command. Because locking suppresses any access to the objects, you cannot unlock them selectively. Instead, you must unlock all the locked objects at the same time. The operations of locking selected objects and unlocking all at once are modelled after a similar functionality in Adobe Illustrator® software.

Do not confuse locking and unlocking one or more graphical objects with locking and unlocking the graph legend (see section 12.5.8). Note that when the legend is locked, all the graph objects contained in it (except the enclosing frame) are also individually locked.

12.6.4 Unlock all

This command unlocks all the previously locked graph objects (see the preceding section for a more detailed explanation). Keyboard shortcut for this command is *Ctrl+U*. The contents

of a locked graph legend provide an exception to this: they are not unlocked with this command. Instead, you must use the *Unlock legend* command (see section 12.5.9).

12.6.5 Graph tool

The four commands in this submenu allow you to select the type of tool, which can be used to work with CanoDraw graph windows. The first of them (*Selection*) is the default choice selected at the start of the CanoDraw program. You are advised to switch to the other graph tools only for the time period they are needed, to add new contents to your graph, and then to switch back to the default *selection* tool. The current graph tool settings are maintained independently for each of the CanoDraw graphs opened in a session.

12.6.5.1 Selection

The mouse pointer has the standard shape of arrow and can be used for various types of selection of graph objects (see section 13.1 for additional details about selection).

12.6.5.2 Arrow

This tool is used to supplement your graphs with additional arrow objects. These can be useful, for example, in situations you must move a label or a group of labels far from the object(s) they label. Then an arrow object might visually re-establish connection between these two (sets of) items.

After you select the arrow tool, move the mouse pointer (which has a changed look) to the point on your graph where the arrow should start. Click the left mouse button once and move the pointer to the place where the arrow should terminate (where the tip of the arrowhead should be placed). While you move the mouse pointer, a preview of the arrow placement is drawn on your graph. When you are satisfied with the arrowhead tip position, click the left mouse button the second time. Note that you cannot cancel the arrow definition once you started with it. If you want to abandon the creation of an arrow, place its tip anywhere and then remove the arrow using the *Edit / Undo Add Arrow* command. Note that the arrow position can be adjusted after it was created by selecting the arrow object, and using either the direction keys on the keyboard or dragging the object to reposition it.

12.6.5.3 Label

This tool is used to add new text items to a CanoDraw graph. You can use it, for example, to add a title to the graph or to provide any supplementary information. You create a new label by clicking with the label tool (marked by a cursor in form of letter A) at the desired label position. CanoDraw displays a dialog where you can enter the text to be placed at the selected location (the default text is *LABEL*). The label is centred at the specified point. You can modify the typeface and font size later.

12.6.5.4 Line

This tool can be used to add straight lines to CanoDraw graphs. The instructions in the section describing use of the *Arrow* tool (section 12.6.5.2) can also be used for this tool. The only difference is that both ends of the lines are identical.

12.7 Window

This menu provides commands for selecting windows, arranging their position in the CanoDraw workspace, and closing particular group of windows.

12.7.1 Cascade

This command disperses the currently opened windows regularly in the application workspace so that they **do overlap**, but it is possible to activate as many of the windows as possible, without moving the other ones.

12.7.2 Tile

This command regularly disperses the currently opened windows in the application workspace, so that they **do not overlap**.

12.7.3 Arrange Icons

This command arranges the position of icons for minimised windows, at the bottom of the CanoDraw application workspace.

12.7.4 Close all

Closes all the currently opened windows, asking about saving them, if their contents changed.

12.7.5 Close graphs of active project

Closes all the graphs, which were created from the currently active CanoDraw project. A project is active if the currently active window relates either directly to this project or to a graph created in this project. If there are any unsaved graphs, CanoDraw first asks whether you want to consider saving their current state. If you answer *No*, the graph windows are closed and any newly created graphs or graphs with modified contents are discarded. If you answer *Yes*, you are asked about saving each graph with non-saved modifications (including newly created graphs, which were not saved yet).

12.7.6 Open graph project **G**

If you select this command, CanoDraw attempts to find the file with the CanoDraw project, which was used for creating the currently active graph. If you changed the project file (*.cdw* file) location, CanoDraw displays a dialog box allowing you to locate the project in its current placement.

12.7.7 List of windows

CanoDraw lists the individual windows, currently opened in the application workspace, at the bottom of the *Window* submenu. If there are too many windows open, CanoDraw shows just part of the windows' list and concludes the list with the *More Windows ...* command. If you select it, you are presented with a dialog, where all the currently opened windows are listed, and you can select one of them and then active it using the *OK* button.

12.8 Help

This menu provides help information about using the CanoDraw for Windows application.

12.8.1 Help Topics

This command opens the Microsoft Help application and provides a list of topics available in the CanoDraw help.

12.8.2 About CanoDraw

This command displays the box with copyright information relevant for the CanoDraw for Windows software, as well as additional information about the actual software version and license.

12.8.3 Tip of the Day

Displays the Tip of the Day dialog box, providing useful hints for using the CanoDraw for Windows program (see Figure 12-69 for an example).

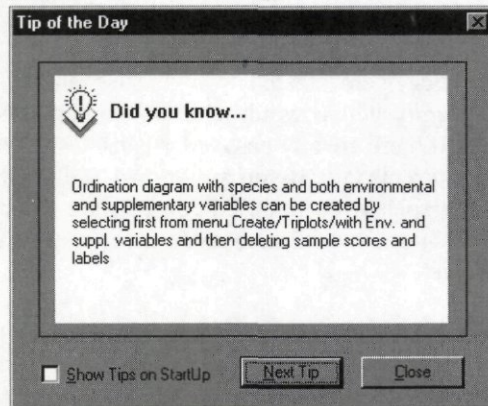


Figure 12-69 Tip of the Day dialog

The controls at the bottom of this dialog box allow you to set or un-set the display of this dialog upon the application startup or to browse through the available tips (using the *Next Tip* button).

13. Working with CanoDraw

The various sections in this chapter describe different aspects of working with the CanoDraw graphs.

13.1 Selecting graph objects

The selection of an appropriate set of objects is the key to efficient modification of the look and feel of CanoDraw graphs. There are several ways how to select desired graph objects and they are described in the following subsections. If you open the *Properties* floating window (either using the *Properties Sheet* command in the *View* menu or using the *F5* keyboard shortcut), its content reflects the selection in the active graph window.

The standard graph window is the place where you select graph objects most of the time. Selected objects are marked either by a red rectangle enclosing the symbols, labels, or pie-symbols or by two red crosses at the opposite ends of a line segments or arrow. For polylines, which are composed of multiple line segments, each node where the individual segments start and end is marked by a red cross on a selected polyline.

Alternatively, some of the selection techniques described below can be also used in the *Graph Contents* window, which displays the graph contents in a hierarchical manner. The selections made here and in the normal graph window can be set and changed independently.

13.1.1 Manual selection with mouse

You can select a single graph object (see section 11.5 for an explanation of this term) by clicking it with the left mouse button. New graph objects can be added to the existing selection by clicking on them with the left mouse button, while keeping the *Ctrl* or *Shift* key pressed at the same time. If you click over the graph background (outside of any graph object), all currently selected graph objects are deselected.

The manual selection is also available in the *Graph Contents* window. There you also select a single item by clicking on it with the left mouse button. A non-contiguous sequence of items can be selected by holding the *Ctrl* key, pressed while clicking over the additional items. A contiguous sequence of items can be selected by selecting the item on one end of the selection range and then clicking the item on the opposite end, while keeping the *Shift* key pressed.

13.1.2 Rubber-band selection

To select all the graph objects in a rectangular area, you can use this type of selection. Position the mouse pointer at one of the corners of the desired selection area, which must be selected so that no graph objects underlay it. Press the left mouse button and keep it pressed. While you move the mouse pointer (keeping the left button pressed) to the opposite corner, an outline of the current selection area is drawn. Once you release the left mouse button, the selection area is defined. All the objects enclosed by the "rubber-band" rectangle are selected.

This selection method cannot be used to extend an existing selection. When you finish a rubber-band selection procedure, any previously selected objects are de-selected and only the objects within the rectangle are selected.

13.1.3 Selection by example: similar and "suchlike" items

In this type of selection, you start by selecting **one** "example" object - the following selection is then based on the properties of this object. If you use the *Select Suchlike* method, only the graph objects of the same type and with very similar visual attributes are selected (see section 12.6.1 for a complete list of attributes used in the *Select Suchlike* command). If you use the *Select Similar* methods, a wider selection is performed: all the graph objects of the same type are selected (e.g. all the symbols, all the labels, or all the line segments). See section 12.6.2 for additional description.

These two commands are also available from the *Graph Contents* view, from the context sensitive menu which appears when you click a selected item using the right mouse button.

13.1.4 Selecting whole class of items

This type of selection is available only from the *Graph Contents* view. You select a logically connected group of items (e.g. all species symbols or all axes labels) by selecting one of the non-terminal items (with bold typeface).

13.1.5 Selecting graph object by its label and vice versa

Graphs produced by CanoDraw for Windows usually present one or more sets of items (samples, species, explanatory variables), with each item in a set represented by a pair of graph objects: a symbol or arrow on the one hand and a label showing the item identity on the other hand. If you plot extensive sets of such items, the visual connection between the symbol (or arrow) and its label is sometimes lost, due to an over-crowded graph area. Also, you may sometimes break this visual connection by moving the label too far from its original position. In other cases, you can see multiple labels attached to just one symbol. There are, in fact, as many symbols as there are labels, but you can see (and select) only the symbol which happened to be on the top and hides the remaining ones.

The two specific selection commands come to your rescue in such situation. Both commands are optionally available from the popup menu, which appears when you click over a selected graph object, using the right mouse button. The command named *Select object's label* is included in this menu if you have only one object selected and that object has an attached label. When you execute the command, the object is deselected and its label is located and selected instead. The other command is named *Select labelled object* and it appears in the menu when your current selection consists of a single label which is attached to some particular object (not all labels have this connection).

13.1.6 Locking and unlocking objects

All the above described selection techniques select all the graph objects which fulfil the rule used for the particular type of selection, with the exception of objects which are **locked**. Object locking is an operation, which adds the currently selected objects to a group of already locked objects (see section 12.6.3). The locked objects are completely excluded from the selection and, therefore, from any manipulation, including a change of their visual attributes, and also a change of their position for the moveable ones (typically labels, but also the legend frame or individual items of an unlocked legend). The currently locked objects can be unlocked only all at the same time, using the *Unlock all* operation (see section 12.6.4).

13.2 Finding particular object

If you want to find a symbol or label for a particular sample or species (for example to highlight its position in the graph), you can do so most easily using the tree-like view of your graph. For example, if you create an ordination diagram representing the scatter of samples in the ordination plane spanned by the first two axes, you can locate the symbol corresponding to a particular sample by selecting the *View / Tree view* menu command. A new view of your graph appears (titled *Graph Contents <GraphName>*) and there you must expand the *Samples scores* section and then the *Symbols* sub-section. A list of available sample symbols appears, stating the text of their associated labels. You can select the particular item and modify the symbol properties using the *Properties* sheet (which can be activated with the *F5* key, when not visible). Note that this method does not work if the symbols are not labelled. Nevertheless, you can explore identity of individual symbols even in such case, by clicking the selected item with the right mouse button. The displayed context menu shows the index of the selected object in the square brackets of the menu command *Summary of xxx [NN]*.

13.3 Modifying graph contents

The contents of a graph is the result of decisions made at three different levels of resolution:

1. First, you decide about the graph contents by selecting the type of graph you want to plot. You do so by selecting one of the commands available from the *Create* submenu. A rough classification of CanoDraw graphs is provided in section 11.4 and more detailed description of the individual commands for creating graphs is available in section 12.5.
2. The selected graph type is combined with the additional settings, which are either project-specific ones or used for all the projects (see section 11.8). As an example, if you execute the menu command for creating a scatter of sample points (*Create / Scatter Plots / Samples*), the contents and appearance of the resulting diagram depend on many **project-specific options**:
 - * which ordination axes are currently specified for plotting and whether flipping of scores along the ordination axes is in place (section 12.4.1.1)
 - * whether the option for plotting sample scores derived from species scores even for constrained analyses is active (section 12.4.1.1)
 - * whether sample symbols should be replaced by pie-symbols (section 12.4.1.1, but beside the *Use Pies ...* option checked there for samples, there also must exist an **active** classification of species)
 - * if the *Draw Envelopes around Classes* option is checked for *Samples* (section 12.4.1.1), then either all the sample symbols are enclosed by a convex polygon (if the samples are not classified) or separate envelopes are placed around the samples belonging to individual sample classes
 - * sample symbols are labelled by their names (with up to eight characters), by their indices (see section 11.2), or they are not labelled, depending on the *Labelling of scores* section in the *Appearance* page (see section 12.4.1.2)
 - * the scatter of sample symbols can be optionally supplemented with a legend, which is particularly handy if the samples have an active classification and / or the envelopes or series are plotted within the graph; legend presence and layout are specified in the *Appearance* page of the *Project Settings* dialog (see section 12.4.1.2)
 - * the set of plotted samples is restricted by the rules from the *Inclusion Rules* page of the project options dialog (see section 12.4.1.3)

- * if the samples are classified (see section 12.4.3), the symbols of samples from different classes are of different type and / or of different color, depending on the visual attribute settings which are application-wide and discussed below
- * if there is any active series collection of samples (see section 12.4.5), it is also displayed in the diagram
- * presence of individual samples in the diagram is further governed (overriding the inclusion rules, mentioned above) by the choices you made in the dialogs invoked by the *Suppress* and *Enforce* commands (see sections 12.4.6 and 12.4.7).

The diagram appearance is further influenced by the following project-wide options:

- * if you plot the results from a Canoco analysis, where scaling of ordination scores was focused on the inter-species distances (in weighted averaging ordination methods) or on inter-species correlations (in linear ordination methods), you have the option to ask CanoDraw to rescale the sample scores for their optimal interpretation, when they are plotted alone (*Rescale sample or species scores to optimality*, see section 12.3.1.1). This rescaling can substantially influence the visual spread of sample points, if the corresponding eigenvalues of the two plotted ordination axes differ much in their extent
 - * the *Axes Tickmarks* and *Axes Layout* options (described in section 12.3.1.2) influence the actual appearance of the axes plotted around the diagram area or even the actual area (if you change the inward tickmarks into a reference grid).
 - * the appearance of individual graph objects within the diagram is directed by the settings accessible from the dialog invoked by the *Visual Attributes* command (section 12.3.2). You can change the thickness of various types of axes lines, font properties of axis labels, type, color, and size of symbols used for samples (whether classified or not), and properties of lines used to draw sample envelopes or representing the sample series.
3. The above two sources of diagram contents (the selection of graph type and the current settings available from various places in the *Project* and *View* submenus) lead to the creation of a new graph. Now you can directly manipulate the graph objects and fine-tune those initial settings.

Probably the most important type of graph modification is the readjustment of item labels, to minimise their overlap and therefore to increase graph readability. You can change the position of any label present in the CanoDraw graph, unless it is locked (see section 13.1.6 for additional information about locking graph objects). You can move a label if it is the only selected object in the graph. Press the left mouse button while the mouse pointer is anywhere within the red rectangle enclosing the selected label, keep the button pressed, and move the mouse pointer. The shifting label position is previewed using a rectangle drawn with a contrasting color. When you release the button, the label is repositioned to the actual mouse pointer position. Alternatively, you can move a singly selected label object using the arrow keys on your keyboard. On each key press, the label position changes in the implied direction by 0.001 in *virtual coordinate units* (see section 11.6). If you keep the *Shift* key pressed at the same time, the change is by 0.050 units. Dragging the selected label using the left mouse button requires two steps: first the label must be selected and then the dragging proceeds. To speed-up the process, CanoDraw also supports dragging of labels without prior selection. For this to work, there cannot be any graph object selected. You position the mouse pointer over the label you want to move, press the left mouse button, and start dragging. You do not see the enclosing rectangle outline for a while, because CanoDraw "waits" with the decision whether you really want to reposition the label for some time. This prevents erroneous shifts of labels while clicking over them. Obviously, this quicker form of repositioning labels is not appropriate for changing label position by a very short distance. It

is also considerably less precise concerning which label will actually move if you start over a group of partially overlapping labels.

You can also extend the graph contents by adding extra arrows, labels, or lines. Section 12.6.5 describes in detail how to add the individual types of graph objects. You can also delete any items from the graph. Like all the changes performed after CanoDraw has created a graph, deletion of graph objects can be undone (see section 12.2.1).

Finally, you can change the visual attributes for any group of graph items, represented by the current selection in the graph window (or in the *Graph Contents* window). Additional information about the visual attributes available for change can be found in section 12.3.4.

13.4 Creating graph legend

A legend in a diagram enables the viewer to recognise various facets of the contents that the graph has. If you plot just a scatter of points in a diagram, there is no pressing need to have a legend alongside it. If you display an ordination diagram containing the arrows for species and symbols for samples, the decision about legend usefulness is more difficult. You can report the fact that the arrows are used for species and the points for samples in the figure caption, but having an example of a short arrow and an example of the symbol used to plot sample positions in the legend, together with short labels saying "Species" and "Samples" probably allows your reader to comprehend the meaning of the graph more quickly. And, finally, if you have a graph where samples belonging to different classes are differentiated by symbol type or response curves for multiple species are to be distinguished by their color, the legend utility cannot be beaten.

CanoDraw produces a legend automatically and there are few opportunities to modify this process. Primarily, you can switch the legend creation on and off and determine the position of the whole legend area, layout of legend sections, and layout of items within the sections (see section 12.4.1.2). The **legend sections** collect **legend items** with qualitatively comparable meaning. For example, if the sample membership in pre-defined classes is coded by the symbol type, a legend section named *SAMPLES* has as many items as there are displayed sample classes, each showing one type of symbol alongside with the corresponding class name. If you add the option to plot envelopes enclosing samples of the same class (see section 12.4.1.1), there is still one legend section named *SAMPLES*, but the class names are preceded not only by symbol examples but also by a short segment demonstrating the color and style of the line for the envelope of that particular class. If you then define one or few series of samples and request plotting that series collection (see section 12.4.5), a new legend section is created and named *SERIES OF SAMPLES*, demonstrating the style and color of lines used to display the individual ordered series of samples. These two sections cannot be merged, because the number of sample classes is generally different from the number of series.

The recommended way of using the facilities for legend creation in CanoDraw is summarised in the following steps:

- * Let CanoDraw create the legend with the graph by enabling legend creation (section 12.4.1.2)
- * Adjust the legend position and the layout of legend sections and items within the sections as needed (section 12.4.1.2) and eventually change the font used for the legend (use the command *Visual Attributes* in the *View* submenu, selecting item *Text* from the *Legend* folder). Note that this font size is also used to determine the size of the example graph objects (if you increase font size, you get larger sample symbols, longer lines, larger patches demonstrating fill style in pie-slices, etc).

- * Recreate the graph with the new settings using the *Recreate graph* command (see section 12.5.6)
- * Repeat the preceding two steps as many times as needed
- * After you are satisfied with the general layout, you can adjust the legend position by dragging it around the diagram area with the mouse
- * You can then unlock the legend (see section 12.5.9), lock its frame, and readjust the position of sample graph objects and labels, or change the legend text
- * Save the finalised graph

13.5 Exploring graph contents

In this section, the graph exploration in CanoDraw is described. Note that by graph exploration, I do not mean understanding what has been actually plotted in the graph. The awareness of what you are doing belongs to the assumptions I have about you, the CanoDraw user. Graph exploration means finding what does the graph tell you about the interesting patterns and relations within the analysed data. This involves both an understanding of how to interpret the contents of the ordination graphs and a deeper insight into the ideas suggested by the present graph contents.

The *ordination diagrams* (which are one type of the graphs produced by CanoDraw, see section 11.4) help you to summarise patterns in your data and to find interesting relations among various variables. The ability to interpret the ordination diagrams in this way results mostly from following one of two simple interpretation rules (named **biplot rule** and **centroid principle** in the Canoco documentation and elsewhere). The consequences of applying such rules to sample, species, and environmental variable scores in the ordination diagrams are listed in detail in two papers (Ter Braak 1994 and Ter Braak & Verdonschot 1995). CanoDraw provides short summaries of the most important rules that can be applied to an ordination diagram. To see the summary, you should click the graph area (outside of any selected graph object) with the right mouse button and then select from the popup menu the *Describe contents* command.

The dialog with title *Graph Description* contains text summarising the graph contents and providing suggestions how to interpret the particular type of items contained in the diagram. The suggestions are usually supplemented with a simple graphical scheme. For example, if you execute the *Describe contents* command over a scatter diagram with sample symbols, you obtain the following description:

Diagram Interpretation

Ordination diagram [Axis 1 x Axis 2] with samples

This diagram contains only one type of scores which can be interpreted as follows:

- * Sample points: the distance between the symbols in the diagram approximates the dissimilarity of their species composition, measured by their Chi-square distance.

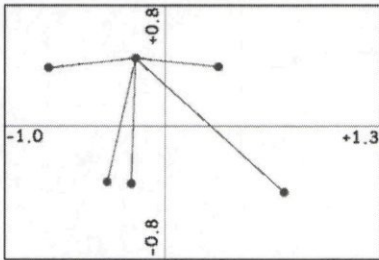


Figure 13-1 Example of Graph Description window contents for a diagram with samples

If, for example, the sample symbols are replaced with pie-symbols, the graph description changes accordingly (the inserted illustration was omitted, being identical with that in Figure 13-1):

Diagram Interpretation

Ordination diagram [Axis 1 x Axis 2] with samples

This diagram contains only one type of scores which can be interpreted as follows:

- * Sample pies: the distance between the symbols in the diagram approximates the dissimilarity of their species composition, measured by their Chi-square distance.
- * Sample symbols are replaced by pie symbols. Segmentation of those symbols into slices is based on the currently active classification of species. Relative size of particular pie-slice corresponds to relative importance (measured either by presence or by quantity) of species from particular class in the corresponding sample..

Figure 13-2 Example of Graph Description window contents for a diagram with sample pie-symbols. The illustration provided within the window was removed.

If there is more than one type of items present in the graph, CanoDraw also provides suggestions how to interpret the relations between the various types of items. Figure 13-3 illustrates one paragraph taken from a description of a biplot diagram with species and environmental variables, focusing on the relation between the species and the nominal environmental variables.

Symbols of individual classes can be projected perpendicularly onto the line overlaying the arrow of the particular species. These projections can be used to approximate the average abundance of that species in individual classes of samples. Projection points are in the order of predicted increase of abundance of the particular species across the classes. Predicted increase occurs in the direction indicated by the arrow. In analyses where centering by species was performed (most analyses), the classes projecting onto the coordinate origin are predicted to have an average value of that species near to the global average value of that species in data.

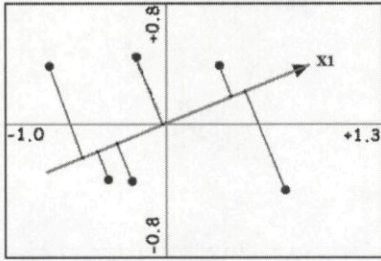


Figure 13-3 Part of the Graph Description window contents for a biplot diagram, suggesting a joint interpretation of species arrows and symbols of nominal environmental variables.

Contents of the Graph Description window can be copied (using the *Copy* button at the window bottom) to the Windows Clipboard and pasted into a word-processing application in RTF format.

Besides suggesting interpretation of the ordination plots, CanoDraw can also provide you with additional data about the samples or about the variables (species or explanatory variables) seen in the plot. To obtain such information, you must first open one or both types of floating windows. CanoDraw has separate windows for providing a summary of variables and for summarising samples. To open the window summarising the variables, select one species or explanatory variable in the ordination diagram and click the right mouse button, while the mouse pointer is over the selected symbol or arrow. From the popup menu select the command *Summary of <variable-type> <variable-name>* (the actual text depends on the variable type and name, for example *Summary of species 'AchMil'*). The *Variable Summary* window (illustrated in Figure 13-4) appears. It will remain open until you click its *Close* button. As you move the mouse pointer through the diagram, each time the cursor goes over a new variable symbol (or arrow), the contents of the window is updated to provide summary information for this variable.

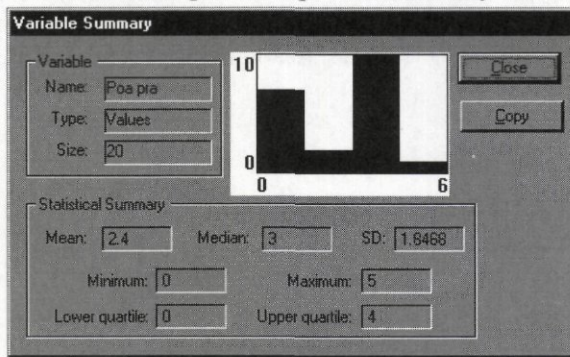


Figure 13-4 Variable Summary floating window

The summary in the *Variable Summary* window includes the name of the variable and the number of observations available for it, and the values of mean, median, standard deviation, minimum, maximum, and upper and lower quartiles. Additionally, a simple frequency histogram

shows the distribution of the variable' values. You can copy the values and related sample indices of the currently summarised variable to the Windows Clipboard, using the *Copy* button.

To see the *Summary of sample* window, you proceed similarly, but you should invoke the pop-up menu over a selected sample and the command to be executed is named *Summary of sample* <sample-label>. The floating window summarising the sample is shown in Figure 13-5.

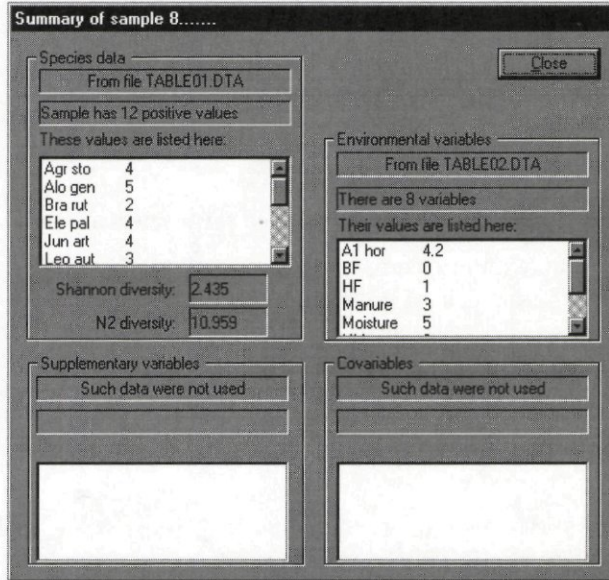


Figure 13-5 Sample summary floating window

This window lists the values available for a sample in the individual Canoco source files. For the species (primary) data, CanoDraw shows only the positive values (the assumed presences of species). CanoDraw displays up to the 500 first rows in any of the data-files. In addition, the estimates of Shannon diversity measure, as well as the N₂ diversity measure (see section 12.4.9) are shown for primary data. In some circumstances, these statistics have little meaning (e.g. if the primary data represent chemical properties of water samples), but they are useful for the typical species composition data.

In addition to providing a continuously updated summary of the individual variables and samples, CanoDraw also provides a quick access to the creation of diagrams summarising, for a selected variable, the distribution of its values through the ordination space. If you click with the right mouse button over a selected variable, a pop-up menu similar to the one shown in Figure 13-6 is displayed.

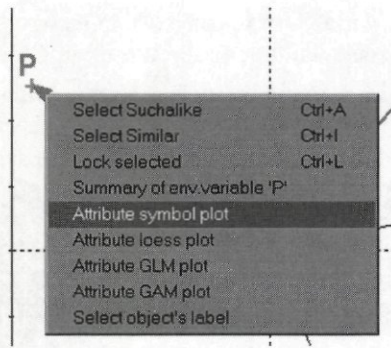


Figure 13-6 Popup menu displayed for a selected variable

The commands starting with the word *Attribute* provide an easy access to creating attribute plots, using the currently active pair of ordination axis (see section 12.4.1.1) as the horizontal and vertical axes and the selected variable as the attribute. The first command (*Attribute symbol plot*) creates the symbol-based attribute plot, where the attribute value determines the size of sample symbols, while the other three commands create contour-based attribute plots, presenting the fitted regression model of the particular type (loess, GLM, GAM). Additional information about specifying regression models is provided in the following section.

13.6 Fitting regression models

Regression models can be added to various types of attribute plots created by CanoDraw. In plots where an attribute value is compared with the values of two other variables, a regression model with two predictors can be fitted, in the other plots (XY diagrams), a regression model with one predictor is fitted.

CanoDraw provides three general families of regression models and you can freely choose among them. Their implementation in CanoDraw and suggested use are summarised in the following sections.

Note that facilities available in CanoDraw for fitting regression models are limited, because their anticipated use is restricted to summarising patterns in diagrams. For any extensive modelling exercise, you are advised to work with general-purpose statistical packages (such as the R, S-Plus, or SAS software).

Implementation of algorithms fitting GLM, GAM, or Loess models was validated by comparing the results with the software S-Plus for Windows 4.5 and S-Plus 2000. Note, however, that the exact values of estimated coefficients, amount of explained variance, etc., may differ to the order of about $1.0e-4$ to $1.0e-5$. This is inevitable, giving the different implementation of the model fitting code.

13.6.1 Generalized Linear Models (GLM)

The GLMs represent a straightforward extension of the classical linear models (McCullagh & Nelder 1989) and their specification in CanoDraw is described in section 12.3.1.3. CanoDraw allows you to specify the expected distributional properties of the response variables but only the **canonical** link functions are available for each type of supported distribution (except the Gamma distribution where the log link function is supported beside the canonical inverse link

function). GLMs represent the most rigid modelling approach among the three regression families offered in CanoDraw. The relation between the response variable and the predictor(s) is described by a few parameters (regression coefficients). With GLMs, you can fit simple, hypothesised statistical models of change of species abundances along the resource gradients (the "species response models"). This includes both the model of linear change of abundance along the gradient, as well as the model of symmetric unimodal response, popularised in ecology in connection with CA and CCA ordination methods. CanoDraw provides a special support for estimating the parameters of fitted unimodal response curves (see below).

When fitting regression models, CanoDraw uses the model specifications representing application-wide defaults. These defaults can be modified in the last three pages of the *Diagram Settings* dialog. Generalized linear models are on the *GLM Options* page (see section 12.3.1.3). If the option *Offer approval of regression model settings ...* in the *Properties 1* page of the same dialog is checked, CanoDraw also presents the model options immediately before a regression model is fitted, so you can adjust the model specification for the particular response variable and predictor(s). Fitted GLM is summarised with a dialog illustrated in Figure 13-7 below.

Fitted Generalized Linear Model

Response variable:

Predictor(s):

Distribution: Link function:

Null model deviance: with residual DFs

Fitted model deviance: with residual DFs

Model significance: F = P = AIC =

Unimodal response curve:

Optimum: S.E.: Conf. interval:

Tolerance: S.E.: Max. value:

Regression coefficients:

Model Term	B	s.e.	T
(Intercept)	1.09234	0.216751	5.03963
Samp.1	-0.78195	0.298081	-2.62328
(Samp.1)^2	-0.62254	0.314117	-1.98187

Figure 13-7 Fitted GLM summary dialog

The dialog shows the names of the response variable and of the predictor(s), the selected type of distribution for the response, and the link function. The total variance in the values of the response variable is displayed in the *Null model deviance* field, together with the corresponding number of degrees of freedom. The fitted model residual variability is shown in the next line, together with the residual degrees of freedom. The next line summarises the fitted model quality using a deviance-based test (with F-ratio statistics, see McCullagh & Nelder 1989) as well as using the AIC statistics (see Hastie & Tibshirani 1990).

The following area labelled *Unimodal response curve* is used only for the specific situation, where a second-order polynomial model with a single predictor variable is fitted and where an appropriate type of link function is chosen (*log* or *logit* link function). In such cases, the dialog shows estimates of the unimodal response curve optimum and curve width (tolerance) on the left side, followed by the estimated standard errors of the estimated optimum and tolerance.

On the right side, the 95% confidence interval for the optimum (i.e. the range of values in which the true value of species optimum lays with a probability of 0.95) is displayed, if it can be estimated. Finally, the predicted abundance (or probability of occurrence) of the species (response variable) is given in the lower right corner of this area. Additional information about calculating various parameters of unimodal response curves from fitted second-order polynomials can be found in Ter Braak & Looman (1986).

The lowest part of this dialog displays the individual estimated regression coefficients, together with the standard errors of such estimates, and an approximate T statistics. The T statistics can, in theory, be used to test hypotheses about a regression coefficient being equal to zero, but the test is very approximate for GLMs.

Note that the standard error estimates for regression coefficients, as well as the standard error estimates for Optimum and Tolerance parameters and the estimates for Optimum confidence interval, do not use the estimated scale parameter for Poisson and binomial distribution families. Rather, the value of scale parameter is assumed equal to 1 in this case.

The *Copy* button can be used to place the model summary in a text format on the Windows Clipboard. You can use the *Skip* button to ask CanoDraw not to plot the response curve (or response surface, with two predictors) corresponding to this model. Note that if there is no additional plot contents, the diagram creation is cancelled.

In the case you selected that your model should be based on a stepwise selection (and you have enabled the option for displaying the fitted model results – see section 12.3.1.1), CanoDraw displays a report about the performed model selection. In the case you asked for model selection using the analysis-of-deviance based test, the report looks like the one illustrated in Figure 13-8.

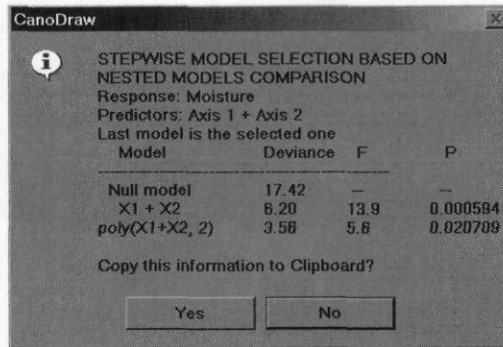


Figure 13-8 Report on stepwise selection of GLM using deviance tests

Note that only the models which were tried and found significantly "better" than the preceding more simple model are displayed. Therefore, if we specified that CanoDraw should check up to quadratic models at most, but the interaction of predictors should be also considered (see section 12.3.1.3), the report in Figure 13-8 means that the model $poly(X1 * X2, 2)$ (i.e. including the interaction between X1 and X2) was tested against the last displayed one ($poly(X1 + X2, 2)$), but the resulting drop of residual deviance was not found to be sufficiently large. See section 12.3.1.3 for additional discussion of the models considered during stepwise model selection.

When you specify model selection based on the AIC statistics, the selection report looks differently (see Figure 13-9).

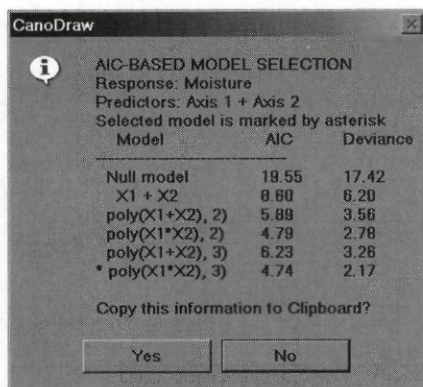


Figure 13-9 Report on GLM selection using AIC statistics

CanoDraw uses a different strategy for model selection here. It performs a stepwise extension of the systematic part of the regression model, but does not stop at the stage where no additional improvement is seen, but goes through all the candidate models and evaluates their parsimony using the AIC. The selected model (the one with the lowest AIC value) is then marked in the report dialog by an asterisk in its front, and all the considered models are shown.

The contents of these report dialogs can be copied to the Windows Clipboard.

13.6.2 Generalized Additive Models (GAM)

GAMs were already discussed in section 12.3.1.4, and an in-depth description can be found in Hastie & Tibshirani (1990). Response curves based on fitted generalized additive models do not have such rigid form as for linear or polynomial GLMs, so their use is recommended in situation where the shape of response curve has to be suggested by the actually observed data or where the assumptions about the response curve shape are being validated.

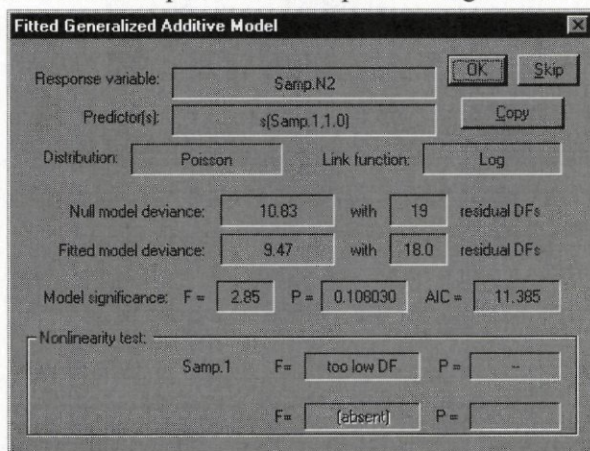


Figure 13-10 Fitted GAM summary dialog

The dialog summarising the fitted GAM is illustrated in Figure 13-10. Its largest part is identical with the dialog used to summarise a generalized linear model (GLM), described in the preceding section. The only difference is that the degrees of freedom of the model and, consequently, also the residual degrees of freedom, may be represented by fractional numbers.

You should also note that the deviance-based test (represented by the F and P field values) is even more approximate than for the GLMs.

The dialog area named *Nonlinearity test* displays the results of the approximate test(s) of "non-linearity" of the smooth terms for the individual predictors. With a cubic smoothing spline (estimated using a penalised form of least squares), a linear component with 1 DF can be "extracted" from it and the amount of variability explained by the non-linear part of the smooth term can be then tested, as suggested by Chambers & Hastie (1992). CanoDraw uses here an F-ratio based test, not the originally suggested χ^2 -based test. In Figure 13-10, the field for the first predictor shows *too low DF*, because the whole smooth term complexity was set to $df=1$, so nothing was left for the non-linear component. The field for the second variable says (*absent*), as this model had only one predictor.

The *Copy* button can be used to place the model summary on the Windows Clipboard in a text format. You can use the *Skip* button to ask CanoDraw not to plot the response curve (or response surface, with two predictors) corresponding to this model. Note that if there is no additional plot contents, the diagram creation is cancelled.

If you select complexity of a generalized additive model based on the AIC statistic, CanoDraw can report selection results with a dialog shown in Figure 13-11.

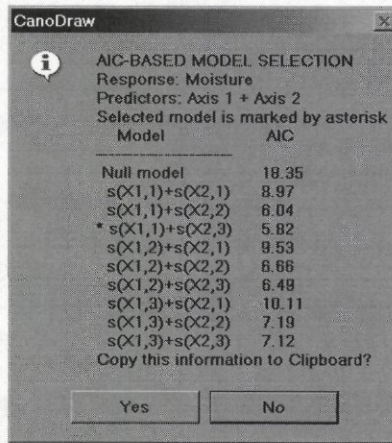


Figure 13-11 Report on GAM selection using AIC statistics

As for generalized linear models, CanoDraw reports here all the candidate models and marks the selected one (with the lowest AIC value) with an asterisk preceding the model description.

13.6.3 Loess (locally-weighted regression) models

A loess smoother represents the most flexible regression model. Unlike the generalized additive models, loess models do not separate the effects of multiple predictors (explanatory variables). If you have two or more predictors, their joint effect is modelled by fitting local regression models, with their definition (which data points are used and what weight they have) changing smoothly across the data space (see Cleveland & Devlin 1988 for the detailed description of the loess method). The implementation of loess used by CanoDraw allows, in the case of two predictors, to define a mixed model standing on the border between a parametric (linear) regression model and the smoother (using so called conditionally-parametric terms – see section 12.3.1.5). After a loess model is fitted, a summary dialog is displayed, as illustrated in Figure 13-12.

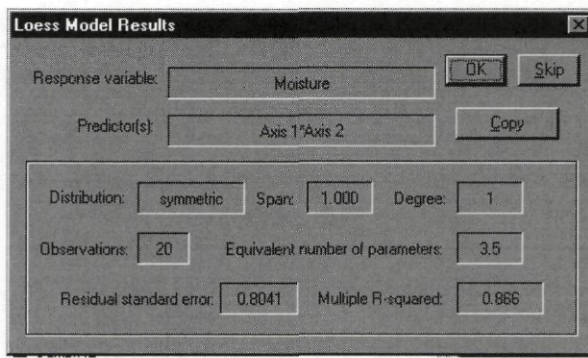


Figure 13-12 Loess Model Results summary dialog

This dialog shows the names of explanatory and response variables, model options (*Distribution*, *Span*, and *Degree*) and the main properties of the fitted model (*Residual standard error*, *Multiple R-squared* – i.e. the amount of explained variability in the response variable).

The *Copy* button can be used to place the model summary in a text format on the Windows Clipboard. You can use the *Skip* button to ask CanoDraw not to plot the response curve (or response surface, with two predictors) corresponding to this model. Note that if there are no additional plot contents, the diagram creation is cancelled.

13.6.4 Regression diagnostics in CanoDraw

Whenever you fit a regression model, you should explore not only the estimated regression coefficients or the fitted regression curve (or surface), but you should also check whether the assumptions made during model selection are fulfilled by your data. The set of techniques developed for checking such assumptions and the fitted model properties are known as **regression diagnostics**. Most (but not all) of the individual regression diagnostic methods work with the **regression residuals**. Regression residuals quantify the discrepancy between the true values of a response variable and the corresponding predictions, made by the regression model (the **fitted** values). The simplest type of the regression residuals, so called **raw residuals** are calculated as the difference between the observed and fitted values of the response variable.

CanoDraw provides only a limited subset of regression diagnostic methods. It allows the plotting of three types of residuals (discussed below, when describing the dialog illustrated in Figure 13-14) against the fitted values and such plots can be used for two purposes: (a) to visualise the changing variability of residuals with the changing predicted values, i.e. response variable **heteroscedasticity**, or (b) to check for grossly underestimated variability – "curvature" of the true response curve / surface.

Additionally, CanoDraw allows you to plot the residuals against the predictor variable(s) of the originally fitted model. Such a plot can detect the inadequate description of the effects a particular predictor has over the response variable. For example, conclusions about using a polynomial term instead of a linear one can be made, based on such a plot.

To create a regression diagnostic plot, called a **residual plot** in CanoDraw, you must start from an existing diagram containing one or more fitted regression models. If you click within this graph outside of any selected graph object with the right mouse button (and the CanoDraw project from which the graph was created is available), the context menu contains command *Residual plots...* and if you select it, CanoDraw displays the dialog illustrated in Figure 13-14. This dialog can be preceded by the dialog shown in Figure 13-13, where you can select which regression model of the multiple models present in a graph, you are interested.

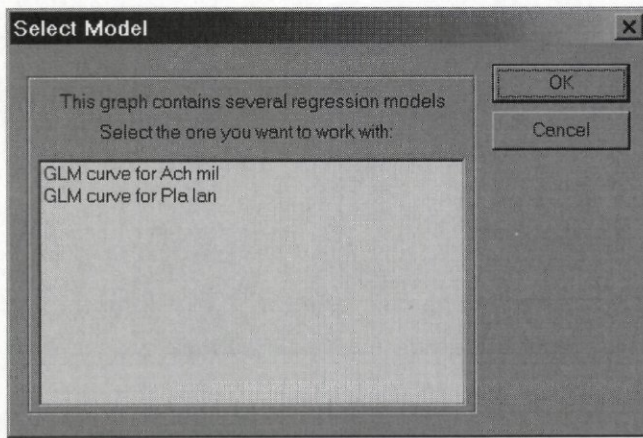


Figure 13-13 Select Model dialog box

The dialog box used to specify residual plot contents is shown in the following figure.

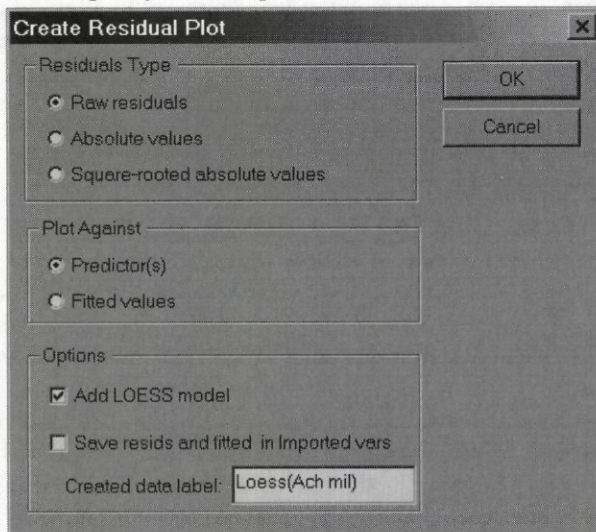


Figure 13-14 Create Residual Plot diagram

The area marked *Residuals Type* specifies how the regression residuals should be transformed. CanoDraw can plot either the original raw residuals, or it can take their absolute values, or the square-root of the absolute residual values. The last type of residuals is good for detecting heteroscedasticity in your model.

The selected type of residuals can be plotted against either the predictor(s) of the original regression model or the fitted (predicted) values of the response variable, and you can select which kind of plot to create in the dialog area named *Plot Against*.

You can generalise the pattern of regression residuals in the residual plot by adding a smooth curve to them, using a loess model. To do so, check the box preceding the *Add LOESS model* option. Note that this option is enforced if you have two predictors in the model and you decide to plot residuals against them.

This dialog can be also used to store permanently the residual values (of the selected type) and the fitted model values in the CanoDraw project, among the **imported** variables. To do so, check the box preceding the *Save resids and fitted in imported vars* option and specify the base

name for the two variables. In our sample dialog, the base name is specified as *Loess(Ach mil)*. The two variables created in the *Imported variables* folder will be therefore named *Loess(Ach mil).Fitted* and *Loess(Ach mil).Residuals*. You can then use the two stored variables to create other types of regression diagnostic plots, with the general XY(Z) attribute diagrams.

14. CanoDraw Examples

In this chapter, several examples of creating graphs from the results of statistical analyses performed with the Canoco software are shown. These examples use selected sample projects described in the Chapter 8 of this manual. The description of the steps needed to produce the graphs tries to be reasonably brief and you should check the suggested parts of the command reference to fully understand them. Each example description starts from a new CanoDraw project created from an existing Canoco sample project. The Canoco sample projects are optionally installed with the Canoco for Windows software, in the *Samples* sub-directory of the Canoco installation folder. You are supposed to create the initial CanoDraw projects yourself, using the suggestions contained in section 12.1.1. CanoDraw samples subdirectories (under the *Samples\CanoDraw* subdirectory) contain the final states of the CanoDraw projects and graphs.

The behaviour of CanoDraw as well as the actual appearance of the graphs you create can differ from the description provided here. This happens if you already worked with the CanoDraw for Windows software before and changed the program settings.

14.1 SPIDER1

Canoco sample directory: *Samples\Unimodal\Spider1*

Description of the corresponding Canoco example starts on page: 226

We will start our exploration of the data used in this example by looking at the results of canonical correspondence analysis (CCA), defined by the Canoco project *spid_cca.con*. You should create a CanoDraw project from it first. There are two possible ways of creating a new CanoDraw project. First, if you work with a Canoco project within the Canoco for Windows application, you can click the *CanoDraw* button to start the CanoDraw program. CanoDraw automatically suggests to create a new CanoDraw project from the currently active Canoco project. If the *CanoDraw* button is disabled in the Canoco project view, the ordination results (in Canoco *sol* file) are probably out of date or absent and you must click the *Analyze* button to actualize them. Alternatively, you can start CanoDraw yourself from the *Start / Programs* menu of Windows and create a new project directly within the CanoDraw program, using the *File / New Project* menu command and specifying the file (with *con* extension in its name) containing the Canoco project on which the new CanoDraw project should be based.

We will continue by creating a triplot diagram, containing symbols for samples and species, and also arrows for environmental variables. You may create the triplot diagram easily using the *Create / Triplots / with Environmental variables* menu command. Note that the resulting diagram (not shown here) is perhaps too overcrowded, particularly due to the many sample labels (containing sample indices). Sample identity is probably not so important for interpreting the ecological relations in these data. Therefore, you need to specify that CanoDraw should not label the symbols representing individual samples. Close the graph first (using the **X** box in its upper right corner; select *No* when asked about saving the graph) and then execute menu command *Project / Settings* and select the *Appearance* page in the displayed dialog. The options for labelling different types of items in ordination diagrams are in the upper part of that dialog page: change the selection in the *Sample labels* row to *None* (see Figure 14-1).

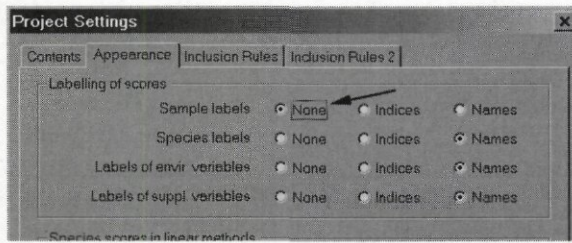


Figure 14-1 Changing labelling style for samples

After you close this dialog with the *OK* button, create the triplot diagram again, using the *Create / Triplots / with Environmental variables* command. The resulting diagram differs somewhat from the one displayed in Figure 14-2.

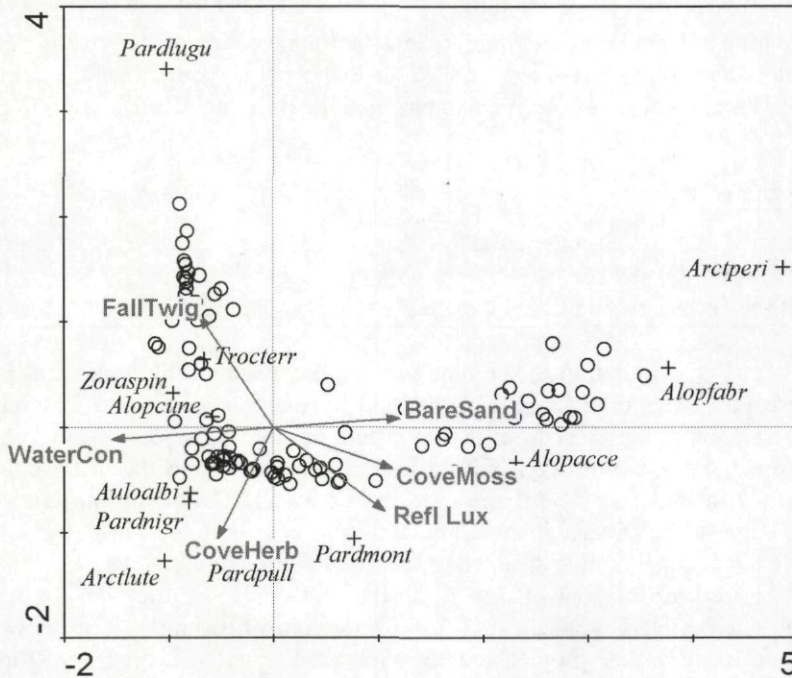


Figure 14-2 Triplot based on SPID_CCA project

The above graph was adjusted after its creation using the following three methods:

- * First, the background of the labels for the environmental variables was made opaque. To do so, you must select one of the labels (say *Refl Lux*, for example) and then select the other labels using the *Select Suchlike* command from the context menu (see section 13.1.3). Then invoke the *Properties Sheet* with the menu command *View / Properties Sheet* or with the *F5* key and change the settings on the *Fill* page as indicated in Figure 14-3. Note that the decision whether to use opaque or transparent background of the labels is often very difficult: while the labels with opaque background become more readable, the underlying symbols and / or labels become more obscured at the same time.

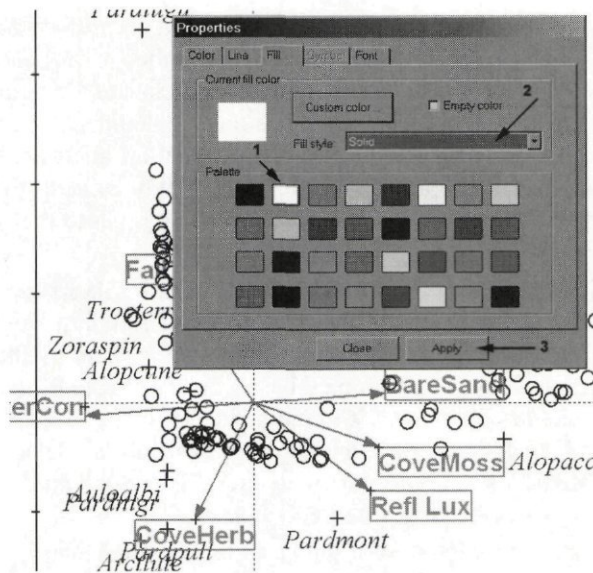


Figure 14-3 Making the background of labels opaque

- * The positions of species and environmental variable labels were adjusted to minimise their overlap. Check the section 13.3 for a description of how to reposition the labels within CanoDraw graphs.
- * After that, the labels of the environmental variables still interfered with the sample symbols. The circles were overwriting the opaque background because they were drawn later than the labels. To override that, you must open the additional *Graph Contents* window, using the *View / Tree View* command. Then select the group named *Sample scores*, click it with the right mouse button, and from the popup menu select the command *Move group upwards* (see Figure 14-4). Close or minimize the *Graph Contents* window to return back to the CanoDraw graph.

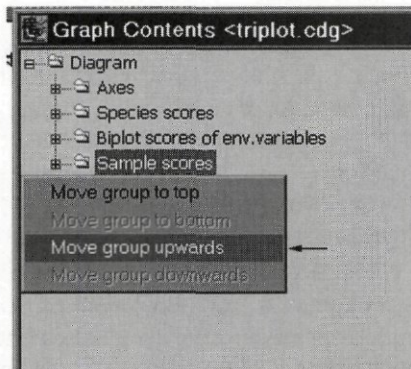


Figure 14-4 Moving group of graph objects upward in the hierarchy

The triplot diagram in Figure 14-2 plots both the active and the supplementary samples in the same way (using the empty circles). You can try to differentiate the active samples from the samples with missing environmental data by following the additional steps. This exercise is not recommended unless you already have some experience with the CanoDraw for Windows program.

- * Create a group of active samples by the *Project / Define groups of / Samples* command and in the displayed dialog box select the *By Rule* button in the *Create* section. In the dialog box *Modify Group* select the *weight in analysis* option and change the value in the *FROM* field, replacing the zero value with *0.1*. This excludes the supplementary samples from the group, because they have a zero weight in constrained ordination analysis. Close the dialog box with the *OK* button. You can optionally rename the newly created group of samples, using the *Rename* button in the *Sample Groups Manager* dialog. Close then this dialog box using the *Close* button
- * The created sample group can be then used in the *Inclusion Rules* page of the dialog invoked by the *Project / Settings* command to plot only the active samples. But we will prefer to plot both kinds of samples in the same diagram, differentiating them by their appearance. To do so, define a new classification of samples based on the created group of active samples. Using the *Project / Classify / Samples* command, open the *Available Classifications of Samples* dialog and click there the *New from group* button. Select the recently created sample group from the list (confirming with the *OK* button) and a new classification is created. Check the box titled *Use this classification in diagrams*.
- * If you want to change symbols used for the active and supplementary samples (contained in the first and in the second class, respectively), you can do so in the dialog invoked by the *View / Visual Settings* menu command. In this dialog, select in the *Attribute category* field the section *Samples / Symbols* and modify there the attribute settings for *Class 1* and *Class 2*.

Before we continue our tutorial, you might like to close the created graph(s) and the CanoDraw project that we worked with until now. This can be done simply by clicking the close (X) buttons of their windows. It is useful to close the graph windows first, so that the still opened project may notice the names of files, in which the graphs were saved. If you created several graphs from a project, the *Window / Close graphs of active project* command can provide an useful short-cut (see section 12.7.5). For each graph being closed, CanoDraw asks you whether to save it. Graphs can be re-opened and modified later on, with or without their parental project.

Additionally, you can print each graph (using the *File / Print* command, see section 12.1.9) or export it into a file with a different format (using the commands in the *File / Export* submenu, see section 12.1.7).

The section about the *SPIDER1* Canoco example (p. 226) suggests to compare the sample scores of two different DCAs, using the Canoco project *dca28100.con*. The one DCA uses 28 samples (the ones with the environmental data available), while the other DCA uses all 100 samples. In the project *dca28100.con*, the axes of the latter analysis are used as supplementary environmental variables with the former DCA. The correlation between the scores on the first ordination axes, for example, can be then shown using the arrow of the supplementary variable *AX1*. Here we show not only this graph, but also two alternative ways of visualising the closeness of the sample scores of the two analyses.

You must start by creating a new CanoDraw project called *dca28100.cdw* from the Canoco project *dca28100.con*. The simplest presentation of the relations between the ordination axes of the two detrended correspondence analyses can be seen from the graph in Figure 14-5.

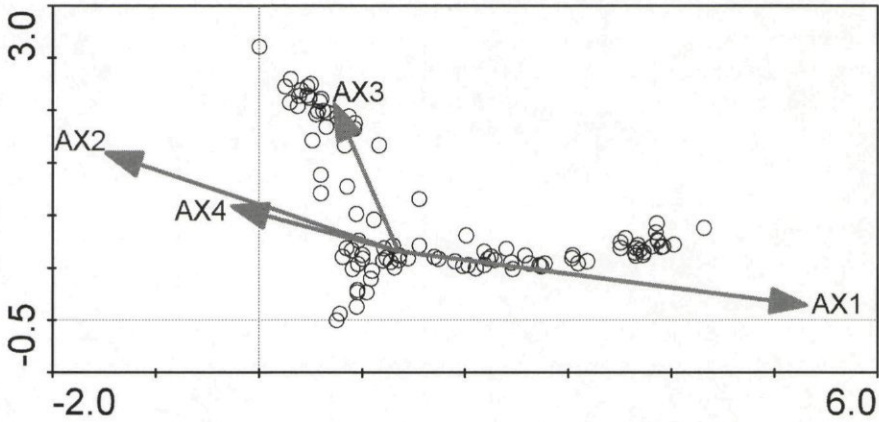


Figure 14-5 Comparison of two DCAs

To create this graph, select the menu command *Create / Biplots and Joint Plots / Samples and suppl. variables*. The sample scores are initially labelled by indices, but you can get rid of the sample labels by selecting one of them, pressing *Ctrl-H* (the shortcut to *Select suchlike* command), and then pressing *Delete (Del)* key to remove all the labels. This diagram shows a good correlation between the first ordination axes of both analyses, but not such a good correlation between the scores on the second ordination axes. But do not jump to conclusions about correlations among the variables represented by the arrows (i.e. correlation between the scores on different axes of the DCA with 100 active samples). The correlations among supplementary (or environmental) variables are not presented optimally in such an ordination plot. Note also that there are 100, not 28 circles plotted in the diagram. This is because even in the analysis with environmental variables, there were 100 samples, but 72 of them were passive.

One alternative way to show the relation between, say, the first ordination axis of the DCA with 28 samples and the first few ordination axes of the DCA with 100 active samples can be seen in Figure 14-6.

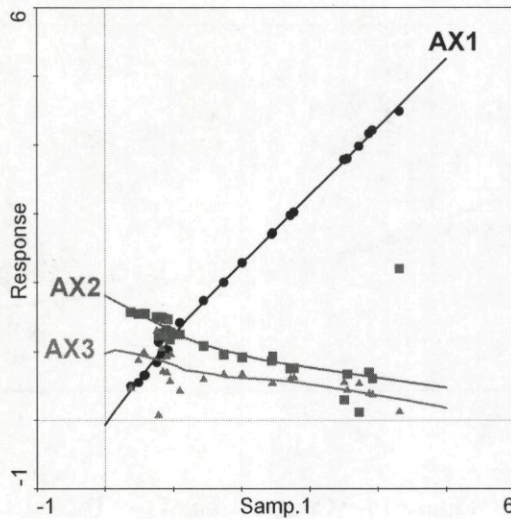


Figure 14-6 Relation between first axis scores of one analysis and scores on first three axes of another analysis, visualised using a loess smoother

In this graph (stored as *dc28100a.cdg*), only the scores of 28 samples, which were active in both analyses, are shown. To achieve this, you must exclude the supplementary (passive) samples, i.e. the samples with zero weights. This can be done in the *Inclusion Rules* page, which is part of the dialog invoked by the *Project / Settings* menu command (see Figure 14-7).

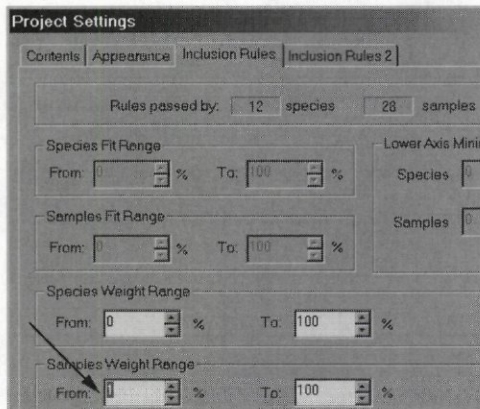


Figure 14-7 Selecting only active samples

Specifying value 1 in the indicated field excludes the samples with a weight less than 1% of the largest sample weight in the data.

To create the actual plot, use the *Create / Attribute Plots / XY(Z) Plot* command and make the following choices in the setup dialog (Figure 14-8):

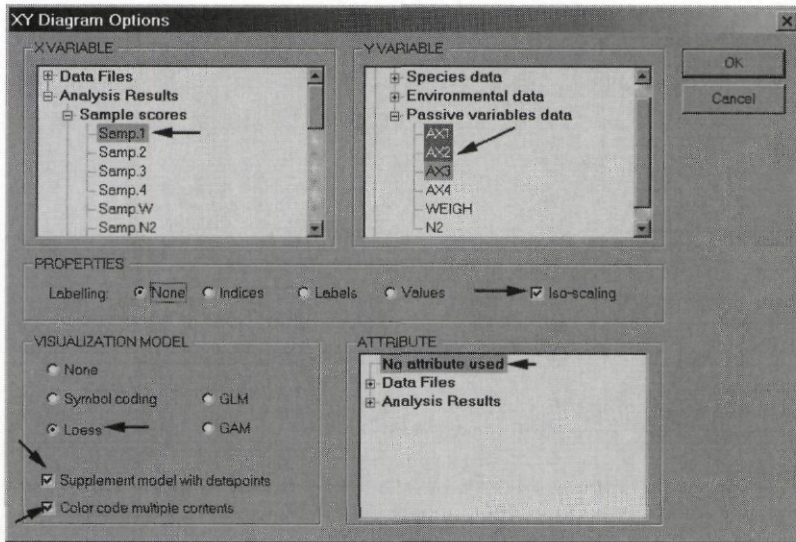


Figure 14-8 Creating XY diagram with multiple response variables

The selection in the *X VARIABLE* list can be achieved by expanding the *Analysis Results* item (clicking on the + box in its front) and then the *Sample scores* item. Similarly, you must expand the *Data Files* item and the *Passive variables data* subitem in the *Y VARIABLE* list. Worth of mentioning is the multiple selection in the *Y VARIABLE* list: only this list supports the multiple selection. To select additional items there, combine the left mouse button with the *Ctrl* key. Also note the choices made in the *VISUALIZATION MODEL* area of which the *Loess* model selection is the most important one. The effect of the *Color code multiple contents* option is difficult to see in our gray illustration, but the color-coding provides very distinctive graphs on screen or color printer print-outs.

After clicking the *OK* button, three additional dialogs are displayed, corresponding to loess model options for each of the three response variables. Leave the default choices in the dialogs: linear local regression, span equal to 0.67, robust algorithm. The resulting diagram was slightly modified for printing purposes (increased size of labels, increased width of lines, and differentiation of symbols which were originally differentiated only by their color).

Finally, we will demonstrate how to show the relation between two axes of the current analysis and one axis of another analysis. We do so using a contour-based attribute plot. From the menu, select the *Create / Attribute Plots / Data Attribute Plot* and specify its contents as illustrated in Figure 14-9.

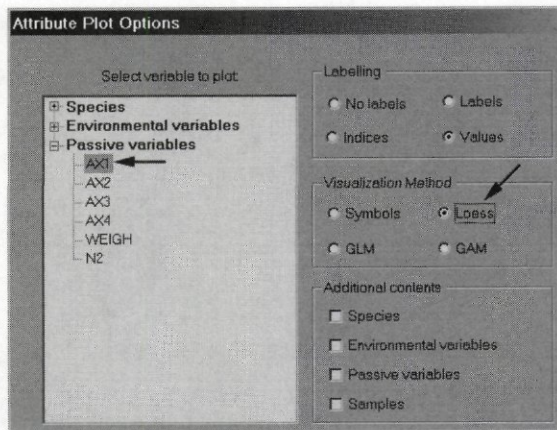


Figure 14-9 Data Attribute Plot dialog

After you click the *OK* button, CanoDraw displays a dialog where the loess settings can be adjusted. Keep the default values for local polynomial degree (linear) and for span (0.67), and make sure that the option *Normalize scale for two predictors* is not checked (because the scores on two ordination axes are already on the same scale). CanoDraw then displays a summary of the fitted model and suggests values for the contour levels (from 0.5 to 5.0, with step 0.5). Confirm them with the *OK* button. The resulting attribute plot is shown in Figure 14-10.

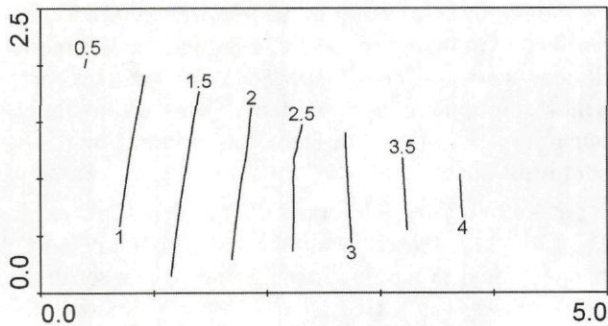


Figure 14-10 Contour-based attribute plots displaying the patterns of first axis scores of the other DCA

Obviously, the largest correlation is seen between the scores on the first axes of both analyses. Contour lines do not fill completely the area, only the interpolated part of the fitted loess surface is shown. This can be changed using the *Plot also the extrapolated values of response* in the *Diagram Settings* dialog box (see section 12.3.1.1).

To remove the currently open project and the graphs created from it, you can close the individual windows separately or you can use the *Window / Close all* menu command.

14.2 SPIDER2

Canoco sample directory: *Samples\Unimodal\Spider2*

Description of the corresponding Canoco example starts on page:: 230

This example focuses on interpreting the relation between species optima (represented by weighted averages) and individual environmental variables, using the variable Water Content as an example. We will visualise the relation of spider species to an environmental variable by fitting species response models.

Start by creating the CanoDraw project *spider.cdw* from the Canoco project *spider.con*. We want to see which species have a good relation with the substrate water contents. As this variable is strongly correlated with the first CCA axis, you can base your decision on the strength of the relation of individual species with the first ordination axis. To do so, define a group of species based on their fit statistics on axis 1. The fit measures the percentage of variability in the species values which can be explained by the species symbol position on the axis, with respect to the positions of individual samples. To define the group, select the *Project / Define Groups of / Species* command. In the dialog with title *Species Groups Manager*, select in the *Create* area the button *By Rule* (which is immediately below the *Close* button). Another, larger dialog appears (titled *Modify Group*) and there you should select the criterion used to define group membership for individual species. Our criterion is the *fit on first ordin. axis* option in the first column. As you click it, the values in the *FROM* and *TO* fields on the right side change, respectively, to *0.0968* and *0.8145*. The former value represents the smallest value found among the 12 species, while *0.8145* is the value of the species with the best fit. The *FROM* and *TO* fields specify the allowed range of the criterion for a species to be a member of the group. Therefore, initially all the species are group members. We will limit the group range by increasing the value in the *FROM* field somewhat, for example to *0.33*. This implies that a species must have at least one third of the variability in its values explained by the first ordination axis to be a group member. You can see in the lower right area that five species pass this condition, and you can check their names. Close this dialog using the *OK* button and return to the groups manager dialog. The new group is named *Group of species 1*. Change this name to something like *Good fit on axis 1* by clicking the *Rename* button and entering the desired group name. Then close the groups manager dialog using the *Close* button.

To specify that only the species from the group just created should be plotted, you must go to the dialog invoked by the *Project / Settings* command and select the dialog page named *Inclusion Rules*. At the bottom is the area named *Limit to Group*. Select there the group name instead of the *DO NOT USE* value. **Note that these two steps (creating species group based on their fit and specifying that only group members are plotted) can be replaced by one, simpler step: you would enter on this page the value of 33 (instead of default 0) in the *Species* field of the *Lower Axis Minimum Fit* area. We followed the more complicated way because you will use the defined group also at another occasion. Also note that the alternative approach does not refer to fit on the first axis, but rather on the horizontal axis, whatever ordination axis it represents.**

Before you close the *Project Settings* dialog, note that the *Contents* page contains an option named *Plot SAMP scores even for constrained axes*, and that by enabling this option, you would have *Samp* scores (instead of *SamE* scores) used in the ordination diagrams created from this project. Note however, that enabling this option is probably not appropriate for our current example. Close the settings dialog with the *OK* button and create a biplot diagram with species and environmental variables, using the command *Create / Biplots and Joint Plots / Species and*

env. variables. This diagram (illustrated in Figure 14-11) displays only the five species having a good fit with the first ordination axis.

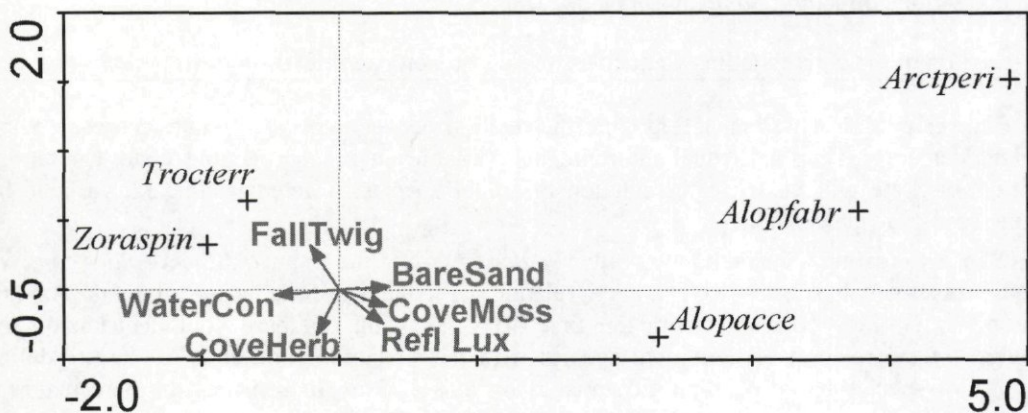


Figure 14-11 Biplot of well-fitting species and environmental variables

Now we would like to describe the relation between the species and the variable *WaterCon* (or the sample scores on the first ordination axis: we assume these two predictors are closely correlated). To do so, we will use generalized additive models, as these models do not substantially prescribe the shapes of fitted response curves (compared, e.g. with second-order polynomials we will use later on). In the additive models, we will determine only their complexity, measured with degrees of freedom. We will use the model selection procedure to select the optimum model complexity separately for each of the species. You should start by selecting from the CanoDraw menu the *Create / Attribute Plots / Species response curves* command. The displayed dialog is illustrated in Figure 14-12, including the required settings.

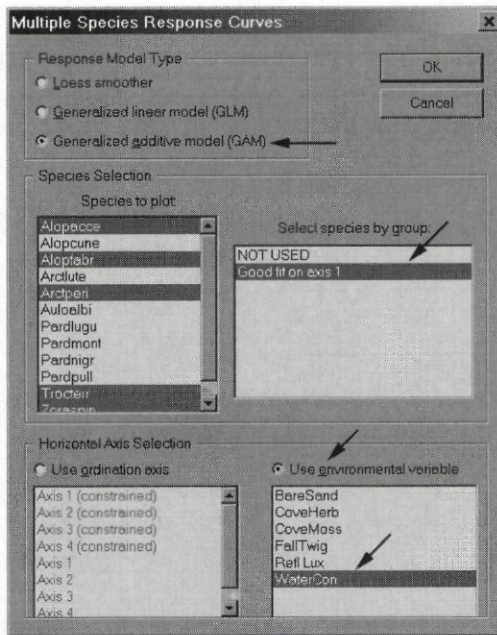


Figure 14-12 Species Response Curves with GAM

In the upper part of the dialog, select the *Generalized additive model (GAM)*. Then select the set of response variables (species) by selecting the name of species group in the middle right dialog area. Additionally, you should change the predictor selection from ordination axes to environmental variables, and select the *WaterCon* variable. After clicking the *OK* button, a new dialog is shown, where you specify settings for the generalized additive models (see Figure 14-13).

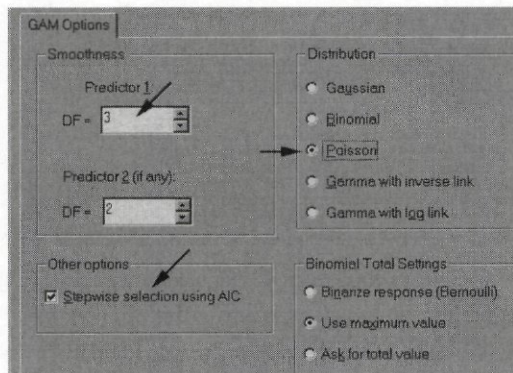


Figure 14-13 GAM Options for species response curves

Given the fact that we ask for model selection using AIC (see checked box in the lower left corner), the specified complexity for *Predictor 1* (DF=3) represents an upper limit for the evaluated model complexities. When selecting the model, CanoDraw evaluates the model performance using AIC statistics (see section 12.3.1.4 for additional details) for varying complexity of model terms for each predictor. It starts with the null model, and continues with a model term with complexity corresponding to one degree of freedom and increasing by one DF until the specified upper limit is reached or exceeded. Therefore, in our case with an upper limit equal to 3.0 and only one predictor, four alternative model specifications are compared: a null model, a model where the *WaterCon* predictor has a complexity $df=1$, a model where the

WaterCon term has a complexity $df=2$, and a model with the *WaterCon* term complexity $df=3$. You should also specify that a Poisson distribution of the species values is expected, with the logarithmic link function implied by CanoDraw.

After you close this dialog with the *OK* button, CanoDraw reports on fitting individual regression model with two dialogs for each of the species. We will illustrate them for the first species, *Alopacce*. CanoDraw first summarises the model complexity selection process (see Figure 14-14).

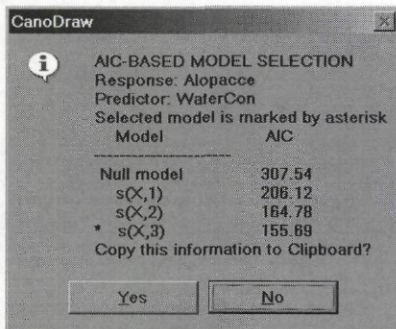


Figure 14-14 Report on regression model selection

The most complex form of describing the effects of soil water contents upon this species has the lowest AIC value, so it was selected. You can copy this report in text format to the Windows Clipboard. Note that the default answer is *No* (i.e. nothing is copied to Clipboard). After you close this dialog, the fitted GAM summary dialog appears (Figure 14-15).

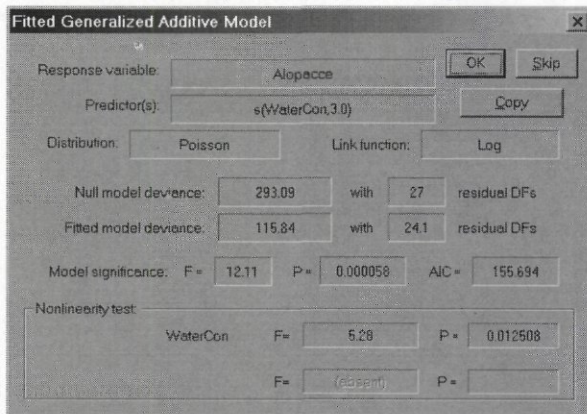


Figure 14-15 Fitted GAM model summary

You can see in this dialog that the fitted model fares much better than the null model (more than 60% of total variability was explained by the model) and that it is also better than a generalized linear model with a linear form of dependence of *Alopacce* upon *WaterCon* ($P=0.0125$). If you want the response curve for this model to become part of the diagram, click the *OK* button. If you prefer suppression of this particular model, click the *Skip* button. This is useful in the situation where the model selection procedure selects a null model (displayed by a horizontal line in the diagram). In our example, all species have a strong relation to soil water content, so we include all of them in the resulting graph.

If you are not interested in seeing the reports on model selection and the summaries of the fitted regression models, you can suppress them by disabling the *Show summary of each fitted regression model* option in the *Properties 1* page of the *Diagram Settings* dialog.

The created graph (Figure 14-16) was additionally adjusted for non-color printing by changing the line style of some of the response curves.

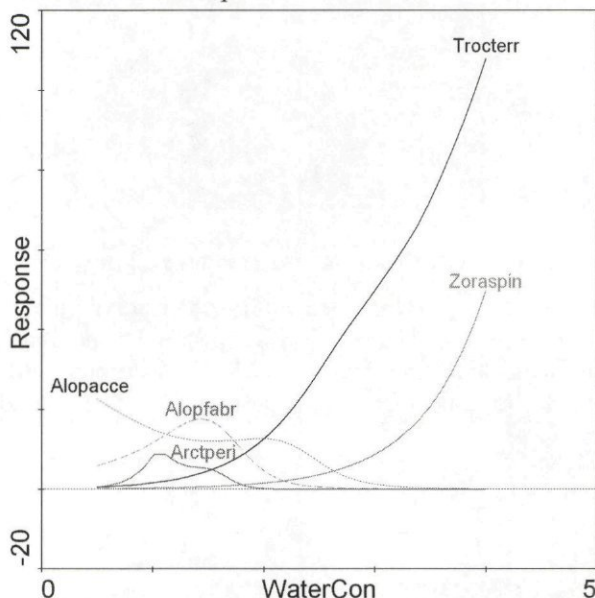


Figure 14-16 Species response curves fitted using generalized additive models

Note that the maximum value for the vertical axis is perfectly sound, the species *Trocterr* has a maximum value of 118 in sample "Pitf 24".

Now we will use an alternative approach involving fitting parametric generalized **linear** models (GLMs). We will again use model selection, this time a stepwise selection where the null model is compared, in turn, with the linear and second-order polynomial models. The selection is based on a parametric test using an analysis of deviance (see section 12.3.1.3). You must start, again, with the *Create / Attribute plots / Species response curves* command, but in the first dialog, select the *Generalized linear model (GLM)* option, instead of GAM. Set the other options as the last time, when you fitted generalized additive models. In the following dialog with GLM options (Figure 14-17), select the following settings:

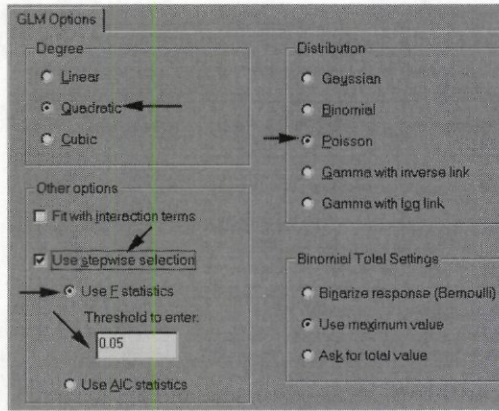


Figure 14-17 GLM options for species response curves

The *Degree* option has value *Quadratic*, so a third-order polynomial will not be considered during the stepwise model selection. The selection is specified to be done using the *F statistics* based test, and the significance threshold value is *0.05*. The distribution of response variables is set to *Poisson*. Model selection and model summaries are again reported before the graph is shown.

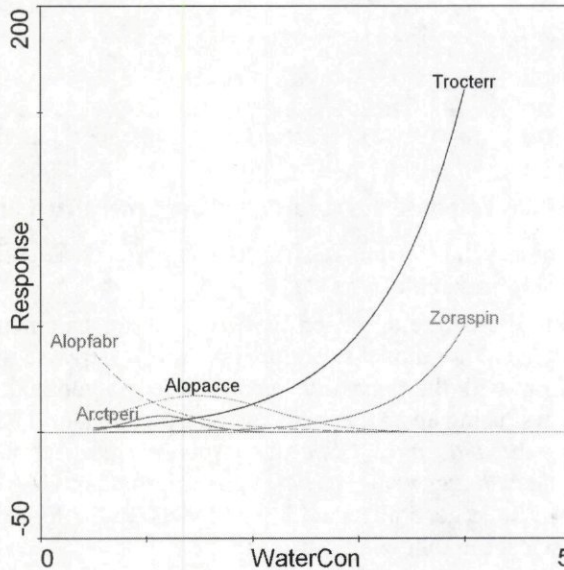


Figure 14-18 Species response curves fitted using generalized linear models

Comparing the linear models (Figure 14-18) with the alternative additive models (Figure 14-16) shows, for example, that for *Alopfabr* a linear model was judged better than the alternative unimodal shape, although the selected GAM suggests an unimodal response. This might be due to the non-symmetric character of the unimodal response, with the increase of *Alopfabr* values being slower than their subsequent decrease along the gradient of increasing water contents.

14.3 DYKE

Canoco sample directory: *Samples\Unimodal\Dyke*

Description of the corresponding Canoco example starts on page:: 234

In this example, we will illustrate the problems mentioned in the original Canoco example: how to specify the plotting of nominal explanatory variables by centroids, how to select which species appear in an ordination diagram, and how to display the scores on the third ordination axis. Start with creating a CanoDraw project *dyke_cca.cdw* from the Canoco project *dyke_cca.con*.

The three variables representing soil type are nominal variables, classifying the samples into three distinct classes, and are better represented in ordination diagrams by centroid scores. You arrange for this in the CanoDraw project using the dialog invoked by the *Project / Nominal variables / Environmental variables* command. In the dialog, select the three variables *Peat*, *Sand*, and *Clay* in the left list and transfer them into right list using the *Select* button.

Next, you will specify the plotted species using the rules suggested in the Canoco example text (p. 235 of this manual): species should appear in the ordination diagram only if their fit to the diagram is 5% or more, and if they occur 10 or more times in the data. While the first condition is easy to specify, the limitation by the number of occurrences is more difficult to set. You should start by defining a group of species occurring in 10 or more samples.

You define a new species group using the command *Project / Define Groups of / Species*. In the group manager dialog, select the *By Rule* button in the *Create* area. A new dialog, titled *Modify Group*, appears and there you should select the *# of nonzero values* criterion. You can see from the range of values (1 to 64) that the species with the highest number of occurrences is present in 64 samples. Change the value in *FROM* field from 1 to 10. You can see that there are 61 species with 10 or more presences in the data. Close this dialog using the *OK* button and in the *Species Group Manager* dialog rename the group to "*More than 9 occurrences*" and close the dialog with the *OK* and then the *Close* button.

Now you can set the restrictions in the *Project Settings* dialog (invoked by *Project / Settings* command). In the *Inclusion Rules* page, start first with the species group you just created. At the bottom of this dialog page is the area called *Limit to Group* and there you must select the group in the *Species* list. Then you should change in the *Species Fit Range* area the zero value in the *From* field to 5. You can see that 23 species pass both inclusion rules (see Figure 14-19).

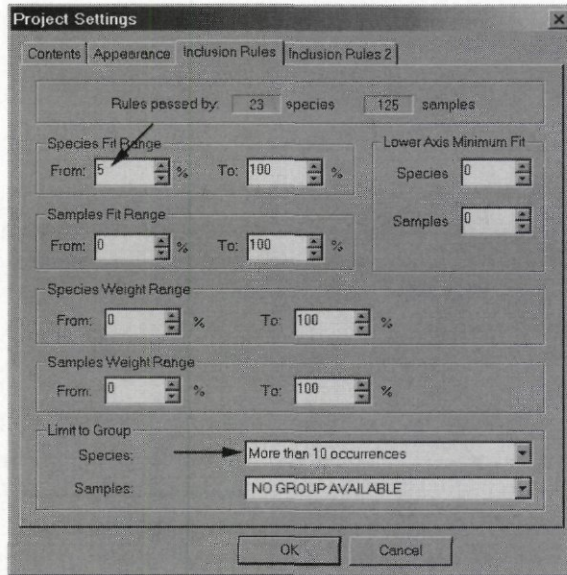


Figure 14-19 Limiting species presence in diagrams

Before you leave this dialog, switch to its first page (*Contents*) and take a look at the value of the *Axes to plot* option. The value is set to *First and second axis* and it can be easily changed to any combination of the first four axes. When you change it, the new settings remain in effect until reset again. Therefore, as we need to make just one diagram (species – environmental variables biplot) with a change in the ordination axes used, you will use an existing shortcut. From the *Create* menu, select the first command (*Simple Ordination Plot*), keep the default selection in its upper part (*Species and environmental variables*), but change the *Ordination Axes to Plot* option in the lower part, to the value *First and third*. There is one additional problem with the graph – the arrows of quantitative environmental variables are very short. You can change their rescaling with respect to the plotted species scores and the centroids for soil types. To do this, select the menu command *View / Diagram Settings*, and in the dialog first page (*Properties 1*), enable the option *Show rescaling coefficients for composite ordination diagrams*. Now you should create the biplot again (note that you **cannot** use the *Recreate graph* command, because it would apply the permanent selection of plotting first and second ordination axes), as described before. This time, a new dialog is shown before the diagram is created (see Figure 14-20). You should change the edit field in the *Explanatory variables* area from the original value 0.64965 to a larger coefficient value, e.g. 5.0. Note that the dialog shows a scaling value also for samples, despite the fact that this diagram shows the environmental variables and species. This is because the centroids of environmental variables are, in fact, the weighted averages of the sample scores, so their rescaling must be synchronised with the sample scores.

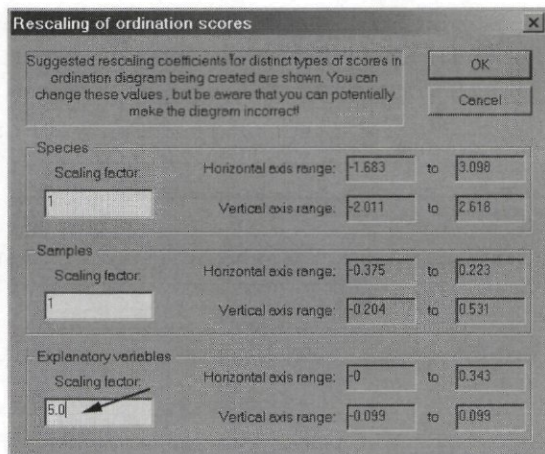


Figure 14-20 Changing scaling of ordination scores

The final graph (after repositioning of species labels and making their background opaque, with white fill color) is shown in Figure 14-21.

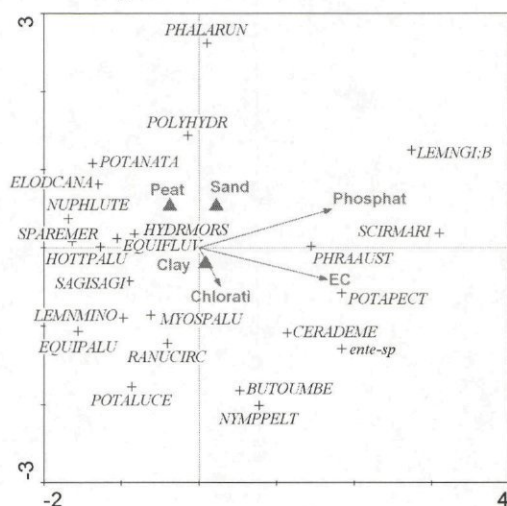


Figure 14-21 Species - env. variables biplot with first and third CCA axis

14.4 DUNEBOOK

Canoco sample directory: *Samples\Unimodal\Dunebook*

Description of the corresponding Canoco example starts on page:: 241

The data describing vegetation of dune meadows is a classical example, thoroughly discussed in Jongmann et al. (1995). The intermediate level of data heterogeneity allows us to demonstrate both the linear and unimodal ordination methods with the same data-set. We will limit our example to just one kind of analysis, represented by the Canoco project file *rda_spe.con*. In this project, a constrained linear method (redundancy analysis, RDA) is used to summarise information about the meadow vegetation composition, explainable by three of the

of different type and/or color in CanoDraw and you will use this feature to visualise the pattern of farming types across the plane spanned by the first two ordination axes.

Start with the menu command *Project / Classify / Samples* which displays the dialog illustrated in Figure 14-23 and described in more detail in section 12.4.3.

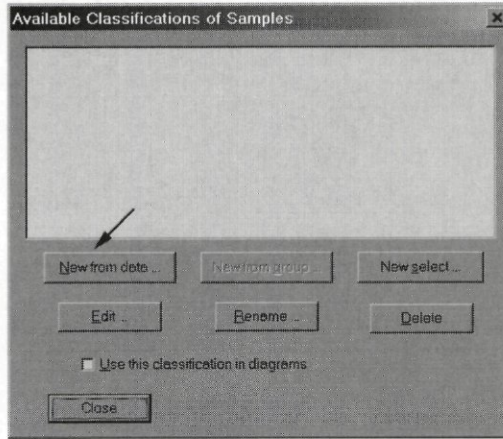


Figure 14-23 Dialog for management of sample classifications

There are no classifications available in the newly created project (the listbox is empty), and you will create the new classification using the *New from data* button. CanoDraw displays yet another dialog (shown in Figure 14-24) listing the variables which can be used to classify samples. The lower part of the dialog shows five strategies how the values of selected variable(s) can be used to define sample classes. The first four strategies are available if exactly one variable is selected. The fifth strategy (*Combine dummy variables*) is, on the other hand, available only if multiple variables were selected. CanoDraw allows multiple selection of variables (achieved by depressing the *Ctrl* key at the same time an item is clicked) only if each of the variables has just two distinct values – zero (0) and one (1). In our example, four such 0 / 1 variables were selected and the middle part of the dialog box contains a message confirming the selection of four nominal variables. You must then specify the classification *Strategy* (even when this is the only one permissible for a multiple selection of variables) and click the *Create* button.

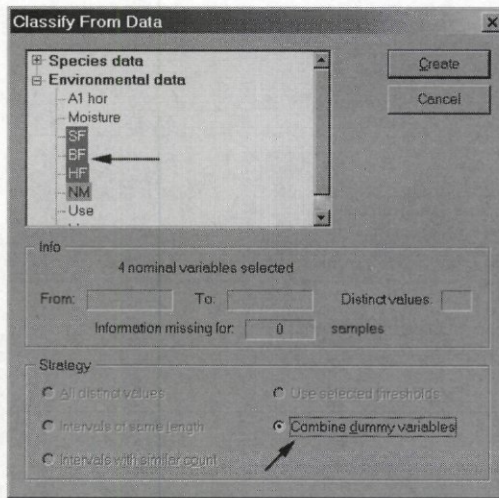


Figure 14-24 Classify From Data dialog

CanoDraw then creates the classification, creating separate class for each selected nominal environmental variable, and displays the classification in a new dialog (Figure 14-25). This dialog is titled *Manual Classification* because, at this point, you can adjust the classification by moving members from one class to another, remove classes or merge two or more existing classes into one class. Note, however, that before you close this dialog, all available items (samples, in this case) must be members of one of the defined classes.

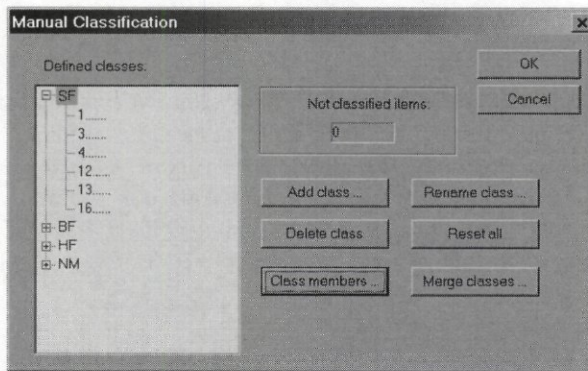


Figure 14-25 Manual Classification dialog

After you click the *OK* button, you are back in the original dialog for managing sample classifications. The new classification appears in the list and you can perhaps change its name now, to be more self-explanatory (use, for example, the name *Type of farming*). Also, you need to check the box in front of the *Use this classification in diagrams* option. Close this dialog using the *Close* button. Now you will check what symbols are used to represent samples from individual classes in ordination diagrams. To do so, create a biplot with samples and environmental variables using the command *Create / Biplots and Joint Plots / Samples and env. variables*. This initial attempt to present the relation between the type of farming and the explanatory variables representing soil development, soil moisture, and amount of manure is shown in Figure 14-26.

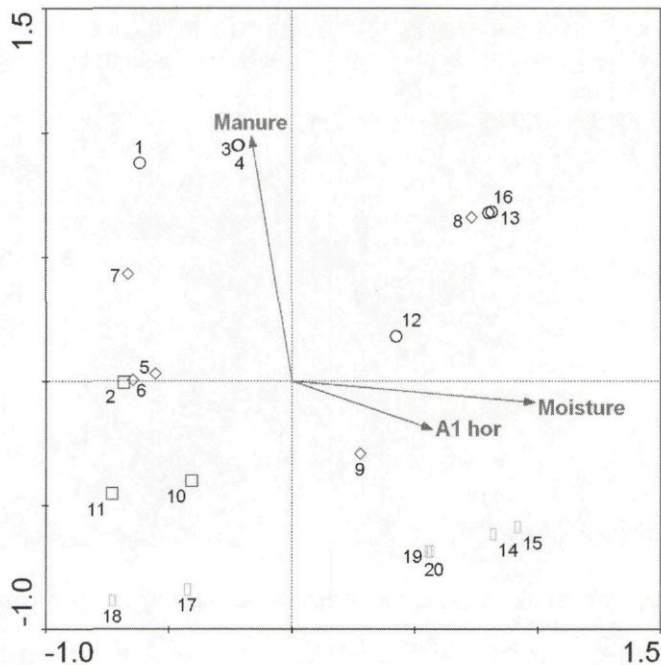


Figure 14-26 Samples and environmental variables biplot with sample symbols appearance coding the type of farming

There are several problems with this graph, listed below in the approximately decreasing order of their importance:

1. Symbols are too small and their types too similar, so it is difficult to group visually all the symbols representing one class.
2. There is no key that would enable us to recognise which type of farming corresponds to which type of item.
3. The indices used to label the sample symbols are perhaps redundant, if the primary purpose of this graph is to see the patterns in farming type distribution, not to identify individual samples.
4. As we present this graph without colours, it will be probably better to abandon the original color-coding and assign black drawing color to all the symbols.

To solve these problems, you need to make the following changes in the CanoDraw settings. These changes correspond one-to-one to the issues listed above:

1. We have four classes of samples, so you need to modify the visual attributes of sample symbols for the first four classes. To do so, open the *Visual Attributes Settings* sheet, using the menu command *View / Visual Attributes*. On its left side, locate the group named *Samples* and open it by clicking on the + sign preceding the label. Within the *Samples* group, open (expand) the subgroup named *Symbols*. This subgroup contains 64 items corresponding to attributes of symbols representing up to 64 sample classes. You will start with the first class, by clicking on the *Class 1* item. Select the *Symbol* tab in the right part of the dialog. Keep the *Symbol Type* setting (*Circle*) for this class, but increase the symbol size from *0.008* to *0.015*. The resulting state of this attribute page is illustrated in Figure 14-27. For the *Class 2*, you will again increase the symbol size to *0.015* and change the symbol type from *Square* to *Cross*. You may then switch to *Color* tab and change the violet color to black. For the *Class 3*, you will change the green color to the black one in page *Color* and in the *Symbol*

page, set the size to *0.015*, and change the symbol type from *Diamond* to *Up-triangle*. And, finally, for *Class 4* set again the black drawing color, and change the symbol type from *Box* to *Star* and its size from *0.008* to *0.015*.

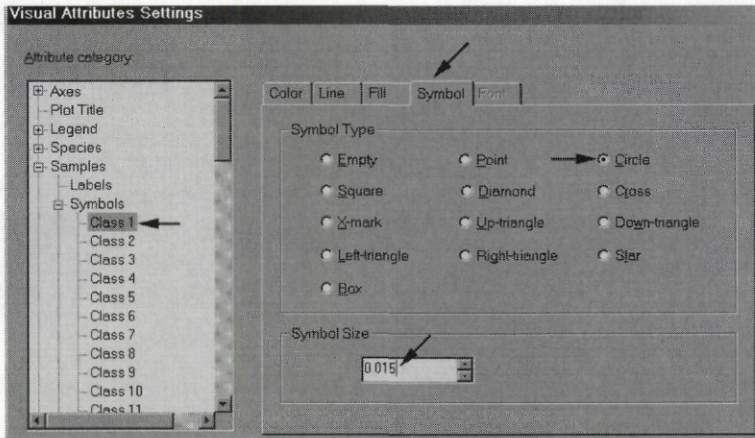


Figure 14-27 Visual Attribute Settings for sample symbols of first class

2. To display a legend for the diagram, use the *Project / Settings* command, and in the *Appearance* page of the *Project Settings* dialog, check (enable) the option *Insert legend into created diagrams*. The recommended setting for the *Legend position* is *Right side*, *Sections layout* should be *Vertical* and adjacent *Wrap after* value should be 2 or more. The layout of items in sections should be set to *Vertical* or *Vertical w. heading*, and the value *Wrap after* should be 4 or more.
3. To suppress the plotting of sample indices, you should change in the same dialog page, which you used in the preceding paragraph, one option in the *Labelling of scores* area: *Sample labels* should be set to *None* instead of *Indices*. Close then this dialog with the *OK* button.
4. The last change (concerning the black color of all symbols) was already made in the point 1 above. Note that you have manipulated only the drawing (outline) color of the symbols. This was sufficient because all the selected symbol types were either pre-set to be empty or they cannot be filled (the cross symbol). To change the fill settings, the options in the *Fill* attribute page would need to be modified.

Finally, create the new biplot diagram (Figure 14-28) using the command *Create / Biplots and Joint Plots / Samples and env. variables*. You can alternatively use the *Create / Recreate graph* command, but this would not change the labelling of sample symbols as this setting is invariable for an existing diagram.

The settings made to *Visual Attributes* in step 1 above are application-wide and will be automatically saved at the end of CanoDraw session. If you want to preserve the current visual attributes used in CanoDraw graphs before changing them, you must use the *File / Visual Attributes / Save* command to store them in a file. You can then revert to the stored settings using the *File / Visual Attributes / Load* command.

Additional support for easy visual assessment of symbols belonging to one sample class can be achieved with the envelopes. To enable plotting of polygons enclosing symbols of individual classes, you must select the *Project / Settings* command and in the dialog page named *Contents*, you need to check the *Samples* option in the area named "Draw Envelopes around Classes of". Note, however, that the color of envelope lines is based on the drawing color of sample symbols, so after our unification of drawing color, all the envelopes will have an identical black color.

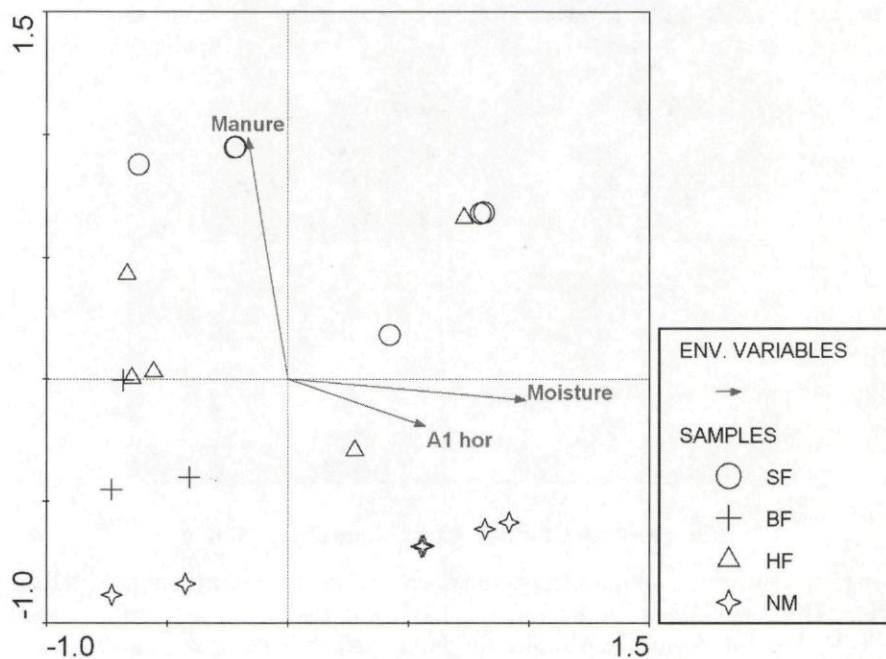


Figure 14-28 Final look of the biplot with samples and environmental variables

As another example of exploring the multivariate data-sets based on the ordination results, we might note that the first ordination axis is mostly correlated with soil moisture. We can therefore wonder, how the occurrence (and abundance) of individual species relates to this environmental factor. There are multiple approaches one can take when addressing such a question, including fitting species response curves along the moisture gradient (see section 14.2 for an example). Here we will use an alternative approach and we will start by classifying samples depending on soil moisture. We arbitrarily decide to have three classes of samples, corresponding to increasing soil moisture. Finally, we will visualise the distribution of species values over the three sample classes, using a diagram with the pie symbols.

To start, define a new classification for samples. Display the dialog with classifications of samples using the *Project / Classify / Samples* menu command (the dialog should already list the classification we created in the preceding part of this section) and click the *New from data* button. In the dialog that appears, select the variable *Moisture* in the *Environmental data* group. You can see in the middle part of the dialog box that this variable has four distinct values, the smallest one being 1 and the largest being 5. We would like to divide the range of its values into three so that there would be a comparable number of samples in each of them. To do so, select the option *Intervals with similar count* at the bottom of this dialog box and then click the *Create* button. CanoDraw asks you about the number of intervals you want to divide the range of the variable values into. Obviously, there cannot be more than four intervals, as there are just four distinct values available for this variable. Change the suggested value 4 to 3. After clicking the *OK* button, CanoDraw displays a new dialog titled *Confirm Class Boundaries*, illustrated in Figure 14-29.

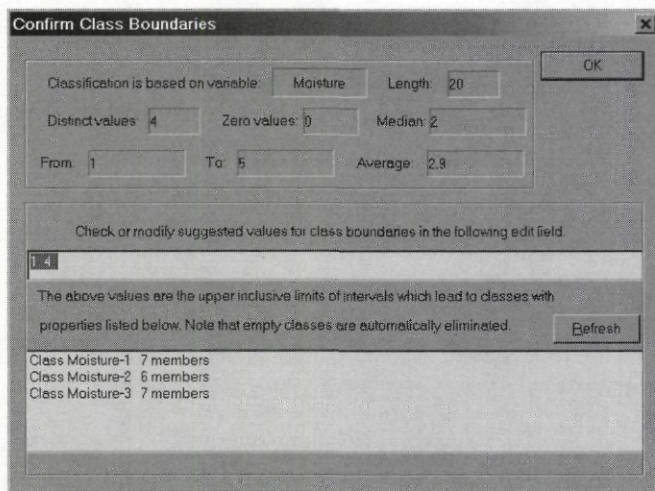


Figure 14-29 Confirm Class Boundaries dialog

The upper part of this dialog again summarises the variable representing the base for this classification. The essential part of the dialog is the line in its middle. The values entered in this line define the threshold values separating the individual regions of the range of the variable values, corresponding to individual prospective sample classes. The algorithm you selected (creation of intervals with similar counts of items) was used to suggest the default values in this line, but you can freely change them. The values represent the upper inclusive values of the interval borders. The outer margins - the lower value of the first interval and the upper value of the last interval - are implied, respectively, by the minimum and maximum value of the variable. In our example the first interval starts and ends at the value 1, the second interval includes values 2 – 4, while the third interval obviously includes only the samples with *Moisture* value equal to 5.

The list in the lower part of this dialog shows the size of individual classes implied by the thresholds currently specified in the line above it. Note, however, that when you change these values, you must indicate to CanoDraw that you have finished the changes by clicking the *Refresh* button: only then is the content of the list changed. You can see that a remarkable equitability of counts of samples in three intervals was achieved. Confirm this classification by clicking the *OK* button. As in our previous example in this section, CanoDraw continues by displaying the *Manual Classification* dialog, where you can fine-tune the created classification manually. You do not need that, however, so close this dialog using the *OK* button.

After you return to the dialog named *Available Classifications of Samples*, you can see our new classification listed with the name *Classification from values of Moisture*. You can also see that it is not active – the older classification is still the one enabled for use in diagrams. You should therefore select the new classification in the list and then check the box *Use this classification in diagrams*. Then close the dialog using the *Close* button.

There are two additional changes you must make before you can create the diagram. Open the dialog with project-specific settings using the command *Project / Settings*. In the *Appearance* page, enable (check) the option named *Display species as symbols even in linear ordination methods*. While displaying species scores as symbols is generally an inappropriate choice with linear ordination methods, it is needed for the specific type of graph we want to create. Then switch to the *Contents* page of the same dialog and specify that you want to replace species symbols with pie symbols. To do so, check the *For species* option in the area named *Use Pies instead of Symbols*. There is another setting you need to consider, named *With slices based on*. Select the option *Values*, so that the size of slices representing individual sample classes will

report what fraction of total **abundances** of a species can be found in samples from the particular class. Selecting the other option (*Presences*) would result in displaying fractions of the total number of **occurrences**.

Close this dialog using the *OK* button and create the diagram (Figure 14-30) using the command *Create / Scatter Plots / Species*.

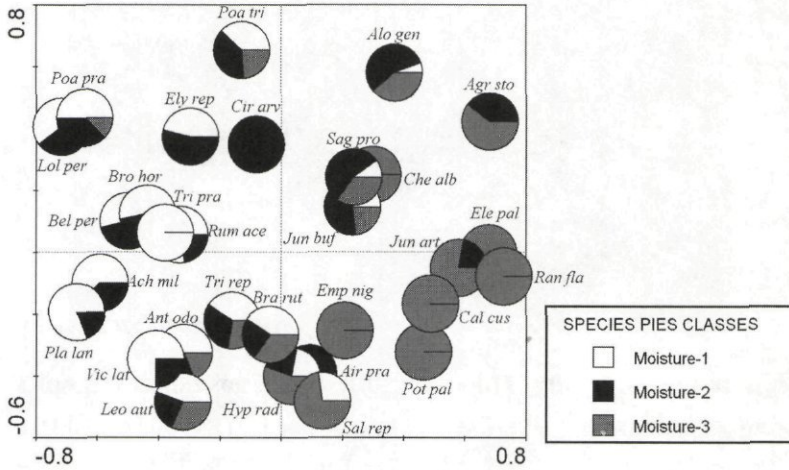


Figure 14-30 Pie symbols plot visualising distribution of species over classes of samples with different soil moisture

Note that the above diagram has been substantially adjusted by shifting the labels of the individual pie symbols. This presents a certain difficulty because several labels overlay the pies and it is difficult to select them before displacement. To facilitate easy shifting of the labels, all the pie symbols were selected (by clicking one of them and then executing the *Select Suchlike* command from the *Object* menu) and then locked (by clicking over one of the pies with the right mouse button and selecting the *Lock selected* command from the popup menu).

14.5 WEEDS

Canoco sample directory: *Samples\Unimodal\Weeds*

Description of the corresponding Canoco example starts on page: 245

This example demonstrates the use of spatial coordinates of samples in a Canoco project. We will show here the possibilities for visualising the spatial variation of community composition using CanoDraw. You should begin by creating a new CanoDraw project (*trend.cdw*) from the Canoco project *trend.con*. Next, you will visualise the position of samples in space. To do so, you will create a XY diagram based on the X and Y coordinates of the individual samples. Select the command *Create / Attribute Plots / XY(Z) Plot* and specify the options in the dialog as shown in Figure 14-31.

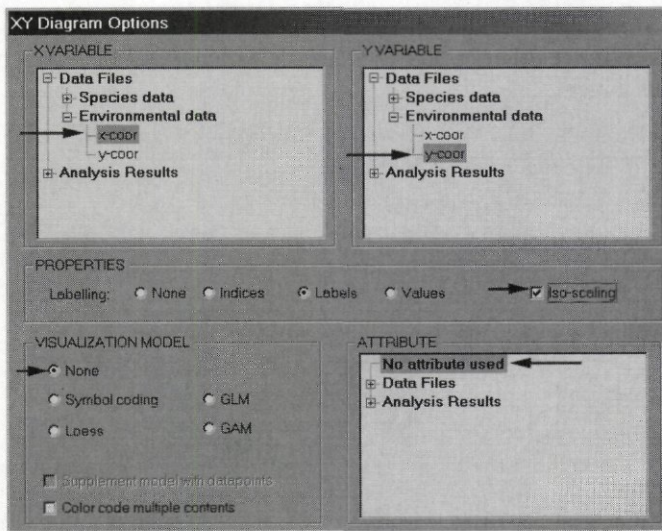


Figure 14-31 Creating a diagram with spatial positions of samples

The resulting graph is shown in Figure 14-32 (after deleting the sample labels).

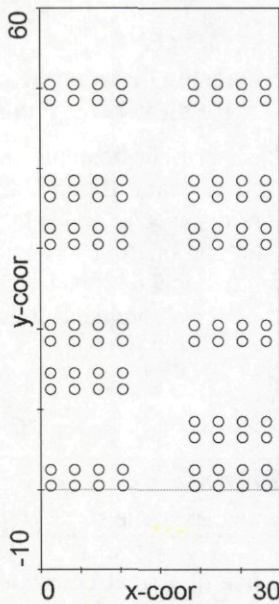


Figure 14-32 Spatial coordinates of samples

You do not need to stop here, however. You can visualise the sample scores on the first ordination axis within this diagram. If you code the position of samples on the first ordination axis by the size of sample symbols, you obtain the symbol attribute plot, like the one shown in Figure 14-33.

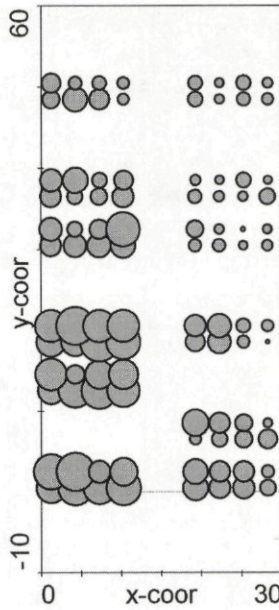


Figure 14-33 Symbol attribute plot showing the pattern of CCA Axis 1 coordinates in space

To create this symbol attribute plot, use again the *XY(Z) Plot* command and choose the same options in the upper part of the dialog, but make a different selection in the lower part of the dialog box (keeping the selection in lists for X and Y variables identical) – as illustrated in Figure 14-34.

Note that we had to create a completely new graph here. This is needed because we are changing the graph contents, not only its appearance. Using the *Create / Recreate graph* menu command would not help us, as this action only applies the current options to an a graph with already defined contents (see also section 11.8). If you make a mistake during new graph creation, for example you forget to enable iso-scaling, you may close it using the *File / Close* menu command (select *No* button when asked about saving the changes) and start again from scratch (with the *Create / Attributes Plots / XY(Z) Plots* command).

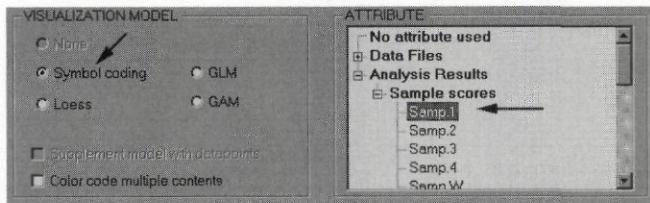


Figure 14-34 XY(Z) Plot settings for creating a symbol attribute plot

Alternatively, you can formalise the same pattern using a regression model – probably the best one for this purpose is the loess smoother. The required settings in the lower part of the *XY(Z) Plot Options* dialog are illustrated in Figure 14-35. The settings in the upper part are the same as illustrated in Figure 14-31.

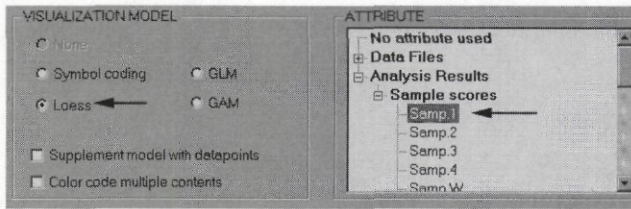


Figure 14-35 XY(Z) Plot settings for creating a contour-based attribute plot

The resulting diagram is shown in Figure 14-36.

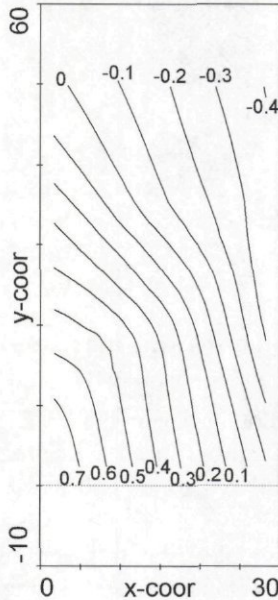


Figure 14-36 Contour-based attribute plot displaying the change of sample scores on the first CCA ordination axis throughout the sampling area

Similar graphs can be then made for the second CCA axis, as well as for the following, unconstrained ordination axes, if needed.

14.6 SEASHORE

Canoco sample directory: *Samples\Unimodal\Seashore*

Description of the corresponding Canoco example starts on page:: 247

In this project, temporal change in vegetation composition along transects running from the seashore is quantified and explained by the sample altitude, measured at the first sampling date. There are four transect lines and two sampling dates. We can visualise the temporal and spatial relatedness of individual samples in the ordination diagrams by displaying the ordering of samples in individual transect lines in CanoDraw series, using separate series for the two years of sampling.

Start with creating a CanoDraw project *uplift.cdw* from the Canoco project *uplift.con*. Then you must define a new series collection. To do so, select the menu command *Project / Define Series of / Samples*. CanoDraw displays a dialog titled *Series Collections for Samples* which allows you to manage existing series collections. We do not have any such collection yet, so

start with the button *Create* to establish a new one. CanoDraw asks first for the name of your new series collection. You can replace the default text with the title "*Line Transects in Two Years*" and click the *OK* button. Now you need to define each series separately: there are eight series in total (four transect lines times two sampling years). We will illustrate the necessary steps for the first series. The newly displayed dialog has two listboxes. The left one shows the existing series and if one of the series is selected (active) there, the right-hand list box displays the samples, which are members of that series. The ordering of samples is important here, too, and the right-hand list supports a change of sample order.

You must first click the *Add* button below the left-hand list (named *Defined series*). The new series, named *Series 1*, is placed into the list. Change its name to, say, *Trans 1 - 1978*, using the *Rename* button. Now you can add samples to this series using the other *Add* button - the one placed below the right-hand list. In the *Add Series Items* dialog, you need to select the sample names starting with *L1* and ending with *78*. As these samples are interspersed with the samples collected at the same transect in the year 1984 (labels ending with *84*), you must make your selection using the left mouse button combined with the *Ctrl* key. After clicking the *Add* button, you return to the previous dialog and the right-hand list contains the samples assigned to this series. They already have the appropriate order, but you can still exercise how to change their ordering, if that would be needed. This is illustrated in Figure 14-37. Let us assume you need to relocate the first sample (*L1-03-78*) to follow the sample *L1-06-78*. To do so, select the sample to be relocated, click it with left mouse button, keep the button pressed, and drag the mouse pointer downwards. CanoDraw indicates the position where the item being relocated would be placed if you release the mouse button. You do so in the position indicated in our snapshot image. Remember to return the relocated item back to its original position before continuing with this example.

You should define the remaining seven series (up to *Trans 4 - 1984*) using the similar procedure as described above. Note that the samples which were already assigned to previous series are still offered for inclusion in the new ones. Each sample can be member of more than one series in a series collection. When finished, close the dialog for editing series using the *OK* button. In the dialog for managing series collection, you should see the name of the just created series collection highlighted. Enable (check) the option named *This collection is used in the plots*. Only one series collection can be used (for the particular type of items) at the same time in diagrams.

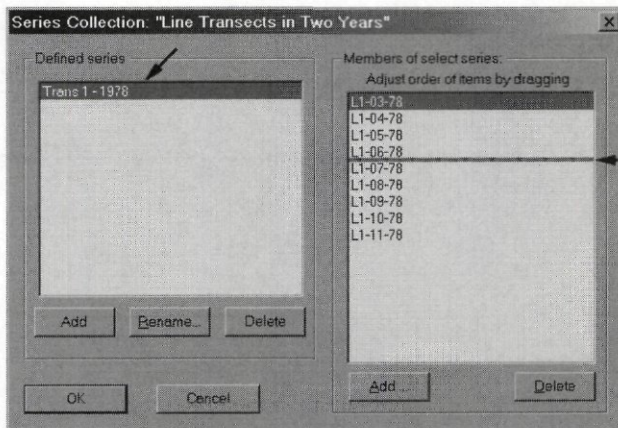


Figure 14-37 Defining new series and changing position of item in a series

We will start our data exploration by examining the changes in vegetation composition, as portrayed by the sample scores on the first ordination axis, with the changes in altitude. Here we will use the *Samp.1*, not *SamE.1* scores, because the former scores represent best the vegetation

composition. Before creating the graph, you must be sure that you will be able to recognise individual series lines. You must enable legend creation for this plot. From the *Project Settings* dialog, select the second page (named *Appearance*) and enable (check) the option *Insert legend into created diagrams*. You can see that the default layout of sections is vertical, while the items within sections are laid out horizontally. This might not be appropriate for our situation (with 8 items within the series legend section, each one with a rather long label). Therefore, keep the sections layout vertical, but allow two sections to be in one column, and specify a vertical arrangement of section items with up to eight items in a single column (see Figure 14-38).

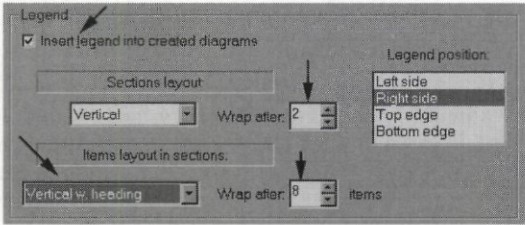


Figure 14-38 Legend options for a plot with series collections

To create the graph, select the *Create / Attribute Plots / XY(Z) Plot* command and set the options in the dialog as shown in Figure 14-39. The resulting graph is shown in Figure 14-40.

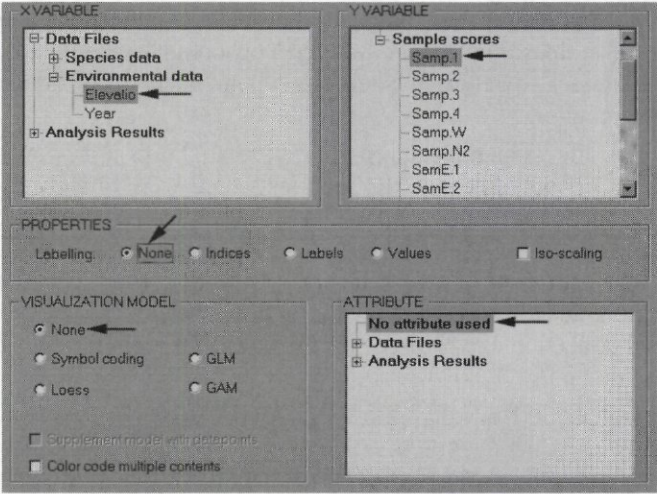


Figure 14-39 XY(Z) diagram options needed to produce the following graph.

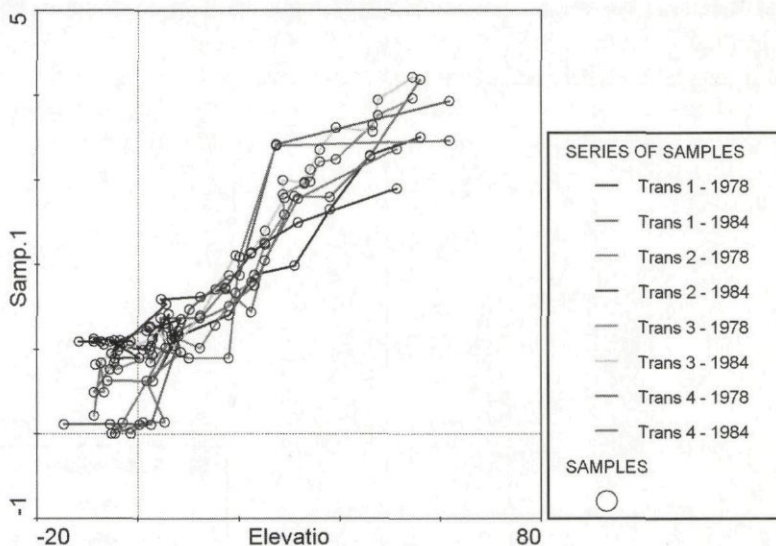


Figure 14-40 Change of CCA Axis 1 scores with site elevation, for individual series

However beautiful the graph is, there are actually four pairs of series, each pair representing the shift of vegetation composition on particular line transects between the two sampling years, and we need to concentrate our attention on these shifts. This is somewhat difficult in our graph, with the large overlap of lines. We probably need to plot separate graphs for each transect, sampled in the two years. There are several ways to achieve this in CanoDraw, probably the quickest one is to suppress plotting of samples except those samples on a particular transect line. As an example, you can show the compositional changes, as reflected by the scores on the first ordination axis, for the samples on the transect number 3. To do so, you need to suppress all the other samples. Select the command *Project / Suppress / Samples* and move all the samples not beginning with L3 to the right-hand list by selecting them in the left-hand listbox and clicking the *Select>>* button. Before you leave this dialog, it should look similar to Figure 14-41.

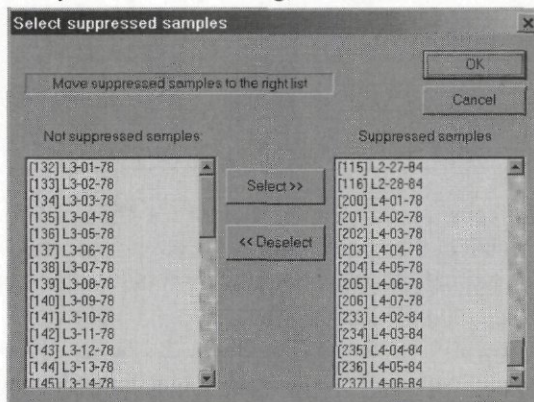


Figure 14-41 Suppression of samples not in transect line 3

After that, go to the already drawn diagram (illustrated in Figure 14-40), click on it with the right mouse button and select from the context menu the command *Recreate graph*. The figure clearly shows that a 1978 sample has a lower score on the first CCA axis than a corresponding

1984 sample at the same elevation, indicating a systematic shift in community composition in time.

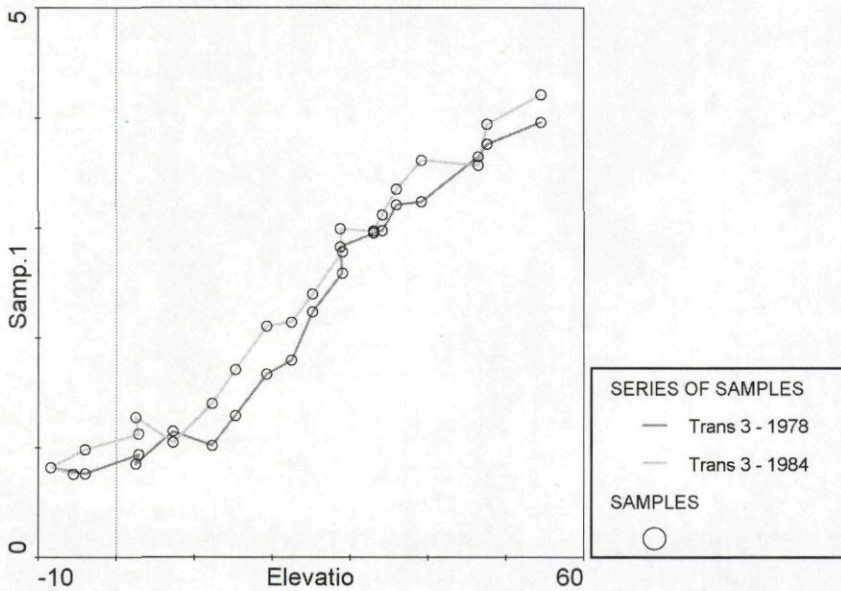


Figure 14-42 Final XY diagram

If you were interested in identity of individual samples, displayed in the Figure 14-42, you can either check the information displayed in the status bar of the graph window, as you move the mouse pointer over the plot, or you can add sample labels to this diagram. To do so, you should select the labelling options within the *XY Diagram Options* dialog (using the choices in the *PROPERTIES* area): the labelling choices specified in the *Appearance* page of the *Project Settings* dialog (see section 12.4.1.2) are **not** used by CanoDraw for the XY and XYZ diagrams.

14.7 DISEASES

Canoco sample directory: *Samples\Unimodal\Diseases*

Description of the corresponding Canoco example starts on page:: 255

This Canoco example focuses on the exploration of the effects of socio-economic status and other factors upon the incidence of various types of diseases. It is used here to demonstrate the creation of special types of ordination diagrams – the regression biplot and T-value biplot.

We will limit our discussion to the Canoco project named *fig1.con*, which was used to produce Figures 1 to 3 in the original paper (Ter Braak & Looman 1994). You must first create from it a CanoDraw project, named *fig1.cdw*. To create a diagram similar to Fig. 2 in the original paper (Unimodal Models booklet, p. 250), you must specify which variables are plotted. To do so, select the menu command *Project / Suppress / Env. variables* and select all the variables except the interactions between *SES* and *P1* to *P4* groups (i.e. except the variables with numbers 8 to 11). Move the selected variables into right-hand list box using the *Select* button (Figure 14-43).

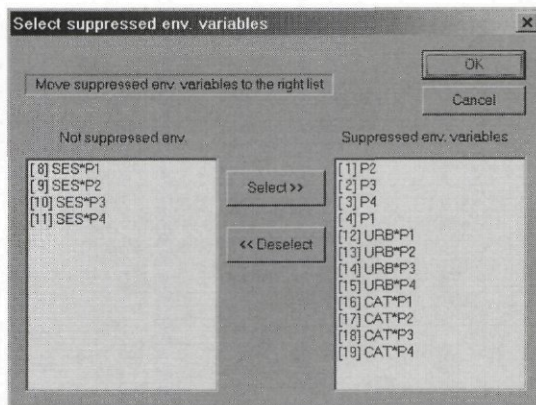


Figure 14-43 Select suppressed env. variables dialog box

Next, you can optionally suppress the legend for this diagram by selecting the *Project / Settings* command and un-checking the option *Insert legend into created diagrams* on the *Appearance* page. Finally, create the diagram using the *Create / Biplots and Joint Plots / T-values Biplot* command. Before the diagram is created, CanoDraw presents a dialog (Figure 14-44) where you can select plotting of Van Dobben circles for one of the available variables. This time, keep the default setting (*None*) in the upper part of this dialog and un-check (disable) the option in the lower part, named *Plot explanatory variables as symbols*. After you click the *OK* button, the diagram is created.

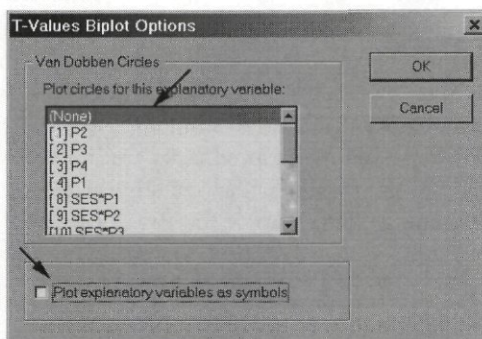


Figure 14-44 T-Values Biplot Options dialog

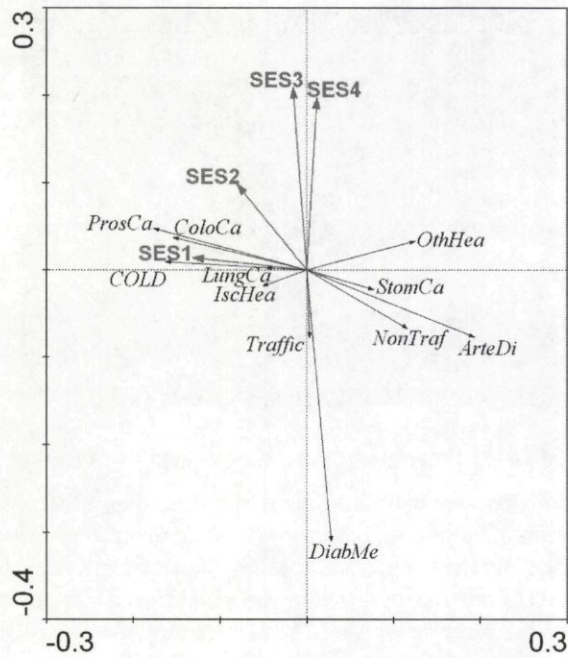


Figure 14-45 T-values Biplot diagram

Note that this diagram shows only part of the information displayed in Fig. 2 of the Ter Braak & Looman paper. Particularly, the arrows for *SESx* variables represent only the solid thick lines in the original figure, while the arrows for the response variables (disease types) represent only the dashed parts of the corresponding arrows in the original Fig. 2. Nevertheless, this is the part of information used to infer significant predictors for individual response variables (or, alternatively, to infer which response variables respond significantly to a change in the values of a particular predictor). The complementary part of the Fig. 2 (i.e. the full arrows of the response variables, including the solid parts, and the full arrows of the predictors, including the dashed segments) is a regression biplot and can be obtained with CanoDraw using the *Create / Biplots and Joint Plots / Regression biplot* command.

The T-values biplot illustrated in Figure 14-45 is also somewhat different from the diagram you can see in the CanoDraw workspace now. To obtain a comparable graph, you must go through the following steps:

1. You must change the range of axes, adding 0.1 unit on the left side of the horizontal axis (i.e. to change the horizontal range from $-0.2, +0.3$ to $-0.3, +0.3$). To do so, click the diagram using the right mouse button and select the *Range of axes* command from the popup menu. CanoDraw displays a dialog where you should change the option in its upper part from *Reset range of axes* to *Use ranges specified below* and modify the value in *From* field for the *Horizontal Axis* (Figure 14-46).

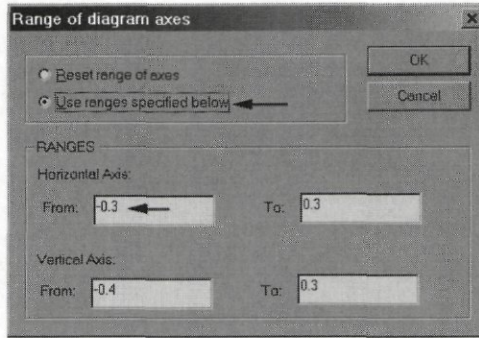


Figure 14-46 Changing range of horizontal axis for T-values biplot

2. Also, you must change the names of explanatory variables. CanoDraw uses labels indicating these are the interactions between the quantitative predictor *SES* and the four nominal variables *P1* to *P4*. To modify the text of a label, select one label, right-click it and select the *Change text* command from the pop-up menu. In Ter Braak & Looman (1994), the arrows are labelled *SES₁* to *SES₄*. Note, however, that you cannot format lower indices within CanoDraw labels. To achieve such effect, you would need to either add another label for the digit, using a smaller font and aligning it properly with the *SES* label, or you must export the diagram in Adobe Illustrator format and modify the label formatting there.

Ter Braak & Looman paper also shows so-called Van Dobben circles, e.g. in Fig. 3. The pair of Van Dobben circles, helping to identify response variables responding significantly to the particular predictor either in a positive or a negative way, can be easily obtained with CanoDraw, using the settings in the upper part of the *T-Values Biplot Options* dialog, which was illustrated in Figure 14-44.

14.8 PRC_SIM

Canoco sample directory: *Samples\Permutio\Prc_sim*

Description of the corresponding Canoco example starts on page:: 287

This example illustrates how to create diagrams with principal response curves (PRC), described in Van den Brink & Ter Braak (1999). For additional details you can also refer to section 12.4.8.3 of this manual.

We start by creating a CanoDraw project, using the analysis defined in the *prc.con* Canoco project file. In this new CanoDraw project (named *prc.cdw*), we begin by importing the PRC scores. To do so, select the *Project / Import variables / Setup PRC scores* menu command. A dialog box appears where you select the environmental variables defining the interaction terms between the experimental treatments levels (except the control treatment) and the individual time points in which the community composition was measured. You should also specify the number of principal curves to import. In our analysis, we have just one principal curve available. Finally, you should specify in this dialog where the analysis log for the original Canoco project is located. In our case, the file has name *prc.log* and it is located in the same place as the Canoco project file was. Figure 14-47 illustrates the final appearance of the dialog box.

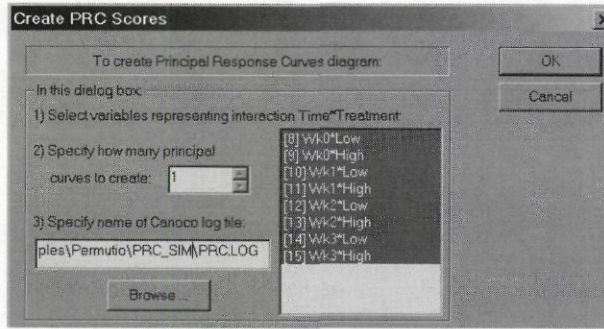


Figure 14-47 Create PRC Scores dialog box

After you click the *OK* button, CanoDraw imports the scores which can be used to plot the first response curve and stores them in the *PRC1* variable, among the *Imported variables*.

To use this variable, however, you must perform several additional steps. First, the calculated *PRC1* variable represents only the vertical coordinates of the PRC diagram. You must define the horizontal scores, too. To do this, the easiest method is to copy the labels of environmental variables, together with their indices to the Clipboard and paste them into an empty spreadsheet, assuming the labels are informative enough to identify the associated sampling time. To copy the labels of environmental variables, display the *CanoDraw Project Details* window (using the *View / Project Details* menu command) and navigate there into *Project Results / Labels* folder. Then you must right-click the item named *EnvV.Labels*. The *Variable Summary* floating window is shown, with individual labels being listed. You should click the *Copy* button and then close this window using the *OK* button. This procedure is illustrated in Figure 14-48.

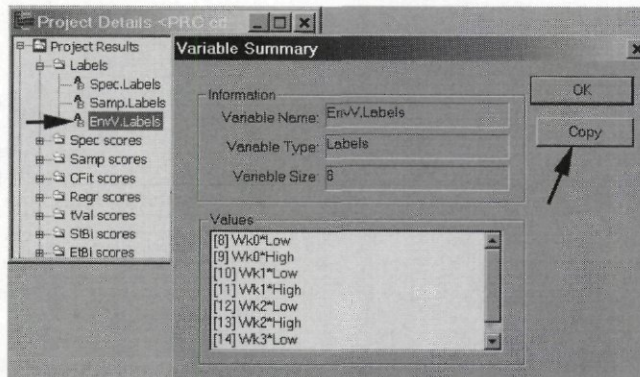


Figure 14-48 Copying labels of environmental variables onto Clipboard

After that, you must open a spreadsheet program and select the *Paste* command to include the copied data into the spreadsheet. The first column contains the indices of environmental variables, the second contains their labels. Because we cannot import back the labels, it is suggested you write the horizontal scores of points for the PRC diagram (the quantitative time values) into the third column, based on the guidance provided by the labels in the second column, and then you delete the no-longer-needed second column. The new column should contain the quantitative description of the horizontal (time) axis. The Figure 14-49 illustrates how the data looks in your spreadsheet after you added the quantitative time values. Note that the column with labels was not deleted here. It was relocated beyond the *Time* column and will not be copied back to the Clipboard.

	A	B	C
1	Indices	Time	EnvV.Labels
2	8	0	Wk0*Low
3	9	0	Wk0*High
4	10	1	Wk1*Low
5	11	1	Wk1*High
6	12	2	Wk2*Low
7	13	2	Wk2*High
8	14	3	Wk3*Low
9	15	3	Wk3*High
10			

Figure 14-49 Adding *Time* variable in spreadsheet program

Now you should select the two columns (the indices and the quantitative time values, the latter called *Time*), copy them to the Clipboard (using for example the *Ctrl-C* shortcut in Microsoft Excel®), and import them back into CanoDraw project. To do so, select the *Project / Import variables / From Clipboard* menu command. The displayed dialog box already lists the two variables found, but you must also change the settings in the left part of the dialog. The individual rows in the spreadsheet correspond to environmental variables, not to samples (see Figure 14-50). Click the *Import* button then.

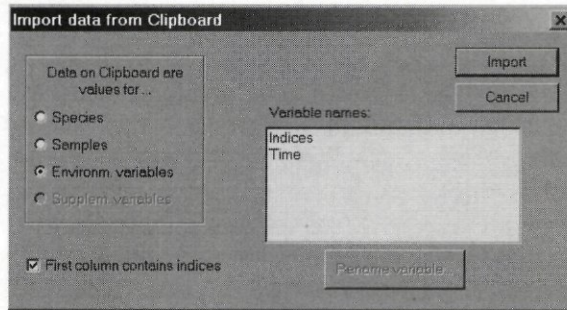


Figure 14-50 Importing the *Time* variable back into CanoDraw

The next step needed for plotting the imported response curve is to define the series of the environmental variables (actually series of the interactions between time and treatment), which would connect, in the temporal order, the interaction terms corresponding to a particular treatment level. That means that each level of the experimental treatment will be represented by a separate series in the series collection. Because we want to plot not only the series lines, but also the symbols for individual dummy environmental variables (similarly to Figure 8-3 in this manual, p. 288), we will start with classifying our environmental (explanatory) variables. Select the *Project / Classify / Env. variables* menu command. In the dialog box named *Available Classifications of Env. Variables* click the *New select* button. Another dialog appears, with the name *Manual Classification*. Click the *Add class* button in this dialog and specify the name *Low* for this class. Leave the box *Continue with Class Members dialog* checked, so that you can specify this class' members after you click the *OK* button in this dialog box. In the *Class Members* dialog, select the items in the left-hand list, containing the word *Low*, and transfer them to the right-hand list using the *Select>>* button. To select non-contiguous items, you must combine the left mouse click with pressing the *Ctrl* key on the keyboard. Close this dialog box with the *OK* button and proceed similarly (by adding a new class and specifying its four members) with the *High* class. At the end, the *Manual Classification* dialog should look as illustrated in Figure 14-51. Close the dialog with *OK* button.

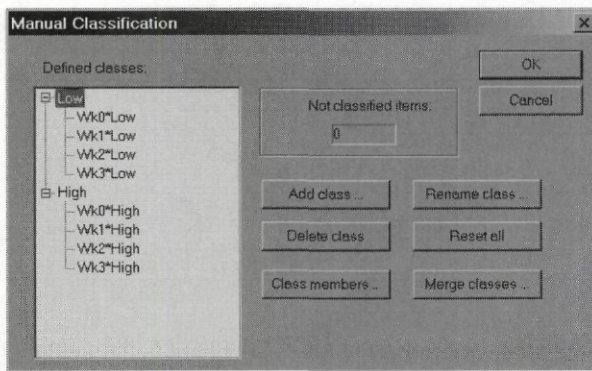


Figure 14-51 Classifying environmental variables by treatment

In the dialog box, to which you return, rename the classification from *New Classification* to *Treatments* and check the box below the list, labelled *Use this classification in diagrams*. Then you can close the dialog using the *Close* button.

We can use the existing classification to simplify the creation of the series collection we need to plot the curves. To do so, select the *Project / Define Series of / Env. variables* menu command. In the dialog named *Series Collections for Env. Variables* click the *From class* button. Confirm the use of *Treatments* classification in the subsequently displayed dialog by clicking the *OK* button. CanoDraw defines the new series collection, based on the selected classification of environmental variables and displays it for optional further editing. The important task is to check that the items in both series defining this series collection are in the appropriate temporal order. This is the case in our sample data, but might not be so in other projects. You can change the ordering of series items in the right-hand list using the mouse pointer (see section 12.4.5). After you close this dialog box using the *OK* button, you return to the original *Series Collections for Env. Variables* dialog box. Check there the box labelled *This collection is used in the plots* and leave this dialog using the *Close* button.

After you have performed all these preliminary steps, you can plot the PRC by creating an XY diagram with the imported *Time* variable on the horizontal axis and the first PRC scores (imported variable *PRC1*) on the vertical axis. Before you do that, change the project options so that the diagram will be supplemented with a legend. Select the *Project / Settings* command and in the *Appearance* property page enable the legend creation by checking the *Insert legend into created diagrams* box. This and the other choices to be made in the lower part of this dialog page are illustrated in Figure 14-52.

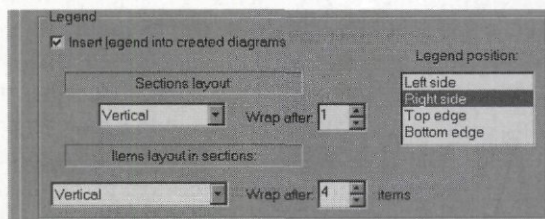


Figure 14-52 Selecting legend options

Then select the *Create / Attribute Plots / XY(Z) Plot* menu command and make the choices there as shown in Figure 14-53.

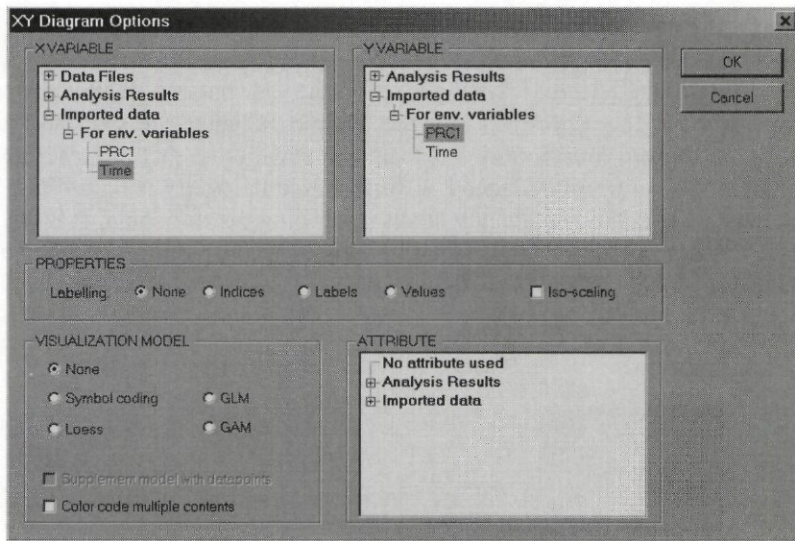


Figure 14-53 Creating the XY diagram with first PRC

The resulting diagram is illustrated in Figure 14-54.

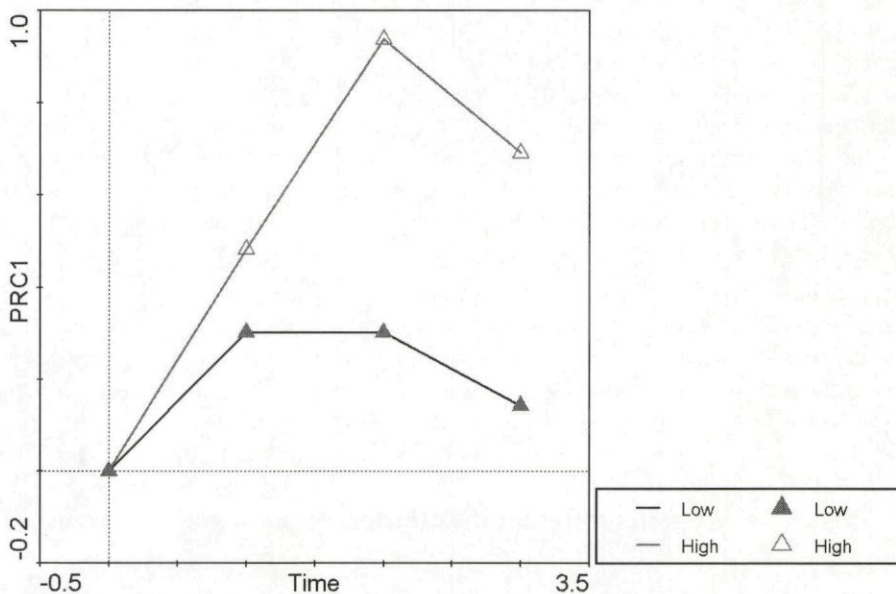


Figure 14-54 PRC diagram with PRC1

You can also attempt to simulate the one-dimensional diagram with species scores along the first RDA axis, as shown in the Figure 8-3. To create it, you should first make sure that the created diagram is not un-necessarily wide, because the horizontal direction brings no information and it only provides space for species labels. Therefore, you should start with the *View / Diagram Settings* menu command and check the *Adjust graph aspect ration by banking to 45 degrees* option on the dialog page named *Properties 1*. Close this dialog box and select the *Create / Attribute Plots / XY(Z) Plot* menu command to display the *XY Diagram Options* dialog. The coordinates for the horizontal axis should be constant and you achieve this in the *X VARIABLE* list-box by expanding the *Analysis Results / Species scores* folder and selecting

the *Constant* item near the end of the list for this folder. In the *Y VARIABLE* list, select the *Spec.1* variable from the *Analysis Results / Species scores* folder. Make sure that the *Labelling* area has the *Labels* option selected. After you click the *OK* button, a small dialog is shown, requesting you to specify the aspect ratio for the diagram. Change the suggested value to **4.0**. This will result in a diagram four times taller than it is wide. Note that the resulting diagram is far from perfect: we do not need the legend, so you should delete it (using the *Del* key after its selection). Likewise, you should delete any labels describing the horizontal axis (there are three labels). The species labels in this project can be simplified to their look in Figure 8-3, their positions re-adjusted to prevent overlap, and their font size somewhat increased. The resulting diagram is shown in Figure 14-55.

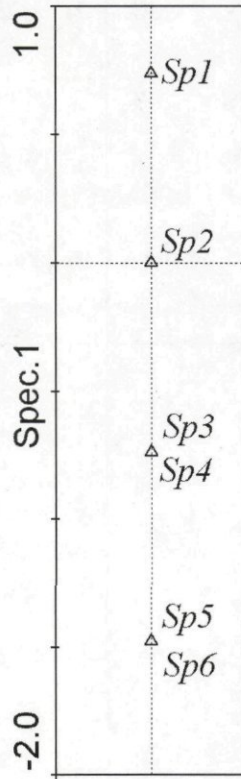


Figure 14-55 Species scores for the principal response curve diagram

15. References

- Aitchison, J. (1982) The statistical analysis of compositional data. *J. R. Statist. Soc. B*, 44: 139-177.
- Aitchison, J. (1984a) The statistical analysis of geochemical compositions. *Mathematical Geology*, 16: 531-564.
- Aitchison, J. (1984b) Reducing the dimensionality of compositional data sets. *Mathematical Geology*, 16: 617-635.
- Aitchison, J. (1986) The statistical analysis of compositional data. Chapman and Hall, London.
- Aitchison, J. (1990) Relative variation diagrams for describing patterns of compositional variability. *Mathematical Geology*, 22: 487-511.
- Anderson, E. et al. (1999) LAPACK Users' Guide. Third Edition. SIAM, Philadelphia.
- Anderson, J.A. (1984) Regression and ordered categorical variables. *J. R. Statist. Soc. B*, 46: 1-30.
- Anderson, M.J. (2001) Permutation tests for univariate or multivariate analysis of variance and regression. *Canadian Journal of Fisheries and Aquatic Sciences*, 58: 626-639.
- Anderson, M.J. & Legendre, P. (1999) An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation*, 62: 271-303.
- Anderson, M.J. & Robinson, J. (2001) Permutation tests for linear models. *Australian & New Zealand Journal of Statistics*, 43: 75-88.
- Anderson, M.J. & Ter Braak, C.J.F. (2002) Permutation tests for multi-factorial analysis of variance. *Journal of Statistical Computation and Simulation* in press.
- Baar, J. & Ter Braak, C.J.F. (1996) Ectomycorrhizal sporocarp occurrence as affected by manipulation of litter and humus layers in Scots pine stands of different age. *Applied Soil Ecology*, 4: 61-73.
- Batterink, M. & Wijffels, G. (1983) Een vergelijkend vegetatiekundig onderzoek naar de typologie en invloeden van het beheer van 1973 tot 1982 in de Duinweilanden op Terschelling. Report Agricultural University, Department of Vegetation Science, Plant Ecology and Weed Science. Wageningen. 101 pp.
- Besag, J. Clifford, P. (1989) Generalized Monte Carlo significance tests. *Biometrika*, 76: 633-642.
- Birks, H.J.B., Peglar, S.M. & Austin, H.A. (1996): An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986-1993. *Abstracta Botanica*, 29: 17-36.
- Borcard, D., Legendre, P. & Drapeau, P. (1992) Partialling out the spatial component of ecological variation. *Ecology*, 73: 1045-1055.
- Both, J.C. & Van Wirdum, G. (1981) Waterhuishouding, bodem en vegetatie van enkele Gelderse natuurgebieden, RIN-report nr. 81/18. Rijksinstituut voor Natuurbeheer, Leersum.
- Cade, B.S. & Richards, J.D. (1996) Permutation tests for least absolute deviation regression. *Biometrics*, 52: 886-902.
- Carnes, B.A. & Slade, N.A. (1982) Some comments on niche analysis in canonical space. *Ecology*, 63: 888-893.

- Carpenter, S.R., Frost, T.M. & Heisey, D.K.T.K. (1989) Randomized intervention analysis and the interpretation of whole-ecosystem experiments. *Ecology*, 70: 1142-1152.
- Chambers, J.M. & Hastie, T.J. [eds] (1992) *Statistical Models in S*. Wadsworth & Brooks, Pacific Grove, CA.
- Chessel, D., Lebreton, J.D. & Yoccoz, N. (1987) Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie. *Revue de Statistique Appliquée*, 4: 55-72.
- Chessel, D., Lebreton, J.D. & Prodon, R. (1982) Mesures symétriques d'amplitude d'habitat & de diversité intra-échantillon dans un tableau espèces-relevés: cas d'un gradient simple. *C.R. Acad. Sc. Paris, Series III*, 295: 83-88.
- Cleveland, W.S. & Devlin, S.J. (1988) Locally-weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83: 596-610.
- Cleveland, W.S. (1994) *The Elements of Graphing Data*. Hobart Press, Summit, New Jersey.
- Cleveland, W.S. & Grosse, E. (1991) Computational methods for local regression. *Statistics and Computing*, 1: 47-62
- Collins, M.F. (1987) A permutation test for planar regression. *Australian Journal of Statistics*, 29: 303-308.
- Corsten, L.C.A. & Gabriel, K.R. (1976) Graphical exploration in comparing variance matrices. *Biometrics*, 32: 851 - 863.
- Cox, D.R. (1958) *Planning of experiments*. J. Wiley, New York, USA.
- Cramer, W. & Hytteborn, H. (1987) The separation of fluctuation and long-term change in vegetation dynamics of a rising sea-shore. *Vegetatio*, 69: 155-167.
- Davies, P.T. & Tso, M.K.S. (1982) Procedures for reduced-rank regression. *Applied Statistics*, 31: 244-244.
- De Lange, L. (1972) *An ecological study of ditch vegetation in the Netherlands*. Dissertation. University of Amsterdam, Amsterdam, The Netherlands.
- Dolédec, S. & Chessel, D. (1994) Co-inertia analysis: an alternative method for studying species-environment relationship. *Freshwater Biol.*, 31: 277-294.
- Dueser, R.D. & Shugart, H.H. (1978) Microhabitats in a forest-floor small-mammal fauna. *Ecology*, 59: 89-98.
- Dueser, R.D. & Shugart, H.H. (1979) Niche pattern in a forest-floor small-mammal fauna. *Ecology*, 60: 108-118.
- Dueser, R.D. & Shugart, H.H. (1982) Reply to comments by Van Horne and Ford and by Carnes and Slade. *Ecology*, 63: 1174-1175.
- Edgington, E.S. (1995) *Randomization tests*. 3rd ed., M. Dekker, New York.
- Escoufier, Y., Roberts, P. (1979) Choosing variables and metrics by optimizing the RV-coefficient. In: Rustagi J.S.[ed.], *Optimizing methods in Statistics*. Academic Press, New York, pp. 205-219.
- Everts, J.W. (1990) Sensitive indicators of side-effects of pesticides on the epigeal fauna of arable land. PhD thesis, Agricultural University, Wageningen.
- Everts, J.W., Aukema, B., Hengeveld, R. & Koeman J.H. (1989) Side-effects of pesticides on ground-dwelling predatory arthropods in arable ecosystems. *Environmental Pollution*, 9: 203-225.
- Farjon, A. & Wiertz, J. (1989) Milieu- en vegetatieverandering in het schraalland van Koolmansdijk (gemeente Lichtenvoorde) 1952-1988. RIN-rapport 89/18. Rijksinstituut voor Natuurbeheer (Leersum).

- Feoli, E. & Orlóci, L. (1979) Analysis of concentration and detection of underlying factors in structured tables. *Vegetatio*, 40: 49-54.
- Fisher, N.I. & Hall, P. (1990) On bootstrap hypothesis testing. *Austral. J. Statist.*, 32: 177-190.
- Freedman, D.A. & Lane, D. (1983) A nonstochastic interpretation of reported significance levels. *Journal of Business & Economic Statistics*, 1: 292-298.
- Fricke, G. & Steubing, L. (1984) Die Verbreitung von Makrophyten und Mikrophyten in Hartwasser-Zuflüsse des Ederstausees. *Archiv für Hydrobiologie*, 101: 361-372.
- Gabriel, K.R. (1981) Biplot display of multivariate matrices for inspection of data and diagnosis. In: Barnett, V.[ed.], *Interpreting multivariate data*. J. Wiley, New York, pp. 147-173.
- Gauch, H.G. (1982) *Multivariate analysis in community ecology*. Cambridge University Press, Cambridge. 298 pp.
- Gifi, A. (1990) *Nonlinear multivariate analysis*. Wiley, New York.
- Gittins, R. (1985) *Canonical analysis. A review with applications in ecology*. Springer-Verlag, Berlin.
- Goodman, L.A. (1981) Association models and canonical correlation in the analysis of cross-classifications having ordered categories. *J. Amer. Statist. Ass.*, 76: 320-334.
- Gordon, A.D. (1981) *Classification. Methods for exploratory analysis of multivariate data*. Chapman and Hall, London.
- Gower, J.C. (1966) Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53: 325-338.
- Gower, J.C. & Hand, D.J. (1996) *Biplots*. Chapman and Hall, London, 224 pp.
- Green, R.H. (1971) A multivariate statistical approach to the Hutchinsonian niche: bivalve molluscs of central Canada. *Ecology*, 52: 543-556.
- Greenacre, M.J. (1984) *Theory and applications of correspondence analysis*. Academic Press, London.
- Hall, P. & Titterton, 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.
- Hastie, T.J. & Tibshirani, R.J. (1990) *Generalized Additive Models*. Chapman and Hall, London.
- Higler, L.W.G. & Repko, F. (1981) The effects of pollution in the drainage area of a Dutch lowland stream on fish and macro-invertebrates. *Verh. Int. Verein. Limnol.*, 21: 1077-1082.
- 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. Section of Ecology and Systematics, Cornell University, Ithaca, New York.
- Hill, M.O. & Gauch, H.G. (1980) Detrended correspondence analysis, an improved ordination technique. *Vegetatio*, 42: 47-58.
- Hope, A.C.A. (1968) A simplified Monte Carlo significance test procedure. *J. Roy. Statist. Soc. Series B*, 30: 582-598.
- Ihm, P. & Van Groenewoud, H. (1984) Correspondence analysis and Gaussian ordination. *COMPSTAT Lectures*, 3: 5-60.
- Israëls, A.Z. (1984) Redundancy analysis for qualitative variables. *Psychometrika*, 49: 331-346.

- Jongman, R.H.G., Ter Braak, C.J.F., Van Tongeren, O.F.R. (1987) Data analysis in community and landscape ecology. Pudoc, Wageningen, The Netherlands. Reissued in 1995 by Cambridge University Press, Cambridge, xix + 299 pp.
- Juggins, S. (1998) WINTRAN. Department of Geography, University of Newcastle, UK. See URL: <http://www.staff.ncl.ac.uk/stephen.juggins>
- Juggins, S. & Ter Braak, C.J.F. (1993) Calibrate - a program for species-environment calibration by [weighted-averaging] partial least squares regression. Environmental Change Research Centre, University College London, London.
- Kendall, M.G. & Stuart, A. (1973) The advanced theory of statistics. Vol. II. Inference and relationship. Griffin, London.
- Kenkel, N.C. & Orłóci, L. (1986) Applying metric and nonmetric multidimensional scaling to ecological studies: some new results. *Ecology*, 67: 919-928.
- Kennedy, P.E. & Cade, B.S. (1996) Randomization tests for multiple regression. *Communications in statistics. Simulation and computation*, 25: 923-936.
- Knox, R.G. (1989) Effects of detrending and rescaling on correspondence analysis: solution stability and accuracy. *Vegetatio*, 83: 129-136.
- Kooijman, S.A.L.M. (1977) Species abundance with optimum relations to environmental factors. *Annals of System Research*, 6: 123-138.
- Krzanowski, W.J. (1988) Principles of Multivariate Analysis. Clarendon Press, Oxford.
- Kunst, A.E., Looman, C.W.N. & Mackenbach, J.P. (1990) Socio-economic mortality differences in the Netherlands in 1950-1984: a regional study of cause-specific mortality. *Soc. Sci. Med.*, 31: 141-152.
- Lebreton, J.D., Chessel, D., Prodon, R. & Yoccoz, N. (1988) L'analyse des relations especes-milieu par l'analyse canonique des correspondances. I. Variables de milieu quantitatives. II. Variables de milieu qualitatives. *Acta Oecologia, Oecologia generalis*, 9: 53-67 and 9: 137-151.
- Legendre, P., Oden, N.L., Sokal, R.R., Vaudor, A. & Kim, J. (1990) Approximate analysis of variance of spatially autocorrelated regional data. *Journal of Classification*, 7: 53-75.
- Legendre, P. & Legendre, L. (1998) Numerical ecology. Second English edition. Elsevier, Oxford.
- Legendre, P. & Anderson, M.J. (1999) Distance-based redundancy analysis: testing multi-species responses in multi-factorial ecological experiments. - *Ecological Monographs*, 69: 1-24
- Legendre, P. & Gallagher, E.D. (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129: 271-280
- Line, J.M., Ter Braak, C.J.F. & Birks, H.J.B. (1994) WACALIB version 3.3 - a computer program to reconstruct environmental variables from fossil assemblages by weighted averaging and to derive sample-specific errors of prediction. *Journal of Paleolimnology*, 10: 147-152.
- Manger, R. & Schouten, A.J. (1989) Onderzoek naar de effecten van bekalking op de nematodenfauna van drie bosopstanden in Boswachterij St. Antonis (Peel-regio), rapport 718823001. RIVM, Bilthoven.
- Manly, B.F.J. (1983) Analysis of polymorphic variation in different types of habitat. *Biometrics*, 39: 13-27.
- Manly, B.F.J. (1991) Randomization and Monte Carlo methods in biology. Chapman and Hall, London, 281 pp.

- Manly, B.F.J. (1997) Randomization, Bootstrap and Monte Carlo methods in biology. Chapman and Hall, London.
- Mardia, K.V., Kent, J.T. & Bibby, J.M. (1979) Multivariate analysis. Academic Press, London, 521 pp.
- McArdle, B.H. & Anderson, M.J. (2001) Fitting multivariate models to community data: a comment on distance-based redundancy analysis. *Ecology*, 82: 290-297
- McCullagh, P. & Nelder, J.A. (1989) Generalized linear models. Second Edition. Chapman and Hall, London.
- McCune, B. (1997) Influence of noisy environmental data on canonical correspondence analysis. *Ecology*, 78: 2617-2623.
- Miller, A.J. (1990) Subset Selection in Regression. - Chapman and Hall, London, 229 pp.
- Minchin, P. (1987) An evaluation of the relative robustness of techniques for ecological ordination. *Vegetatio*, 69: 89-107.
- Montgomery, D.C. & Peck, E.A. (1982) Introduction to linear regression analysis. Wiley, New York, 504 pp.
- Noy-Meir, I. (1973) Data transformation in ecological ordination. I. Some advantages of non-centering. *J. Ecol.*, 61: 329-341.
- Noy-Meir, I., Walker, D. & Williams, W.T. (1975) Data transformation in ecological ordination. II. On the meaning of data standardization. *J. Ecol.*, 63: 779-800.
- Økland, R.H. & Eilertsen, O. (1994) Canonical correspondence analysis with variation partitioning: some comments and an application. *J. Veg. Sci.*, 5: 117-126.
- Oksanen, J. & Minchin, P.R. (1997) Instability of ordination results under changes in input data order: explanations and remedies. *Journal of Vegetation Science*, 8: 447-454.
- Opdam, P.F.M., Kalkhoven, J.T.R. & Phillippona, J. (1984) Verband tussen Broedvogelgemeenschappen en Begroeiing in een Landschap bij Amerongen. Pudoc, Wageningen.
- Post, B.J. (1986) Factors of influence on the development of an arable weed vegetation. Proc. EWRS Symposium 1986, Economic Weed Control: 317-325.
- Prentice, I.C. (1980) Multidimensional scaling as a research tool in quaternary palynology: a review of theory and methods. *Rev. Palaeobot. Palynol.*, 31: 71-104.
- Purata, S.E. (1986) Studies on secondary succession in Mexican tropical rain forest. Acta Univ. Ups. Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science, 19. Almqvist and Wiksell International, Stockholm.
- Rao, C.R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhya A*, 26: 329-358.
- Rao, C.R. (1973) Linear statistical inference and its application. 2nd edition. J. Wiley, New York. 625 pp.
- Robinson, P.M. (1973) Generalized canonical analysis for time series. *Journal of Multivariate Analysis*, 3: 141-160.
- Sabatier, R., Lebreton J.D. & Chessel D. (1989) Multivariate analysis of composition data accompanied by qualitative variables describing a structure. In: Coppi R. & Bolasco S. [eds.], *Multiway data tables*. North-Holland, Amsterdam, pp. 341-352.
- Seber, G.A.F. (1977) Linear regression analysis. Wiley, New York.
- Snoeijs, P.J.M. & Prentice, I.C. (1989) Effects of cooling water discharge on the structure and dynamics of epilithic algal communities in the northern Baltic. *Hydrobiologia*, 184: 99-123.

- Stapel, M. & Ter Braak, C.J.F. (1994) Randomization and bootstrap tests in factorial experiments: Does analysis follow from design? Dutch-German Biometrics Meeting, 15-18 May 1994, Munster.
- Stewart-Oaten, A., Murdoch, W.W. & Parker, K.P. (1986) Environmental impact assessment: "pseudoreplication" in time? *Ecology*, 67: 929-940.
- Tamm, C.O., Nilsson, A. & Wiklander, G. (1974) The optimum nutrition experiment Lisselbo. A brief description of an experiment in a young stand of Scots Pine (*Pinus sylvestris* L.). Report Royal College of Forestry, 18. Stockholm. 25 pp.
- Tausch, R.J., Charlet, D.A., Weixelman, D.A. & Zamudio, D.C. (1995) Patterns of ordination and classification instability resulting from changes in input data order. *J. of Vegetation Science*, 6: 897-902.
- Ter Braak, C.J.F. (1980) Binary mosaics and point quadrat sampling in ecology. MSc thesis, Newcastle upon Tyne, UK.
- Ter Braak, C.J.F. (1983) Principal components biplots and alpha and beta diversity. *Ecology*, 64: 454-462.
- Ter Braak, C.J.F. (1985) Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics*, 41: 859-873.
- Ter Braak, C.J.F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67: 1167-1179.
- Ter Braak, C.J.F. (1987) CANOCO - A FORTRAN program for canonical community ordination by [partial][detrended][canonical] correspondence analysis, principal components analysis and redundancy analysis (version 2.1). DLO-Agricultural Mathematics Group, Wageningen.
- Ter Braak, C.J.F. (1990) Interpreting canonical correlation analysis through biplots of structural correlations and weights. *Psychometrika*, 55: 519-531.
- Ter Braak, C.J.F. (1992) Permutation versus bootstrap significance tests in multiple regression and ANOVA. In: Jöckel, K.H., Roethe, G. & Sendler W.[eds.], *Bootstrapping and related techniques*. Springer Verlag, Berlin, pp.79-85.
- Ter Braak, C.J.F. (1994) Canonical community ordination. Part I: Basic theory and linear methods. *Ecoscience*, 1: 127-140.
- Ter Braak, C.J.F. (1995) Non-linear methods for multivariate statistical calibration and their use in palaeoecology: a comparison of inverse (k-Nearest Neighbours, PLS and WA-PLS) and classical approaches. *Chemometrics Intell. Lab. Syst.*, 28: 165-180.
- Ter Braak, C.J.F. (1996) Unimodal models to relate species to environment. DLO-Agricultural Mathematics Group, Wageningen.
- Ter Braak, C.J.F. (1997) Review of "Biplots" by Gower, J.C. & Hand, D.J. (1996), Chapman and Hall, London. *Psychometrika*, 62: : 457-459.
- Ter Braak, C.J.F. & Barendregt, L.G. (1986) Weighted averaging of species indicator values: its efficiency in environmental calibration. *Math. Biosci.*, 78: 57-72.
- Ter Braak, C.J.F. & De Jong, S. (1998) The objective function of partial least squares regression. *Journal of Chemometrics*, 12: 41-54.
- Ter Braak, C.J.F. & Juggins S. (1993) Weighted averaging partial least squares regression (WA-PLS) an improved method for reconstructing environmental variables from species assemblages. *Hydrobiologia*, 269: 485-502.
- Ter Braak, C.J.F. & Looman, C.W.N. (1986) Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio*, 65: 3-11

- Ter Braak, C.J.F. & Looman, C.W.N. (1994) Biplots in reduced-rank regression. *Biom. J.*, 36: 983-1003.
- Ter Braak, C.J.F. & Prentice, I.C. (1988) A theory of gradient analysis. *Advances in ecological research*, 18: 271-317.
- Ter Braak, C.J.F. & 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. & Verdonschot, P.F.M. (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences*, 57: 255-289.
- Ter Braak, C.J.F. & Wiertz, J. (1994) On the statistical analysis of vegetation change: a wetland affected by water extraction and soil acidification. *Journal of Vegetation Science*, 5: 361-372.
- Torgerson, W.S. (1958) *Theory and methods of scaling*. Wiley, New York, 460 pp.
- Underwood, A.J. (1992) Beyond BACI: the detection of environmental impacts on populations in the real, but variable, world. *J. Exp. Mar. Biol. Ecol.*, 161: 145-178.
- Underwood, A.J. (1994) On beyond BACI: sampling designs that might reliably detect environmental disturbances. *Ecological Applications*, 4: 3-15.
- Underwood, A.J. (1996) *Experiments in ecology. Their logical design and interpretation using analysis of variance*. Cambridge University Press, Cambridge, 528 pp.
- Van Dam, H., Suurmond, G. & Ter Braak, C.J.F. (1981) Impact of acidification on diatoms and chemistry of Dutch moorland pools. *Hydrobiologia*, 83: 425-459.
- Van den Brink, P.J. & Ter Braak, C.J.F. (1997) Ordination of responses to toxic stress in experimental ecosystems. *Toxicology and Ecotoxicology News*, 4: 174-178.
- Van den Brink, P.J. & Ter Braak, C.J.F. (1998) Multivariate analysis of stress in experimental ecosystems by Principal Response Curves and similarity analysis. *Aquatic Ecology*, 32: 163-178.
- Van den Brink, P.J. & Ter Braak, C.J.F. (1999) Principal Response Curves: Analysis of time-dependent multivariate responses of a biological community to stress. *Environmental Toxicology and Chemistry*, 18: 138-148.
- Van den Brink, P.J., van Wijngaarden, R.P.A., Lucassen, W.G.H., Brock, T.C.M. & Leeuwangh, P. (1996) Effects of the insecticide Dursban 4E (active ingredient chlorpyrifos) in outdoor experimental ditches: II. Invertebrate community responses and recovery. *Environmental Toxicology and Chemistry*, 15: 1143-1153.
- Van den Wollenberg, A.L. (1977) Redundancy analysis. An alternative for canonical correlation analysis. *Psychometrika*, 42: 207-219.
- Van der Aart, P.J.M. & Smeenk-Enserink, N. (1975) Correlations between distribution of hunting spiders (Lycosidae, Ctenidae) and environmental characteristics in a dune area. *Neth. J. Zool.*, 25: 1-45.
- Van der Leeden, R. (1990) *Reduced rank regression with structured residuals*. DSWO Press, Leiden.
- Van der Voet, H. (1987) Correlaties van vangsten van de populieregglasvlinder met weersvariablen. GLW-nota HVO-87-18.
- Van der Voet, H. (1987) Het bepalen van behandelingseffecten op grond van korte tijdreeksen. Rapport ITI-TNO 87 ITI B 30.
- Van Dobben, H.F., Ter Braak, C.J.F. & Dirkse, G.M. (1999) Undergrowth as a biomonitor for deposition of nitrogen and acidity in pine forest. *Forest Ecology and management*, 114: 83-95.

- Van Horne, B. & Ford, R.G. (1982) Niche breadth calculation based on discriminant analysis. *Ecology*, 63: 1172-1174.
- Verdonschot, P.F.M. & Ter Braak, C.J.F. (1994) An experimental manipulation of oligochaete communities in mesocosms treated with chlorpyrifos or nutrient additions: multivariate analysis with Monte Carlo permutation tests. *Hydrobiologia*, 278: 251-266.
- Welch, W.J. (1990) Construction of permutation tests. *Journal of the American Statistical Association*, 85: 693-698.
- Whittaker, J. (1984) Model interpretation from the additive elements of the likelihood function. *Appl. Statist.*, 33: 52-65.

16. Appendix A: The extended Dune Meadow Data

Table 16.1 shows a slightly modified version of the species data matrix of the Dune Meadow from Ter Braak (1987) and Jongman et al. (1987). Along the top are the sample identification numbers. The sample identification numbers 18, 19 and 20 in the original papers have been modified in to 28, 29 and 30. The samples with identification numbers 20 and 21 have been added and will be treated as supplementary samples that do not influence the analysis. Notice that sample 20 is actually a duplication of sample 17. The species numbered 31-33 have also been added and will be treated as supplementary species. For 20 out of the 22 sites we also have environmental data (Table 16.2). Five "environmental" variables were recorded at these sites, two of which are nominal. The (semi-) quantitative variables are (1) A1: thickness of the A1 horizon (in mm), (2) Moisture: moisture content of the soil scored on a five-point scale, (3) Manure: quantity of manuring, also scored on a five-point scale. The nominal variables are (4) agricultural use, with the three classes hayfield, haypasture and pasture, and (5) management regime, with the four classes standard farming (SF), bio-dynamical farming (BF), hobby farming (HF) and nature management (NM). In the original data there were two missing values. For these the mean of the variable has been imputed. The imputed cells are indicated by an asterisk in Table 16.2. The original numbering of the samples is reflected in the sample code name in Table 16.2. The sample identification numbers correspond to those in Table 16.1.

Table 16.1 Dune meadow vegetation data of the island of Terschelling, The Netherlands. The table shows the abundance values (response values) of 33 plant species (rows) in 22 sample plots (columns of one digit width). The values are on a 1-9 scale and replace the original codes of the Blaun-Blanquet scale. A blank (space) denotes absence. The data is a subset from M. Batterink and G. Wijffels (unpubl.).

Samples	1111111122223																							
species	1234567890123456701890																							
1 <i>Achillea millefolium</i>	13	222	4																		2	2		
2 <i>Agrostis stolonifera</i>	48	43	45	44	7																	5		
3 <i>Aira praecox</i>																					2	2	3	
4 <i>Alopecurus geniculatus</i>	272	53	85	4																				
5 <i>Anthoxanthum odoratum</i>		432	4																		4	4	4	
6 <i>Bellis perennis</i>	3222	2																				5	2	
7 <i>Bromus hordeaceus</i>	4	32	2	4																				
8 <i>Ghenopodium album</i>																							1	
9 <i>Cirsium arvense</i>	2																							
10 <i>Eleocharis palustris</i>																					4		4	
11 <i>Elymus repens</i>	44444	6																						
12 <i>Empetrum nigrum</i>																							2	
13 <i>Hypochaeris radicata</i>																					2	2	5	
14 <i>Juncus articulatus</i>																					44	33	4	
15 <i>Juncus bufonius</i>																					2	4	43	
16 <i>Leontodon autumnalis</i>	52233332352222	2	2562																					
17 <i>Lolium perenne</i>	75652664267	2																						
18 <i>Plantago lanceolata</i>	555	33	2	23																				
19 <i>Poa pratensis</i>	44542344444	2	1413																					
20 <i>Poa trivialis</i>	2765645454	49	2																					
21 <i>Potentilla palustris</i>																							22	
22 <i>Ranunculus flammula</i>																					2	2222	4	
23 <i>Rumex acetosa</i>																					563	2	2	3
24 <i>Sagina procumbens</i>	5	22	242																				3	
25 <i>Salix repens</i>																							335	
26 <i>Trifolium pratense</i>	252																							
27 <i>Trifolium repens</i>	52125223633261	22																						
28 <i>Vicia lathyroides</i>																					12		1	
29 <i>Brachythecium rutabulum</i>	2226222244	44	634																					
30 <i>Calliargonella cuspidata</i>																					4	3	3	
31 <i>Hippophae rhamnoides</i>																					1		21	
32 <i>Poa annua</i>	3364	2	232	3																			4	
33 <i>Ranunculus acris</i>																					232	2	11	

Table 16.2 Dune meadow environmental data of the island of Terschelling, The Netherlands.

The asterisk denotes an imputed value (the mean of the variable).

Sample Code name	Sample Identification Number	A1 horizon	Moisture	Manure	Use	Management regime
Sample 1	1	2.8	1	4	Haypasture	SF
Sample 2	2	3.5	1	2	Haypasture	BF
Sample 3	3	4.3	2	4	Haypasture	SF
Sample 4	4	4.2	2	4	Haypasture	SF
Sample 5	5	6.3	1	2	Hayfield	HF
Sample 6	6	4.3	1	2	Haypasture	HF
Sample 7	7	2.8	1	3	Pasture	HF
Sample 8	8	4.2	5	3	Pasture	HF
Sample 9	9	3.7	4	1	Hayfield	HF
Sample 10	10	3.3	2	1	Hayfield	BF
Sample 11	11	3.5	1	1	Pasture	BF
Sample 12	12	5.8	4	2*	Haypasture	SF
Sample 13	13	6.0	5	3	Haypasture	SF
Sample 14	14	9.3	5	0	Pasture	NM
Sample 15	15	11.5	5	0	Haypasture	NM
Sample 16	16	5.7	5	3	Pasture	SF
Sample 17	17	4.0	2	0	Hayfield	NM
Sample 18	28	4.6*	1	0	Hayfield	NM
Sample 19	29	3.7	5	0	Hayfield	NM
Sample 20	30	3.5	5	0	Hayfield	NM

17. Appendix B: Mathematical derivations

17.1 Environmental biplot scores represent covariances or weighted averages

In this section we show that the environmental biplot scores, given by equation (6.32), can in linear methods be obtained from a weighted regression of the weighted covariances between species and environmental variables on the species scores. For given species scores, the environmental biplot scores thus best approximate the covariances between species and environmental variables. We then show that, in unimodal methods, weighted averages between species and environmental variables are approximated. The formulae in this section hold for both direct and indirect gradient analyses, and also for supplementary environmental variables. For direct gradient analysis, the biplots have even stronger least-squares properties: not only the environmental arrows are optimally positioned but the species points are optimally positioned simultaneously, as shown in the Appendices of Ter Braak (1986, 1987) [Unimodal models: pp 72 and pp 80].

The weighted covariance of species k and environmental variable j is given by the formula

$$(17.1) \quad r_{jk} = \sum_i w_i z_{ij} y'_{ik} / \sum_i w_i$$

If both the species data and environmental data are standardized to unit variance, covariances are equal to correlation coefficients. Let b_k be the species score of species k on a particular ordination axis ($k = 1, \dots, m$). The regression coefficient of the weighted regression of the covariances between the m different species and environmental variable j on the species scores $\{b_k\}$ yields

$$(17.2) \quad c_j^* = \sum_k w_k r_{jk} b_k / \sum_k w_k b_k^2$$

where c_j^* is called the environmental biplot score for environmental variable j on the ordination axis. On inserting (17.1) in (17.2) and interchanging the order of summation, we obtain, consecutively,

$$(17.3) \quad c_j^* = \sum_k w_k (\sum_i w_i z_{ij} y'_{ik} / \sum_i w_i) b_k / \sum_k w_k b_k^2$$

$$(17.4) \quad c_j^* = \sum_i \sum_k (w_i w_k z_{ij} y'_{ik} b_k / \sum_k w_k b_k^2) / \sum_i w_i$$

$$(17.5) \quad c_j^* = \sum_i w_i z_{ij} [\sum_k w_k y'_{ik} b_k / \sum_k w_k b_k^2] / \sum_i w_i$$

The term between square brackets is equal to the species-derived sample score x_i^* (6.19), so that

$$(17.6) \quad c_j^* = \sum_i w_i z_{ij} x_i^* / \sum_i w_i$$

which is the weighted covariance between environmental variable j and the sample scores $\{x_i^*\}$. Equation (6.32) is obtained from (17.6) by noting that the environmental variables are standardized to unit variance.

For unimodal methods the above theory carries over by noting that the weighted covariance (17.1) is actually a weighted average when using the definitions for w_i , given below (6.5) on page 157, and for y'_{ik} , given by (6.16). On inserting these definitions in (17.1), we obtain

$$(17.7) \quad r_{jk} = \frac{\sum_i w_i^* y_{ik} z_{ij}}{y_{+k}} = \frac{\sum_i w_i^* y_{ik} z_{ij}}{\sum_i w_i^* y_{ik}}$$

the weighted average of species k with respect to environmental variable j (cf. (6.11) with $\alpha = 0$). For unimodal methods in biplot scaling, $b_k = u_k$, and the formula for the species-derived sample score x_i^* , equation (6.19), also holds true as was shown in equation (6.25). With these equivalencies the theory carries over. The factor $1-\lambda_k$ in (6.32) for methods in Hill's scaling derives from the term $1/\sum_k b_k^2$ in (17.2) which differs a factor $(1-\lambda)$ between Hill's scaling and the biplot scaling (compare (6.14) with (6.13) with u_k replaced by b_k).

If there are covariables in the analysis, the environmental variables are adjusted for covariables. This means that each environmental variable is residualized with respect to the covariables, i.e. the environmental data are replaced by the residuals of the regression of the environmental data on the covariables (see Unimodal models: pp 134 - 138). After the replacement, the $\{z_{ij}\}$ in equations (17.1) - (17.7) represented the residualized environmental data. The covariance (17.1) is then a **partial covariance** in terms of the original variables. By consequence, if there are covariables in the analysis, the species scores and environmental biplot scores represent, in linear methods, partial covariances and, in unimodal methods, weighted averages with respect to residualized environmental variables.

17.2 Environmental centroid scores represent class means or class totals

Each nominal environmental variable groups samples in to a number of classes. The mean abundance of a species in a class is called its class mean. In this section we show that the centroid scores, given by equation (6.34), can in linear methods be obtained from a weighted regression of the class means on the species scores. For given species scores, the environmental centroid scores thus best approximate the means of species in environmental classes. We then show that, in unimodal methods, relative class totals are approximated. The formulae in this section hold for both direct and indirect gradient analyses and for supplementary environmental classes. For direct gradient analysis, the biplots have even stronger least-squares properties: not only the environmental points are optimally positioned but the species points are optimally positioned simultaneously, as shown in the Appendix of Ter Braak (1994) [Unimodal models: pp 151].

The weighted mean abundance of species k in class j , indicated by the dummy variable z_{ij} , is defined by the formula

$$(17.8) \quad m_{jk} = \frac{\sum_i w_i z_{ij} y'_{ik}}{\sum_i w_i z_{ij}}$$

For class j , the regression coefficient of the weighted regression of the class means $\{m_{jk}\}$ [$k = 1, \dots, m$] on the species scores $\{b_k\}$ [$k = 1, \dots, m$] is given by

$$(17.9) \quad c_j^* = \frac{\sum_k w_k m_{jk} b_k}{\sum_k w_k b_k^2}$$

On inserting (17.8) and (17.9) and interchanging the order of summation, we obtain, analogously to the derivation in the previous section below equation (17.3)

$$(17.10) \quad c_j^* = \sum_i w_i z_{ij} x_i^* / \sum_i w_i z_{ij}$$

which is precisely the centroid scores defined in (6.34).

The class mean (17.8) has a special meaning in unimodal methods. On inserting the definitions for w_i , given below (6.5) on page 157, and for y'_{ik} , given by (6.16), in (17.8), we obtain

$$(17.11) \quad m_{jk} = (y_{j+}/y_{+k}) \sum_i w_i^* z_{ij} y_{ik} / \sum_i w_i^* y_{i+} z_{ij} = y_{jk} y_{j+} / (y_{j+} y_{+k})$$

where, in a loose, but informative notation, y_{jk} is the total abundance of species k in class j and y_{j+} is the total abundance across all species in class j . With this notation, (17.11) is analogous to the transformation (3.6), that is implicit in (canonical) correspondence analysis. The notation emphasizes that a class acts as a super sample: m_{jk} is related to the total abundance y_{jk} in the class as y'_{ik} is related to y_{ik} . On inserting (17.11) in (17.9) yields (17.10) if the biplot scaling is used. The class centroid c_j^* is thus the result of the weighted regression of the relative class totals m_{jk} on the species scores. The class centroids and species scores together thus form a biplot that represents the relative class totals. These totals are fitted by weighted least-squares. The interpretation is precisely as that of the biplot of samples points and species points, namely in terms of either y_{jk}/y_{j+} or y_{jk}/y_{+k} , as described in sections 6.3.4 and 6.3.5, and as summarized on page 171 of Unimodal models.

In Hill's scaling, the class centroids and species scores do not form a biplot. Nevertheless, the plot represents the (relative) class totals by the centroid principle: the centroid scores c_j^* given by (17.10) and the sample scores x_i^* have the same relation to the species scores $\{u_k\}$. In particular, by inserting (6.21) in (17.10), we obtain, analogously to (6.21),

$$(17.12) \quad c_j^* = \lambda^{\alpha-1} \sum_k w_k^* y_{jk} u_k / \sum_k w_k^* y_{jk}$$

The interpretation is thus precisely that of a joint plot of sample points and species scores in Hill's scaling. The derivation shows that the class point is added in very much the same way as a supplementary sample, using as data the values y_{jk} . The same can be shown in linear methods, with the data $\{y'_{ik}\}$ replaced by the means $\{m_{jk}\}$.

The environmental centroids also have an attractive interpretation when there are covariables in the analysis, namely in terms of adjusted means and adjusted totals. Adjusted means and totals are means and totals from which the effects of the covariables have been removed by regression. This is most easily proven by using projection operators and matrix algebra, but the essence of the proof is given here in terms of residuals of regressions on the covariables. If there are covariables in the analysis, the centroids are still calculated by equation (17.10). The species-derived sample scores $\{x_i^*\}$ are uncorrelated with the covariables, because they have been regressed on the covariables (see Step A3 and (A.16) in Unimodal models: pp 136-137). This means that both sides of equation (17.10) can be interpreted as residuals of a regression on the covariables. Carrying this interpretation over from (17.10) to (17.9) implies that the right-hand side of (17.9) should not change if replaced by residuals of a regression on the covariables. Therefore, the $\{m_{jk}\}$ must be uncorrelated with the covariables, i.e. the $\{m_{jk}\}$ are adjusted means. An adjusted mean is the mean of a residualized variable. In the present case, the species data $\{y'_{ik}\}$ are residualized, i.e. y'_{ik} in equation (17.8) is replaced by the residual of the regression of the species data $\{y'_{ik}\}$ on the covariables.

17.3 Sum of all canonical eigenvalues (trace)

The trace, if reported as result of the Monte Carlo test, is equal to the sum of the canonical eigenvalues. The subsection gives the short-cut formula by which the sum of the canonical eigenvalues is calculated, using the notation of the Appendix of Ter Braak & Prentice (1988) [Unimodal models: pp134-138]. Their notation differs from that in this manual in that samples are columns, not rows. The sum of the canonical eigenvalues is calculated in CANOCO as the trace of the matrix

$$(17.13) \quad R^{-1}YZ_2'(Z_2WZ_2')^{-1}Z_2Y'$$

where Z_2 is the “ Z_2 with tilde” in Ter Braak & Prentice (1988).

18. Appendix C: Format of (W)Canolmp files

This appendix gives a formal description of the output files produced by Canolmp and WCanolmp. The condensed-format data files have the following general format (characterized here by a FORTRAN format specification, which appears on the second line of the output data file):

(I5,1X,<NPL>(I6,F<w>.<p>))

where <w> is the total width of the single value field (plus one, to get the values preceded by a blank), <p> is the number of decimal digits, and <NPL> is the number of the data couples per a single line (so that the line is no longer than 80 characters in total).

The full-format data files have the following general format (displayed again as a FORTRAN format specification):

(I5, 1X, <N>F<w>.<p>)

where <w> and <p> have the same meaning as before and <N> is the number of variables in the data set. That form is used if all the values for a single sample fit into one row of the output file. If not, the following format specification is used:

(I5, 1X, <N>F<w>.<p>, <NL>/6X, <N>F<w>.<p>))

where <w> and <p> have the same meaning as before, <N> is the maximum number of value fields fitting on a single line of the output file, and <NL> is the number of the continuation lines needed to display values of the all variables in the data set. Note that the value of (<NL>+1) * <N> might be, in fact, larger than the number of variables in the data set. In any case, the correct number of variables is given on the third line of the output data file and this is the number used by CANOCO when reading the data files.

19. Appendix D: CanoDraw software setup

CanoDraw supports work in a networked environment and generally in any environment with increased security restrictions (e.g. where the installation directory is read-only for the ordinary users). In the following comments, I discuss various aspects of the CanoDraw setup which might be useful for a system administrator who wants to assure unlimited use of CanoDraw in such environments.

- * CanoDraw for Windows can be installed either as part of a new installation of the Canoco for Windows package or as an upgrade of the already installed Canoco package. During the software installation, the CanoDraw setup program modifies settings stored in the *HKEY_LOCAL_MACHINE* sub-tree of the Windows™ registry, so it must be installed by an user with administrative rights.
- * On the other hand, the CanoDraw program itself does not modify the *HKEY_LOCAL_MACHINE* registry sub-tree, but only modifies the entries in the registry sub-tree corresponding to the current user. CanoDraw settings are stored in the key named *HKEY_CURRENT_USER\Software\Smilauer\CanoDraw4.0*, which has several subkeys. Most of the key values stored in this CanoDraw subkey should not be directly modified with the registry editor, because they are accessible using the CanoDraw for Windows user interface (particularly in the dialog boxes invoked by the menu commands *View / Diagram Settings*, *View / Visual Attributes*, and *View / Workspace Settings*). The only exception is discussed in the next paragraph, which is an option, which can be modified by the system administrator.
- * CanoDraw stores most of the current settings corresponding to the options in *Diagram Settings* and *Visual Attributes* dialog boxes at the time of application exit and retrieves them again when the application starts. For Windows NT 4.0, Windows 2000, and Windows XP, these data are stored in the Windows™ registry, as a binary value (type *REG_BINARY*) named *Global*, with an approximate size of 101 KB. The *Global* value is set for the key named *HKEY_CURRENT_USER\Software\Smilauer\CanoDraw\Settings*. The Windows registry does not accept such large data on the operating systems Windows 95, Windows 98, or Windows ME, so CanoDraw must store these defaults in a file. The default file name is *settings.ini* and its location is identical to the folder where the CanoDraw program file (*canodrw4.exe*) is stored. This might cause problems in the situation where the installation directory is not available for writing for the programs started by ordinary users and where user's profile roaming needs to be supported. When looking for the settings file on Windows 9x or Windows ME, CanoDraw first checks a registry key named *HKEY_CURRENT_USER\Software\Smilauer\CanoDraw4.0\Global*, looking for the value named *SettingsFile* (with string type *REG_SZ*). This value is stored in the registry for each user at the time CanoDraw first starts (under the above-listed types of operating systems) if it is not already present there. If found, CanoDraw takes it as a full path (containing both the volume and folder address and the file name and extension) to the file from where the settings should be retrieved at the program start and where the settings are stored at the program exit. You can therefore use this registry value to relocate the position of the file to another place. **Note that you must setup the file location separately for each user, because the contents of the *HKEY_CURRENT_USER* sub-tree changes depending on the identity of currently logged user.** The recommended path is *<WindowsNT-dir>\Profiles\<user-name>\Application Data\CanoDraw\settings.ini*, but you must create the innermost folder (*CanoDraw*) yourself, it is not created by CanoDraw.

20. Index

A

- Active species and samples · 117, 174, 227
- AIC · 344
- Algorithms · 35, 36, 38, 64, 164, 307, 311
- Amplitude ecological · *See* Species tolerance
- Analysis of concentration · 173
- Analysis of variance · 14, 51, 63, 104, 172, 210, 212, 260, 262, 265
- Arch effect · 62, 88, 191, 192, 220, 226, 238
- Attribute plot · 395
- Autocorrelation · 45, 54, 109, 214, 272, 279

B

- Before-After Control-Impact design · 54, 104, 110, 210, 212, 279, 282, 284, 286
- Bell-shaped response curve · *See* Gaussian curve
- Biplot · 34, 194, 390
 - interpolative · 41
 - predictive · 41
 - regression · 41, 54, 56, 164, 180, 181, 255, 394
 - rule · 40, 41, 56, 91, 195, 196, 231, 302, 414
 - T-value · *See* T-value biplot
- Biplot scaling · 41, 92, 144, 147, 151, 159, 169, 196, 242, 480
- Biplot scores of environmental variables · 41, 137, 141, 153, 156, 168, 182, 194, 479
- Blocks · 44, 105, 128, 211, 215, 265
 - design · 104, 105, 211, 215, 262, 265, 286
- Bray-Curtis distance · 307

C

- CA · *See* Correspondence Analysis
- Calibration · 37, 247
- CANOCO
 - console version · 115, 117, 127, 185, 278
 - data files · 117, 189
 - Windows version · 20, 81, 115, 185
- CanoDraw · 19, 28, 112, 185, 223, 227, 241, 256, 311–466
- CanoImp · 26, 67–71, 227, 483
- CanoMerge · 69–71
- Canonical coefficients · 143, 149, 163, 180, 237, 249
 - destandardization · 165, 245
 - standardized · 163
- Canonical Correlation Analysis · 38, 123, 204
- Canonical Correspondence Analysis · 49, 56, 61, 149, 230, 249, 251
- Canonical eigenvalue · 34, 49, 51, 63, 103, 126, 131, 165, 210, 482

- Canonical ordination · 33, 34, 35, 37, 86, 119, 191
- Canonical Variates Analysis · 62, 92, 173, 298
- Categorical data · 61, 118
- CCA · *See* Canonical Correspondence Analysis
- Centering species data · 93, 161, 174, 204, 244
- Centroid of species scores · 158, 170
- Centroid principle · 39, 41, 92, 93, 143, 144, 147, 150, 156, 196, 414, 481
- Centroids of environmental variables · 41, 142, 154, 156, 171, 241, 256, 480
- Centroids of sample scores · 39, 155, 156, 161, 171, 309
- Chi-square distance · 92, 145, 148, 152, 177, 196, 307
- Chi-square statistic · 50, 118, 174, 197
- Choice of methods · 88, 191
- Classifying ordination items · 366–73, 445–50
- Clipboard · 25, 26, 31, 67, 101, 134, 228, 236, 239, 331, 379, 382
- Collinear variable · 118, 122, 131, 166, 212, 276
- Collinearity · *See* Collinear variable
- Community · 33, 34, 38, 249, 291
- Compositional data · 56–61, 60, 93, 295
- Condensed file format · 61, 68, 75–78, 202, 483
- Conditional effect · 31, 101, 129, 164, 169, 172. *See also* Forward Selection, conditional effects
- CON-file · *See* Project file
- Constrained ordination · 33, 37, 55, 86, 123
- Contour lines · 338, 396, 398, 400, 418, 434, 454
- Correlation
 - among environmental variables · 41, 120
 - among species · 29, 41, 138, 141, 195
 - inter-set · 158, 167, 365
 - intra-set · 168, 169
 - table of species-environment · 115, 182, 241
- Correlation biplot · *See* Scaling of ordination scores
- Correlation matrix · 29, 87, 93, 120, 167, 195
- Correspondence Analysis · 36, 59, 61, 144, 226
- Count data · 51, 200, 229, 245, 289, 342
- Covariables · 33, 34, 119, 161, 166, 169, 177, 189
- Covariance biplot · *See* Scaling of ordination scores
- CVA · *See* Canonical Variates Analysis

D

- Data editing options · 27, 95, 176, 223, 228, 235
- Data transformation · *See* Transformation
- DCA · *See* Detrended Correspondence Analysis
- DCCA · *See* Detrended Canonical Correspondence Analysis
- DECORANA · 90, 162, 192, 193, 197, 201, 203
- Default values · 44, 125, 132, 187, 193, 211
- Degrees of freedom · 48, 52, 53, 54, 165, 181, 259, 302
- Deleting
 - covariables · 95, 96, 198
 - environmental variables · 95, 96, 199

samples · 95, 96, 197
species · 95, 96
Detrended Canonical Correspondence Analysis · 90, **149**,
155, 193, 197, 237, 247
Detrended Correspondence Analysis · **145**, **148**, 197, 226
Detrending methods · 88, 90, 125, **192**, 237
Diagnostics · *See* Ordination diagnostics, Regression
diagnostics
Direct gradient analysis · **34**, 86, 119, 191
Discriminant analysis · *See* Canonical Variates Analysis
Dissimilarity data · *See* Principal Coordinates Analysis
Down-weighting of rare species · 94, 203, 218
Dual scaling · *See* Correspondence Analysis
Dummy variable · 61, 99, 105, 110, 131, 171, 173, 212,
262, 480

E

Effective number
 of samples · *See* N2, effective number of occurrences
 of species · *See* N2, diversity index
Eigenvalue equations · 62, 138, 141, 148, 153
Eigenvalues · **34**, 39, 118, 123, 125, 130, 135, 152, 158,
161, 205
Eigenvector scores · 35, **157**, 166, 179
Environmental variables · **34**, 86, 100, 123, 190, 199, 204
Error messages · 71, 115, 189
Euclidean distance · *See* Pythagorean distance
Euclidean distance biplot · *See* Scaling of ordination scores
Example projects · **221**
Experimental design · 34, 46, 52, 54, 104, 209, 211, 265,
275
Explanatory variables · *See* Environmental variables and
Covariables

F

Factor analysis · *See* Principal Components Analysis
Factorial design · 51, 265, 293
Fit into ordination space
 samples · **176**, 197, 363
 species · 91, **174**, 197, 204, 235, 284, 362, 435
Fitted values · **40**, 49, 56, 123, 150, 151, 152, 153, 163,
175
Format
 of CANOCO project file · 82, **217**
 of data files · **71–80**
Forward Selection · 100, 102, 194, 216, 236, 251
 automatic · 30, 100, 101, 129
 conditional effects · 31, 101, 129
 marginal effects · 31, 101, 129
 results · 31, 101
F-ratio · 43, 47, **49**, 50, 52, 53, 63, 103, 210, 258
Free file format · **78–80**
F-test · *See* F-ratio
Full file format · **71–74**, 182, 483
Full model · 48, 100, 104, 127, 189, 211, 259

Fuzzy coding · 62
F-value · *See* F-ratio

G

Gaussian curve · 38, **59**, 419
Generalized Additive Models · 345, **421–22**, 436–39
Generalized Linear Models · 59, 61, 341, **418–21**, 439
Goodness of fit · *See* Ordination diagnostics and Summary
 of analysis
Gradient · **34**, 49, 59, 86, 88, 103, 210, 238, 245. *See also*
 Length of gradient
Gradient analysis · 34, 37, 86, 191
Graphics · *See* CanoDraw
Grids · **45**, 104, 106, 210, 212, 214, 252, 272
Groups of ordination items · **373–76**

H

Hellinger distance · 307
Heterogeneity of samples · *See* Sample heterogeneity
Hill's scaling · 39, **92**, 145, 146, 150, 159, 162, 168, 196,
480
Homogeneity analysis · 61, 191
Hybrid methods · 88, 125, 192, 220
Hypothesis
 alternative · 43, 51, 103, 210, 280
 null · **43**, 44, 47, 49, 103, 165, 210, 280

I

Impact models · 280, 282, 284
Importing data
 from databases · **67**, 380
 from spreadsheets · 26, **67**, 379
Indicator variable · *See* Dummy variable
Indirect gradient analysis · **34**, 42, 86, 191, 226
Inertia · **118**, 123, 175, 194, 260, 268
Influential data · 94, 119, 157, 176, 200, 246, 247, 255
Initialization file · 117, **187–88**, 274
Input data · 67, 87, 117, 189
Installation · 17
Interactions
 covariables · 95, 99, 198
 environmental variables · 95, 99, 200, 230, 252, 275
Interpretation
 of ordination diagrams · 39, **135**, 144, 145, 180, 195,
 414
 of results · 226, 230, 237, 241, 245, 284, 287
Isolines · *See* Contour lines

J

Jaccard similarity coefficient · 307
Jittering · 333

Joint plot · 34, 144, 390, 481

L

Labels · 68, 69, 315, 319, 331, 333, 340, 353, 360, 410
Latent variable · 34, 35, 229
Least-squares · 38, 40, 62, 163, 191, 481
Length of gradient · 90, 125, 196, 197, 199, 227, 276
Level-accuracy · 48, 53, 259, 268, 269
Leverage · 105, 119
Linear combination · 34, 36, 55, 62, 118, 123, 139, 166, 200, 299
Linear Discriminant Analysis · *See* Canonical Variates Analysis
Linear methods · 34, 91, 117
Linear model · 50, 57, 60, 94, 191, 200, 280
Linear transects · 45, 104, 106, 210, 212, 214, 270, 454–58
Loading · 35
Loess smoother · 347, 422
Log file · *See* Log View
Log View · 25, 82, 111, 113, 115, 129
Log window · *See* Log View
Log-ratio analysis · 57, 93, 282, 286, 295

M

Marginal effect · 31, 101, 129, 130, 169, 172, 270
Maximum data size · 13
Means · 41, 53, 62, 94, 120, 124, 173, 206, 245, 298
Metric scaling · 34, 64
Monte Carlo permutation · *See* Permutation tests
Multicollinearity · 122, 131
Multidimensional Scaling · 34, 64
Multinomial logit model · 58
Multiple correlation · 121, 164, 169, 195, 234, 302
Multiple correspondence analysis · 61, 191
Multiple regression · 35, 55, 139, 143, 149, 163, 179, 194, 227, 301
Multivariate regression · 34, 47, 50, 55, 166, 180, 255

N

N2
diversity index · 161, 246, 417
effective number of occurrences · 157, 251, 383
Nested sampling design · 44, 104, 214, 275
Nominal environmental variables · 39, 91, 165, 169, 171, 195, 234
Nominal variables · 38, 61–62, 191, 365
Non-standard analyses · 192, 219

O

Optimum · 35, 36, 60, 88, 92, 196, 220, 335, 419
Ordination diagnostics · 115, 174–79, 197

Ordination diagram · 35, 39, 112, 317, 385–99
ORIGIN in solution file · 161, 302
Origin of ordination diagram · 40, 158, 161, 170
Outliers · 115, 119, 246

P

Partial canonical ordination · 38
Partial correlation · 120, 167, 196
Partial ordination · 37, 126
Partial test · 43, 47, 50, 258, 259, 268, 277, 301
Passive · *See* Supplementary
PCA · *See* Principal Components Analysis
Percentage data · 56, 58, 60, 62, 265, 295
Percentage variance accounted for · 63, 123, 124, 125, 182
Permutation
design-based · 52, 53, 109, 268, 293
model-based · 48, 52, 268, 293
Permutation tests · 43–44, 48, 102, 103, 109, 126, 209, 236
on first canonical eigenvalue · 51, 127
on trace · 51, 127
Permutation types · 44, 104, 106, 210, 262
Permutations
number of · 44, 103, 211
PrCoord · 305–10
Principal Components Analysis · 36, 55, 135, 191, 206, 252, 295
centered · 93, 191
double centered · 191, 206
log-ratio · 191, 295
non-centered · 183, 191, 206
of instrumental variables · *See* Redundancy Analysis
on a correlation matrix · 93, 206
standardized · 191, 206
Principal Coordinates Analysis · 64, 191, 305
Principal Response Curves Analysis · 287, 288, 291, 381, 461–66
Project file · 21, 81, 111, 186, 217
CanoDraw project · 312, 322
Project Setup Wizard · 22, 81
Project View · 24, 82, 85, 111
Pseudo-F statistic · 49, 53, 282
Pythagorean distance · 64, 136, 138, 139, 141, 145, 176, 195, 307

R

RDA · *See* Redundancy Analysis
Reciprocal Averaging · *See* Correspondence Analysis
Reduced model · 48, 52, 100, 103, 127, 211, 259
Reduced rank regression · 38, 54, 255, 288
Redundancy Analysis · 54, 139, 204, 241, 295, 301
distance-based · 305, 308
Regression · 40, 55, 301, 341, 399, 418–25
logit · 38, 343
log-linear · 38, 58, 59
Poisson · 38

Regression biplot · *See* Biplot, regression
Regression coefficients · 118, 137, 145, 149, 163
 destandardization · 165, 245
 standardized · 163
Regression diagnostics · 120, 423
Regression sum of squares · 103, 174, 194, 210, 259
Repeated measurement design · 54, 104, 210, 212, 282,
 284, 286, 288
Rescaling
 nonlinear · 132, 148, 155, 158, 161, 168, 171, 193
 ordination scores · 335, 336, 390, 412, 442
Residual sum of squares · 49, 51, 103, 210, 253, 259, 264
Response model · 36, 37, 60, 88, 125, 158, 178, 193, 226,
 401
Response surface · *See* Contour lines
Response variable · 34, 38, 42, 43, 61, 86, 200, 277, 301

S

Sample heterogeneity · 179, 197
Sample scores
 derived from environmental variables · 123, 139, 145,
 149, 152, 154, 157, 164, 166
 derived from species · 92, 123, 135, 137, 139, 144, 157,
 160, 167, 196, 358
Sampling design · 44, 46, 48, 54, 104, 253
Scaling of ordination scores · 60, 92, 117, 136, 139, 149,
 162, 195–97
SD · *See* Length of gradient
Seasonal variation · 249
Selection of species for plotting · 441
Selection of variables · *See* Forward Selection
Series of ordination items · 376–78
Setup Wizard · *See* Project Setup Wizard
Significance test · 44, 102, 103, 194, 200, 234, 237, 258
Slope parameter · 36, 40, 55, 135, 139, 141, 152, 158, 161
Soerensen similarity coefficient · 308
Solution file · 87, 132, 190
Spatial data · 46, 245, 252, 272, 451–54
Species response curve · 401, 436
Species response surface · *See* Contour lines
Species scores · 35, 91, 157, 195
Species tolerance · 133, 178, 197, 299, 335, 419
Species-by-environment table · 179–82, 204, 230, 241
Split-plot design · 46, 53, 104, 108, 212, 213–14, 262, 275
Standard deviation
 of environmental variables · 120, 245
 of ordination axes · 120, 168
Standardization
 of environmental variables · 41, 163, 164, 166, 167,
 182, 200
 of species data · 93, 158, 174, 204, 244
Statistical test · *See* Significance test
Stratified sampling · 44, 104
Succession · 247
Summary of analysis · 28, 85, 115, 122–26, 165, 174, 243
Supplementary environmental variables · 35, 42, 86, 134,
 143, 169, 171, 173, 209, 228, 314

Supplementary samples · 35, 42, 95, 98, 117, 161, 174,
 177, 179, 227
Supplementary species · 35, 42, 95, 98, 158, 202, 278
Symbols
 list of · 133, 135, 207
System requirements · 13

T

TAU · *See* Total standard deviation
Time series · 45, 104, 106, 108, 109, 110, 210, 212, 214,
 280, 455
Tip of Day window · 20, 83
Tolerance · 65, 335, 419
Total standard deviation · 118, 135, 205, 288, 296, 302
Total sum of squares · 63, 118, 174, 177, 205
Trace statistics · 51, 63, 127, 210, 259, 266
Transformation
 logarithmic · 57, 61, 93, 94, 176, 200, 295
 of environmental data · 228
 of species data · 64, 94, 117, 185, 200
 square root · 94, 176, 200, 229, 246
Transition formulae · 136, 140, 146, 147, 150, 156, 162,
 202
Trend · 45, 245, 273
Triplot · 35, 124, 241, 298, 394, 444
Tutorial · 20–31, 308, 427
T-value biplot · 179–82, 255, 392, 459
T-values of regression coefficients · 118, 122, 163, 165,
 180, 228, 255, 280, 302, 364

U

Unfolding · *See* Correspondence Analysis
Unimodal methods · 92, 117
Unimodal models
 fitting · 341, 345, 401, 435

V

Van Dobben circles · 180, 393
Variance decomposition · 252, 260
Variance Inflation Factor · 120, 121, 167, 180, 232
VIF · *See* Variance Inflation Factor

W

Warnings · 115, 206, 296
Weighted averages
 table of · 41, 115, 154, 182, 232
Weighted averaging · 35, 38, 124, 144, 157, 158, 191, 196,
 479
Weighting
 by error variance · 93, 204
 samples · 95, 97, 203

species · 95, 97, 202
Weights of samples · 98, **156**, **160**, 247, 363

Weights of species · 98, **156**, 157, 363
Whole-plot · **46**, 104, 108, 109, 214, 275

21. List of Figures

Figure 2-1 Where to start the installation of Canoco for Windows.....	17
Figure 2-2 Starting the installer program.....	18
Figure 2-3 User Information dialog box.....	18
Figure 2-4 Select Components page.....	19
Figure 2-5 The CANOCO workspace.....	21
Figure 2-6 The first page of the Project Setup Wizard.....	22
Figure 2-7 Creating new directory in the File Open dialog box.....	23
Figure 2-8 The Project View window.....	24
Figure 2-9 The Log View window.....	25
Figure 2-10 Selecting the data table in the spreadsheet application.....	26
Figure 2-11 Data Editing Choices in the Project Setup Wizard.....	29
Figure 2-12 Interactions of Environmental variables in the Project Setup Wizard.....	30
Figure 2-13 Forward Selection summary.....	31
Figure 3-1 Linear response model in PCA and RDA.....	36
Figure 3-2 Unimodal response model in (D)CA and CCA.....	37
Figure 3-3 Comparison of Gaussian and multinomial logit models.....	59
Figure 4-1 WCanolmp program window.....	67
Figure 4-2 CanoMerge program window.....	70
Figure 5-1 Canoco for Windows empty workspace.....	83
Figure 5-2 Canoco for Windows workspace with Project and Log views.....	84
Figure 5-3 Project View window.....	85
Figure 5-4 Available Data wizard page.....	86
Figure 5-5 Data Files wizard page.....	87
Figure 5-6 Type of Analysis wizard page.....	88
Figure 5-7 Canonical Axes wizard page.....	89
Figure 5-8 Detrending Method wizard page.....	90
Figure 5-9 Scaling: Linear Methods wizard page.....	91
Figure 5-10 Scaling: Unimodal Methods wizard page.....	92
Figure 5-11 Centering and Standardization wizard page.....	93
Figure 5-12 Transformation of Species Data wizard page.....	94
Figure 5-13 Data Editing Choices wizard page.....	95
Figure 5-14 Delete Items wizard page.....	96
Figure 5-15 Set Weights wizard page.....	97
Figure 5-16 Add Items dialog.....	98
Figure 5-17 Supplementary Samples / Species wizard page.....	98
Figure 5-18 Interactions of Variables wizard page.....	99
Figure 5-19 Forward Selection wizard page.....	100
Figure 5-20 Forward Selection report dialog.....	101
Figure 5-21 Forward Selection Step dialog.....	102

Figure 5-22 Permutation Test results dialog.....	102
Figure 5-23 Global Permutation Test wizard page.....	103
Figure 5-24 Permutation Type wizard page.....	104
Figure 5-25 Definition of Blocks wizard page.....	105
Figure 5-26 Permutation Restrictions wizard page.....	106
Figure 5-27 Grid Dimensions wizard page.....	107
Figure 5-28 Split-Plot Design I wizard page.....	108
Figure 5-29 Split-Plot Design II wizard page.....	109
Figure 5-30 Canoco for Windows workspace after analysis.....	112
Figure 5-31 Canoco Options dialog.....	113
Figure 8-1 Dune meadow data: CCA triplot in biplot scaling with focus on species (scaling 2).	243
Figure 8-2 Ordination diagram based on RDA with model block+N*PK in E40.....	267
Figure 8-3 PRC diagram of the simulated data.....	288
Figure 8-4 Ordination diagram based on the redundancy analysis of the coxite data (Aitchison 1984a).	297
Figure 8-5 Triplot based on a CVA of the Fisher's Iris data.....	298
Figure 9-1 User interface of PrCoord program.....	305
Figure 9-2 Specifying input data for db-RDA in Canoco.....	309
Figure 9-3 Diagram from the distance-based RDA, first two axes.....	310
Figure 11-1 Three types of project windows.....	316
Figure 11-2 Two types of graph windows.....	316
Figure 12-1 Open CanoDraw Project dialog box.....	323
Figure 12-2 Open CanoDraw Graph dialog box.....	324
Figure 12-3 Dialog for exporting graph in bitmap format.....	326
Figure 12-4 Dialog for exporting graph in metafile format.....	326
Figure 12-5 Properties 1 dialog page.....	332
Figure 12-6 Labels and Transformations dialog box.....	333
Figure 12-7 Effect of jittering on a XY graph where there is substantial overlap of points.....	334
Figure 12-8 Summary dialog for a fitted generalized linear model.....	335
Figure 12-9 Rescaling of ordination scores dialog.....	336
Figure 12-10 Properties 2 dialog page.....	339
Figure 12-11 GLM Options dialog page.....	341
Figure 12-12 Binomial Total Selection dialog.....	343
Figure 12-13 GAM Options property page.....	345
Figure 12-14 Loess Model Options dialog page.....	347
Figure 12-15 Visual Attributes Settings dialog.....	349
Figure 12-16 Workspace Settings dialog.....	350
Figure 12-17 Color property page.....	351
Figure 12-18 Line property page.....	352
Figure 12-19 Fill property page.....	352
Figure 12-20 Symbol property page.....	353

Figure 12-21 Font property page	353
Figure 12-22 Graph Contents window	354
Figure 12-23 Zoom Level dialog.....	354
Figure 12-24 Project Details window	355
Figure 12-25 Main toolbar.....	356
Figure 12-26 Graph Tools toolbar	356
Figure 12-27 Contents page.....	357
Figure 12-28 Appearance page.....	360
Figure 12-29 Inclusion Rules page.....	362
Figure 12-30 Inclusion Rules 2 page.....	364
Figure 12-31 Available Classifications dialog	366
Figure 12-32 Manual Classification dialog	367
Figure 12-33 Add New Class dialog	367
Figure 12-34 Class Members dialog.....	368
Figure 12-35 Merge Classes dialog	368
Figure 12-36 Classify From Data dialog	369
Figure 12-37 Dialog for specifying number of intervals.....	370
Figure 12-38 Dialog for specifying class thresholds (boundaries).....	371
Figure 12-39 Classification From Group dialog.....	372
Figure 12-40 Group Manager dialog	374
Figure 12-41 Dialog for group definition by a rule	375
Figure 12-42 Series Collections dialog.....	377
Figure 12-43 Dialog for editing series collection	377
Figure 12-44 Dialog for selecting classification.....	378
Figure 12-45 Import data from Clipboard dialog	379
Figure 12-46 Dialog for importing selected variables from Canoco data file.....	380
Figure 12-47 Create PRC Scores dialog.....	381
Figure 12-48 Delete Imported Variables dialog	382
Figure 12-49 Export Statistics dialog	382
Figure 12-50 Manage Dependent Graphs dialog.....	384
Figure 12-51 Dialog for creating simple ordination diagram.....	385
Figure 12-52 Scatter of species symbols	388
Figure 12-53 Scatter of sample symbols, with symbol type coded by the management type	389
Figure 12-54 Scatter of explanatory variables represented by arrows (quantitative variables) and symbols (dummy variables).....	390
Figure 12-55 Biplot with species and environmental variables, based on a Canonical Correspondence Analysis (CCA)	391
Figure 12-56 T-values Biplot Options dialog.....	393
Figure 12-57 T-values biplot diagram	393
Figure 12-58 Regression Biplot diagram from a project based on Canonical Correspondence Analysis.	394
Figure 12-59 Triplot diagram	395

Figure 12-60 Data Attribute Plot Options dialog.....	396
Figure 12-61 Symbol Attribute Plot diagram	397
Figure 12-62 Contour (3-D) Attribute Plot diagram.....	398
Figure 12-63 Results Attribute Plot dialog	399
Figure 12-64 XY Diagram Options dialog box	399
Figure 12-65 XYZ Diagram with Z values coded by symbol sizes.....	400
Figure 12-66 XY scatter diagram with fitted loess model.....	401
Figure 12-67 Dialog for fitting multiple species response curves	402
Figure 12-68 Range of diagram axes dialog.....	403
Figure 12-69 Tip of the Day dialog	408
Figure 13-1 Example of Graph Description window contents for a diagram with samples	415
Figure 13-2 Example of Graph Description window contents for a diagram with sample pie-symbols. The illustration provided within the window was removed.	415
Figure 13-3 Part of the Graph Description window contents for a biplot diagram, suggesting a joint interpretation of species arrows and symbols of nominal environmental variables.	416
Figure 13-4 Variable Summary floating window	416
Figure 13-5 Sample summary floating window.....	417
Figure 13-6 Popup menu displayed for a selected variable	418
Figure 13-7 Fitted GLM summary dialog.....	419
Figure 13-8 Report on stepwise selection of GLM using deviance tests.....	420
Figure 13-9 Report on GLM selection using AIC statistics	421
Figure 13-10 Fitted GAM summary dialog	421
Figure 13-11 Report on GAM selection using AIC statistics	422
Figure 13-12 Loess Model Results summary dialog	423
Figure 13-13 Select Model dialog box.....	424
Figure 13-14 Create Residual Plot diagram.....	424
Figure 14-1 Changing labelling style for samples	428
Figure 14-2 Triplot based on SPID_CCA project.....	428
Figure 14-3 Making the background of labels opaque	429
Figure 14-4 Moving group of graph objects upward in the hierarchy	429
Figure 14-5 Comparison of two DCAs.....	431
Figure 14-6 Relation between first axis scores of one analysis and scores on first three axes of another analysis, visualised using a loess smoother	432
Figure 14-7 Selecting only active samples	432
Figure 14-8 Creating XY diagram with multiple response variables	433
Figure 14-9 Data Attribute Plot dialog	434
Figure 14-10 Contour-based attribute plots displaying the patterns of first axis scores of the other DCA.....	434
Figure 14-11 Biplot of well-fitting species and environmental variables.....	436
Figure 14-12 Species Response Curves with GAM	437
Figure 14-13 GAM Options for species response curves	437
Figure 14-14 Report on regression model selection	438

Figure 14-15 Fitted GAM model summary	438
Figure 14-16 Species response curves fitted using generalized additive models.....	439
Figure 14-17 GLM options for species response curves	440
Figure 14-18 Species response curves fitted using generalized linear models.....	440
Figure 14-19 Limiting species presence in diagrams	442
Figure 14-20 Changing scaling of ordination scores.....	443
Figure 14-21 Species - env. variables biplot with first and third CCA axis.....	443
Figure 14-22 Triplot diagram (including species, environmental variables, and samples) for a redundancy analysis of dune meadow data.....	444
Figure 14-23 Dialog for management of sample classifications	445
Figure 14-24 Classify From Data dialog	446
Figure 14-25 Manual Classification dialog	446
Figure 14-26 Samples and environmental variables biplot with sample symbols appearance coding the type of farming.....	447
Figure 14-27 Visual Attribute Settings for sample symbols of first class.....	448
Figure 14-28 Final look of the biplot with samples and environmental variables	449
Figure 14-29 Confirm Class Boundaries dialog	450
Figure 14-30 Pie symbols plot visualising distribution of species over classes of samples with different soil moisture.....	451
Figure 14-31 Creating a diagram with spatial positions of samples.....	452
Figure 14-32 Spatial coordinates of samples.....	452
Figure 14-33 Symbol attribute plot showing the pattern of CCA Axis 1 coordinates in space	453
Figure 14-34 XY(Z) Plot settings for creating a symbol attribute plot	453
Figure 14-35 XY(Z) Plot settings for creating a contour-based attribute plot	454
Figure 14-36 Contour-based attribute plot displaying the change of sample scores on the first CCA ordination axis throughout the sampling area	454
Figure 14-37 Defining new series and changing position of item in a series.....	455
Figure 14-38 Legend options for a plot with series collections.....	456
Figure 14-39 XY(Z) diagram options needed to produce the following graph.....	456
Figure 14-40 Change of CCA Axis 1 scores with site elevation, for individual series...457	457
Figure 14-41 Suppression of samples not in transect line 3	457
Figure 14-42 Final XY diagram	458
Figure 14-43 Select suppressed env. variables dialog box.....	459
Figure 14-44 T-Values Biplot Options dialog.....	459
Figure 14-45 T-values Biplot diagram	460
Figure 14-46 Changing range of horizontal axis for T-values biplot.....	461
Figure 14-47 Create PRC Scores dialog box.....	462
Figure 14-48 Copying labels of environmental variables onto Clipboard	462
Figure 14-49 Adding <i>Time</i> variable in spreadsheet program	463
Figure 14-50 Importing the <i>Time</i> variable back into CanoDraw.....	463
Figure 14-51 Classifying environmental variables by treatment.....	464
Figure 14-52 Selecting legend options	464

Figure 14-53 Creating the XY diagram with first PRC	465
Figure 14-54 PRC diagram with PRC1	465
Figure 14-55 Species scores for the principal response curve diagram.....	466

22. List of Tables

Table 3.1 Terminology used in CANOCO, with commonly used synonyms.	34
Table 3.2 Expected mean squares.	54
Table 4.1 Command-line options of CANOIMP.EXE compared with WCanImp.	68
Table 4.2 Full format data file with 3 samples and 11 variables (species or environmental variables).	72
Table 4.3 The environmental data of the Dune meadow data in full format. The file is named 'DUNEENV.DTA'	74
Table 4.4 Condensed format data with 3 samples and 11 variables. Same data as Table 4.2.	76
Table 4.5 The species data of the Dune Meadow data in condensed format. The file is named 'DUNE_SPE.DTA'.	76
Table 4.6 The environmental data of the Dune meadow data in condensed format.	77
Table 4.7 Free format data file with 3 samples and 11 variables (species or environmental variables).	79
Table 4.8 Another data file in free format (same data as Table 4.7).	80
Table 6.1 First part of the log-window of a canonical correspondence analysis on the Dune Meadow data.	116
Table 6.2 Codes for the options for the scaling of ordination scores in unimodal methods (CA, CCA and DCCA).	117
Table 6.3 Codes for the options for the scaling of ordination scores in linear methods (PCA and RDA).	117
Table 6.4 Regression diagnostics.	120
Table 6.5 Correlations among environmental variables and ordination axes.	121
Table 6.6 Means, standard deviations and inflation factors of environmental variables.	122
Table 6.7 Summary of a CCA of the Dune Meadow data.	123
Table 6.8 Summary of a RDA of the Dune Meadow data.	124
Table 6.9 Summary of a DCA with detrending-by-segments (with interpretation by the environmental variables).	125
Table 6.10 Summary of a partial CCA of the Dune Meadow data.	126
Table 6.11 Summary of the global permutation test of the relation between species and environment in the Dune Meadow data using CCA.	128
Table 6.12 Monte Carlo permutations to test the significance of the relation between species and environment in the Dune Meadow data using CCA.	128
Table 6.13 Blocks defined by Management type in the Dune meadow data.	129
Table 6.14 The second block of a split-plot design containing 6 whole-plots with 4 split-plots each.	129
Table 6.15 Step 1 in manual forward selection of the Dune Meadow data: the marginal effects of the environmental variables.	130

Table 6.16 Step 2 in a manual forward selection of the Dune Meadow data with Moisture already selected.....	130
Table 6.17 Step 3 of a manual selection of the Dune Meadow data after Moisture and Manure have been selected.....	131
Table 6.18 A significance test in forward selection.....	131
Table 6.19 Heading of each table in linear methods.....	133
Table 6.20 Heading of each table in unimodal methods, followed by a table of species scores.....	133
Table 6.21 Heading of each table with detrending-by-segments.....	133
Table 6.22 The order of tables in the solution file with their codes and symbols.....	133
Table 6.23 Heading of a table for supplementary environmental variables.....	134
Table 6.24 Notation for input data and eigenvalues.....	135
Table 6.25 Transition formulae and scaling in PCA.....	136
Table 6.26 Eigenvalue equations and scaling in PCA.....	138
Table 6.27 Transition formulae and scaling in RDA.....	140
Table 6.28 Eigenvalue equations and scaling in RDA.....	141
Table 6.29 The environmental biplot scores $\{c_j^*\}$ in PCA and RDA.....	142
Table 6.30 The centroid scores $\{c_j^+\}$ of environmental classes in PCA and RDA.....	143
Table 6.31 Canonical coefficients $\{c_j\}$ in RDA.....	144
Table 6.32 Transition formulae and Hill's scaling in CA and DCA-POL.....	146
Table 6.33 Transition formulae and the biplot scaling in CA and DCA-POL.....	147
Table 6.34 Eigenvalue equations and the biplot scaling in CA.....	148
Table 6.35 Transition formulae and Hill's scaling in CCA and DCCA-POL.....	150
Table 6.36 Transition formulae and the biplot scaling in CCA and DCCA-POL.....	151
Table 6.37 Eigenvalue equations and the biplot scaling in CCA.....	153
Table 6.38 The environmental biplot scores $\{c_j^*\}$ in CA, CCA and D(C)CA-POL.....	154
Table 6.39 The centroid scores $\{c_j^+\}$ of environmental classes in CA, CCA and D(C)CA-POL.....	155
Table 6.40 Some transition formulae in DCCA with detrending by segments.....	156
Table 6.41 The environmental biplot scores $\{c_j^*\}$ in DCA and DCCA with detrending by segments.....	156
Table 6.42 The centroid scores $\{c_j^+\}$ of environmental classes in DCA and DCCA with detrending by segments.....	156
Table 6.43 Species scores $\{b_k\}$ in linear methods.....	159
Table 6.44 Species scores $\{u_k\}$ in unimodal methods.....	160
Table 6.45 Species scores $\{u_k\}$ in DCA and DCCA with non-linear rescaling (default in detrending-by-segments).....	160
Table 6.46 Sample scores $\{x_i^*\}$ that are derived from the species scores in linear methods.....	162

Table 6.47 Regression coefficients $\{c_j\}$ of standardized environmental variables for each of the ordination axes.....	164
Table 6.48 t-Values of the regression coefficients $\{c_j\}$ of Table 6.47.....	166
Table 6.49 Sample scores $\{x'_i\}$ that are derived from the environmental variables.....	167
Table 6.50 Inter-set correlation of environmental variables with the ordination axes. ...	168
Table 6.51 Biplot scores $\{c_j^*\}$ of environmental variables.	171
Table 6.52 Centroids $\{c_j^+\}$ of environmental variables in the ordination diagram.....	172
Table 6.53 Centroid scores for supplementary environmental variables.....	173
Table 6.54 Cumulative fit per species as fraction of variance of species.	176
Table 6.55 Squared residual length per sample.	177
Table 6.56 Species tolerances.	178
Table 6.57 Sample heterogeneity.....	179
Table 6.58 t-value biplot: scores for response variables.....	181
Table 6.59 t-value biplot: scores for predictors.	181
Table 6.60 Environment-by-species table in unimodal methods.....	183
Table 7.1 Default CANOCO.INI file in Canoco version 4.0.....	187
Table 7.2 Variants of PCA (also available in RDA if $Q_{42} = 1$ or 3).....	206
Table 7.3 Example permutation file: three permutations of the numbers 1, ..., 20.....	216
Table 7.4 Example CANOCO.CON file with added question numbers.....	219
Table 8.1 Naming convention of files in the examples.	222
Table 8.2 List of examples in Unimodal models (n.a. = not available).....	224
Table 8.3 Weighted averages of hunting spiders with respect to the six standardized environmental variables.....	232
Table 8.4 Forward selection of water chemistry and soil type variables to determine their importance in explaining the occurrence of water plants in fresh-water dykes.....	236
Table 8.5 Species coordinates of the t-value biplot (StBi:).	257
Table 8.6 Environmental coordinates for the t-value biplot (EtBi:).	257
Table 8.7 Variance explained by management type and soil characteristics in the dune meadow data.	260
Table 8.8 Variance decomposition of the effect of management and soil on dune meadow vegetation.....	261
Table 8.9 Experimental design with blocks and treatments coded in three different ways.....	263
Table 8.10 Analysis of variance table of all species simultaneously for experiment E40 obtained by RDA on log-transformed cover percentage data. (df = degrees of freedom, total SS = sum of squares totaled across species, F = F-ratio, P = Monte Carlo significance level, 199 permutations, project = name of project from which results are taken).	266
Table 8.11 Results of the automatic forward selection with line permutation tests for transect 1 (project line1.con) using 9999 permutations for each test.	271
Table 8.12 The variables in the file design.dta.....	276

Table 8.13 Data tables used as input files for the PRC analysis in CANOCO: species.dta (columns S1-S6) and design.dta (columns C, L H and W0 - W3).....	287
Table 8.14 Output of CANOCO for obtaining the PRC's.....	288
Table 8.15 Layout of the N*P experiment E40: 4 blocks with each 4 N-levels by 2 P-levels.....	293
Table 8.16 Multiple regression of y on x1 - x5: standardized regression coefficients (Regr:) and associated t-ratios (Tval:) as given by CANOCO.....	301
Table 11-1 Contents of individual window types.....	317
Table 11-2 Graph object types in CanoDraw graphs.....	319
Table 12-1 Transformation of fonts used by CanoDraw (TrueType™ fonts) into Adobe PostScript™ fonts. Typeface variants in parentheses correspond to italics, bold, or bold-italics font styles, respectively.....	327
Table 12-2 Recommended choices for the <i>Distribution</i> field in GLM Options or GAM Options property pages.....	342
Table 12-3 Attributes for individual graph object types, checked for identity in <i>Select Suchlike</i> command.....	405
Table 16.1 Dune meadow vegetation data of the island of Terschelling, The Netherlands.....	476
Table 16.2 Dune meadow environmental data of the island of Terschelling, The Netherlands.....	477

