DATA ANALYSIS IN COMMUNITY AND LANDSCAPE ECOLOGY

Edited by

R. H. G. JONGMAN, C. J. F. TER BRAAK &
O. F. R. VAN TONGEREN

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521475747

First published by Pudoc (Wageningen) 1995
New edition with corrections published by Cambridge University Press 1995
Tenth printing 2005

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
Data analysis in community and landscape ecology/R.H.G. Jongman,
C.J.F. ter Braak, and O.F.R. van Tongeren, editors. – New ed.
p. cm.
Includes bibliographical references (p.) and index.
ISBN 0 521 47574 0 (pbk.)
1. Biotic communities–Research–Methodology. 2. Landscape
ecology–Research–Methodology. 3. Biotic communities–Research–
Statistical methods. 4. Landscape ecology–Research–Statistical
methods. 5. Biotic communities–Research–Data processing.
6. Landscape ecology–Research–Data processing. I. Jongman, R. H. G.
II. Braak, C. J. F. ter. III. Van Tongeren, O. F. R.
QH541.2.D365 1995
574.5′0285–dc20 94–31639 CIP

This reprint is authorized by the original publisher and
copyright holder, Centre for Agricultural Publishing and
Documentation (Pudoc), Wageningen 1987.

ISBN 978-0-521-47574-7 paperback

Transferred to digital printing 2007

Dune Meadow Data

In this book the same set of vegetation data will be used in the chapters on ordination and cluster analysis. This set of data stems from a research project on the Dutch island of Terschelling (Batterink & Wijffels 1983). The objective of this project was to detect a possible relation between vegetation and management in dune meadows. Sampling was done in 1982. Data collection was done by the Braun-Blanquet method; the data are recorded according to the ordinal scale of van der Maarel (1979b). In each parcel usually one site was selected; only in cases of great variability within the parcel were more sites used to describe the parcel. The sites were selected by throwing an object into a parcel. The point where the object landed was fixed as one corner of the site. The sites measure 2x2 m^2. The sites were considered to be representative of the whole parcel. From the total of 80 sites, 20 have been selected to be used in this book (Table 0.1). This selection expresses the variation in the complete set of data. The names of the species conform with the nomenclature in van der Meijden et al. (1983) and Tutin et al. (1964-1980).

Data on the environment and land-use that were sampled in this project are (Table 0.2):

- thickness of the A1 horizon
- moisture content of the soil
- grassland management type
- agricultural grassland use
- quantity of manure applied.

The thickness of the A1 horizon was measured in centimetres and it can therefore be handled as a quantitative variable. In the dunes, shifting sand is a normal phenomenon. Frequently, young developed soils are dusted over by sand, so that soil development restarts. This may result in soils with several A1 horizons on top of each other. Where this had occurred only the A1 horizon of the top soil layer was measured.

The moisture content of the soil was divided into five ordered classes.It is therefore an ordinal variable.

Four types of grassland management have been distinguished:

- standard farming (SF)
- biological farming (BF)
- hobby-farming (HF)
- nature conservation management (NM).

The grasslands can be used in three ways: as hayfields, as pasture or a combination of these (intermediate). Both variables are nominal but sometimes the use of the

grassland is handled as an ordinal variable (Subsection 2.3.1). Therefore a ranking order has been made from hay production (1), through intermediate (2) to grazing (3).
The amount of manuring is expressed in five classes (0-4). It is therefore an ordinal variable.

All ordinal variables are treated as if they are quantitative, which means that the scores of the manure classes, for example, are handled in the same way as the scores of the A1 horizon. The numerical scores of the ordinal variables are given in Table 0.2. There are two values missing in Table 0.2 . Some computer programs cannot handle missing values, so the mean value of the corresponding variable has been inserted. The two data values are indicated by an asterisk.

Table 0.1. Dune Meadow Data. Unordered table that contains 20 relevées (columns) and 30 species (rows). The right-hand column gives the abbreviation of the species names listed in the left-hand column; these abbreviations will be used throughout the book in other tables and figures. The species scores are according to the scale of van der Maarel (1979b).

		00000000011111111112 12345678901234567890	
1	Achillea millefolium	13..222..4.....2...	Ach mil
2	Agrostis stolonifera	..48...43..45447...5	Agr sto
3	Aira praecox	.....2.3.	Air pra
4	Alopecurus geniculatus	.272...53..85..4....	Alo gen
5	Anthoxanthum odoratum	....432..4.....4.4.	Ant odo
6	Bellis perennis	.3222....2.....2..	Bel per
7	Bromus hordaceus	.4.32.2..4.....	Bro hor
8	Chenopodium album	.....1.....	Che alb
9	Cirsium arvense	...2.....	Cir arv
10	Eleocharis palustris	.....4.....458...4	Ele pal
11	Elymus repens	44444...6.....	Ely rep
12	Empetrum nigrum	.....2.	Emp nig
13	Hypochaeris radicata	.....2.....2.5.	Hyp rad
14	Juncus articulatus	.....44.....33...4	Jun art
15	Juncus bufonius	.....2.4..43.....	Jun buf
16	Leontodon autumnalis	.52233332352222.2562	Leo aut
17	Lolium perenne	75652664267.....2..	Lol per
18	Plantago lanceolata	....555..33.....23..	Pla lan
19	Poa pratensis	44542344444.2...13..	Poa pra
20	Poa trivialis	2765645454.49..2....	Poa tri
21	Potentilla palustris	.....22.....	Pot pal
22	Ranunculus flammula	.....2....2222...4	Ran fla
23	Rumex acetosa	....563.2..2.....	Rum ace
24	Sagina procumbens	...5...22.242.....3.	Sag pro
25	Salix repens	335	Sal rep
26	Trifolium pratense	....252.....	Tri pra
27	Trifolium repens	.52125223633261..22.	Tri rep
28	Vicia lathyroides	.....12.....1..	Vic lat
29	Brachythecium rutabulum	..2226222244..44.634	Bra rut
30	Calliergonella cuspidata	.....4.3...3	Cal cus

Table 0.2. Environmental data (columns) of 20 relevées (rows) from the dune meadows. The scores are explained in the description of the Dune Meadow research project above; asterisk denotes mean value of variable.

Sample number	A1 horizon	Moisture class	Management type	Use	Manure class
1	2.8	1	SF	2	4
2	3.5	1	BF	2	2
3	4.3	2	SF	2	4
4	4.2	2	SF	2	4
5	6.3	1	HF	1	2
6	4.3	1	HF	2	2
7	2.8	1	HF	3	3
8	4.2	5	HF	3	3
9	3.7	4	HF	1	1
10	3.3	2	BF	1	1
11	3.5	1	BF	3	1
12	5.8	4	SF	2	2*
13	6.0	5	SF	2	3
14	9.3	5	NM	3	0
15	11.5	5	NM	2	0
16	5.7	5	SF	3	3
17	4.0	2	NM	1	0
18	4.6*	1	NM	1	0
19	3.7	5	NM	1	0
20	3.5	5	NM	1	0

5 Ordination

C.J.F. ter Braak

5.1 Introduction

5.1.1 Aim and usage

Ordination is the collective term for multivariate techniques that arrange sites along axes on the basis of data on species composition. The term ordination was introduced by Goodall (1954) and, in this sense, stems from the German 'Ordnung', which was used by Ramensky (1930) to describe this approach.

The result of ordination in two dimensions (two axes) is a diagram in which sites are represented by points in two-dimensional space. The aim of ordination is to arrange the points such that points that are close together correspond to sites that are similar in species composition, and points that are far apart correspond to sites that are dissimilar in species composition. The diagram is a graphical summary of data, as in Figure 5.1, which shows three groups of similar sites. Ordination includes what psychologists and statisticians refer to as multidimensional scaling, component analysis, factor analysis and latent-structure analysis.

Figure 5.1 also shows how ordination is used in ecological research. Ecosystems are complex: they consist of many interacting biotic and abiotic components. The way in which abiotic environmental variables influence biotic composition is often explored in the following way. First, one samples a set of sites and records which species occur there and in what quantity (abundance). Since the number of species is usually large, one then uses ordination to summarize and arrange the data in an ordination diagram, which is then interpreted in the light of whatever is known about the environment at the sites. If explicit environmental data are lacking, this interpretation is done in an informal way; if environmental data have been collected, in a formal way (Figure 5.1). This two-step approach is indirect gradient analysis in the sense used by Whittaker (1967). By contrast, direct gradient analysis is impossible without explicit environmental data. In direct gradient analysis, one is interested from the beginning in particular environmental variables, i.e. either in their influence on the species as in regression analysis (Chapter 3) or in their values at particular sites as in calibration (Chapter 4).

Indirect gradient analysis has the following advantages over direct gradient analysis. Firstly, species compositions are easy to determine, because species are usually clearly distinguishable entities. By contrast, environmental conditions are difficult to characterize exhaustively. There are many environmental variables and even more ways of measuring them, and one is often uncertain of which variables the species react to. Species composition may therefore be a more informative

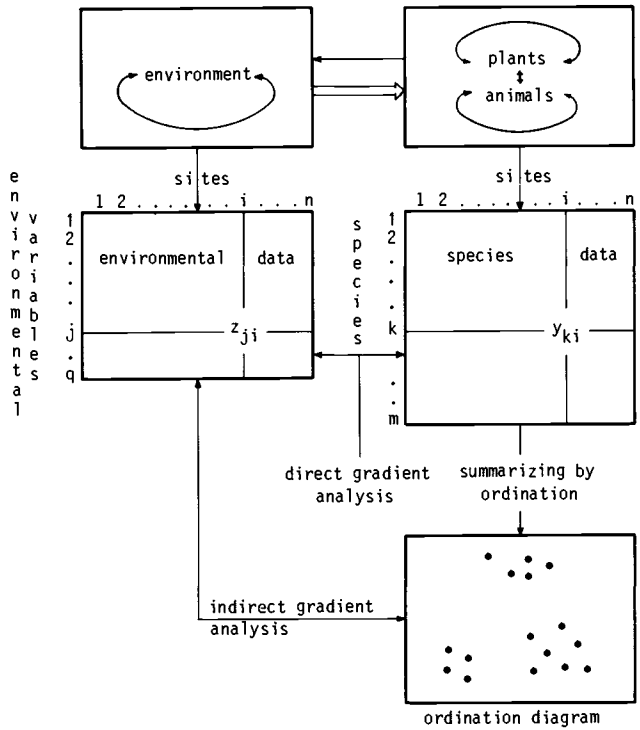


Figure 5.1 Outline of the role of ordination in community ecology, showing the typical format of data sets obtained by sampling ecosystems and their analysis by direct gradient and indirect gradient analysis. Also shown is the notation used in Chapter 5. Point of site in the ordination diagram (•).

indicator of environment than any given set of measured environmental variables. Ordination can help to show whether important environmental variables have been overlooked: an important variable has definitely been missed if their is no relation between the mutual positions of the sites in the ordination diagram and the measured environmental variables.

Secondly, the actual occurrence of any individual species may be too unpredictable to discover the relation of its occurrence to environmental conditions by direct means (Chapter 3) and therefore more general patterns of coincidence of several species are of greater use in detecting species-environment relations.

Thirdly, for example in landscape planning, interest may from the onset be focused more on the question of which combinations of species can occur, and less on the behaviour of particular species. Regression analysis of single species then provides too detailed an account of the relations between species and

environment. The ordination approach is less elaborate and gives a global picture, but – one hopes – with sufficient detail for the purpose in hand.

Between regression analysis and ordination (in the strict sense) stand the canonical ordination techniques. They are ordination techniques converted into multivariate direct gradient analysis techniques; they deal simultaneously with many species and many environmental variables. The aim of canonical ordination is to detect the main pattern in the relations between the species and the observed environment.

5.1.2 Data approximation and response models in ordination

Ordination techniques can be viewed in two ways (Prentice 1977). According to one view, the aim of ordination is to summarize multivariate data in a convenient way in scatter diagrams. Ordination is then considered as a technique for matrix approximation (as the data are usually presented in the two-way layout of a matrix). A second, more ambitious, view assumes from the beginning that there is an underlying (or latent) structure in the data, i.e. that the occurrences of all species under consideration are determined by a few unknown environmental variables (latent variables) according to a simple response model (Chapter 3). Ordination in this view aims to recover that underlying structure. This is illustrated in Figure 5.2 for a single latent variable. In Figure 5.2a, the relations of two species, A and B, with the latent variable are rectilinear. In Figure 5.2c they are unimodal. We now record species abundance values at several sites and plot the abundance of Species A against that of Species B. If relations with the latent variable were rectilinear, we would obtain a straight line in the plot of Species B against Species A (Figure 5.2b), but if relations were unimodal, we would obtain a complicated curve (Figure 5.2d). The ordination problem of indirect gradient analysis is to infer about the relations with the latent variable (Figures 5.2a,c) from the species data only (Figure 5.2b,d). From the second viewpoint, ordination is like regression analysis, but with the major difference that in ordination the explanatory variables are not known environmental variables, but 'theoretical' variables. These variables, the latent variables, are constructed in such a way that they best explain the species data. As in regression, each species thus constitutes a response variable, but in ordination these response variables are analysed simultaneously. (The distinction between these two views of ordination is not clear-cut, however. Matrix approximation implicitly assumes some structure in the data by the mere way the data are approximated. If the data structure is quite different from the assumed structure, the approximation is inefficient and fails.)

The ordination techniques that are most popular with community ecologists, are principal components analysis (PCA), correspondence analysis (CA), and techniques related to CA, such as weighted averaging and detrended correspondence analysis. Our introduction to PCA and CA will make clear that PCA and CA are suitable to detect different types of underlying data structure. PCA relates to a linear response model in which the abundance of any species either increases or decreases with the value of each of the latent environmental variables (Figure 5.2a). By contrast, CA is related, though in a less unequivocal way, to a unimodal response model (Figure 5.2c). In this model, any species occurs in a limited range

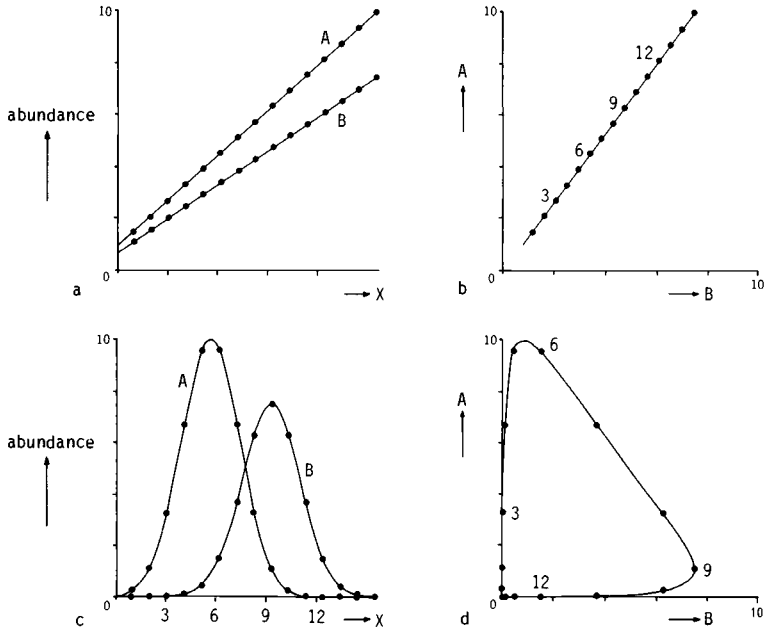


Figure 5.2 Response curves for two species A and B against a latent variable x (a, c) and the expected abundances of the species plotted against each other (b, d), for the straight line model (a, b) and a unimodal model (c, d). The numbers refer to sites with a particular value for x. The ordination problem is to make inferences about the relations in Figures a and c from species data plotted in Figures b and d.

of values of each of the latent variables. PCA and CA both provide simultaneously an ordination for the sites and an ordination for the species. The two ordinations may be plotted in the same diagram to yield 'joint plots' of site and species points, but the interpretation of the species points is different between PCA and CA.

PCA and CA operate directly on the species data. By contrast, multidimensional scaling is a class of ordination techniques that operate on a table of dissimilarity values between sites. To apply these techniques, we must therefore first choose an appropriate dissimilarity coefficient to express the dissimilarity in species composition between any two sites (Subsection 6.2.2). After choosing one, we can calculate the dissimilarity values of all pairs of sites required as input for multidimensional scaling. CA and PCA may also be considered as multidimensional scaling techniques, but ones that use a particular dissimilarity coefficient.

5.1.3 Outline of Chapter 5

Section 5.2 introduces CA and related techniques and Section 5.3 PCA. Section 5.4 discusses methods of interpreting ordination diagrams with external (environmental) data. It is also a preparation for canonical ordination (Section 5.5). After a discussion of multidimensional scaling (Section 5.6), Section 5.7 evaluates the advantages and disadvantages of the various ordination techniques and compares them with regression analysis and calibration. After the bibliographic notes (Section 5.8) comes an appendix (Section 5.9) that summarizes the ordination methods described in terms of matrix algebra.

5.2 Correspondence analysis (CA) and detrended correspondence analysis (DCA)

5.2.1 From weighted averaging to correspondence analysis

Correspondence analysis (CA) is an extension of the method of weighted averaging used in the direct gradient analysis of Whittaker (1967) (Section 3.7). Here we describe the principles in words; the mathematical equations will be given in Subsection 5.2.2.

Whittaker, among others, observed that species commonly show bell-shaped response curves with respect to environmental gradients. For example, a plant species may prefer a particular soil moisture content, and not grow at all in places where the soil is either too dry or too wet. In the artificial example shown in Figure 5.3a, Species A prefers drier conditions than Species E, and the Species B, C and D are intermediate. Each of the species is therefore largely confined to a specific interval of moisture values. Figure 5.3a also shows presence–absence data for Species D: the species is present at four of the sites.

We now develop a measure of how well moisture explains the species data. From the data, we can obtain a first indication of where a species occurs along the moisture gradient by taking the average of the moisture values of the sites in which the species is present. This average is an estimate of the optimum of the species (the value most preferred), though not an ideal one (Section 3.7). The average is here called the species score. The arrows in Figure 5.3a point to the species scores so calculated for the five species. As a measure of how well moisture explains the species data, we use the dispersion ('spread') of the species scores. If the dispersion is large, moisture neatly separates the species curves and moisture explains the species data well. If the dispersion is small, then moisture explains less. To compare the explanatory power of different environmental variables, each environmental variable must first be standardized; for example by subtracting its mean and dividing by its standard deviation.

Suppose that moisture is the 'best' single environmental variable measured in the artificial example. We might now wish to know whether we could in theory have measured a variable that explains the data still better. CA is now the technique that constructs the theoretical variable that best explains the species data. CA does so by choosing the best values for the sites, i.e. values that maximize the dispersion of the species scores (Figure 5.3b). The variable shown gives a larger

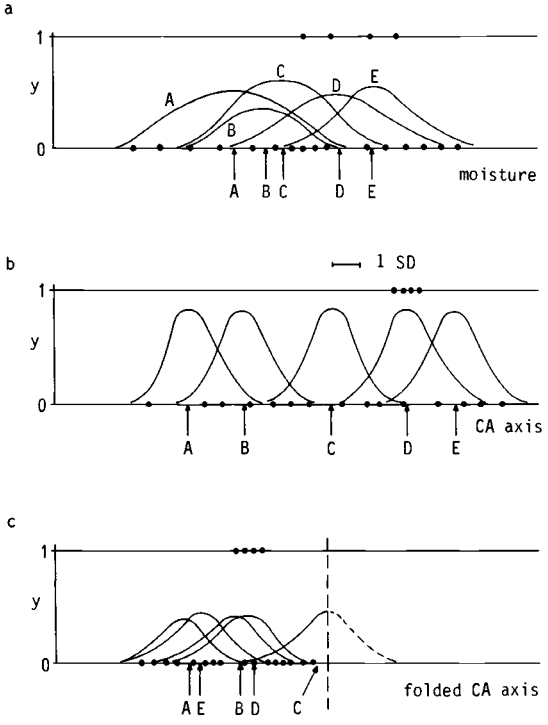


Figure 5.3 Artificial example of unimodal response curves of five species (A-E) with respect to standardized variables, showing different degrees of separation of the species curves. a: Moisture. b: First axis of CA. c: First axis of CA folded in this middle and the response curves of the species lowered by a factor of about 2. Sites are shown as dots at $y = 1$ if Species D is present and at $y = 0$ if Species D is absent. For further explanation, see Subsections 5.2.1 and 5.2.3.

dispersion than moisture; and consequently the curves in Figure 5.3b are narrower, and the presences of Species D are closer together than in Figure 5.3a.

The theoretical variable constructed by CA is termed the first ordination axis of CA or, briefly, the first CA axis; its values are the site scores on the first CA axis.

A second and further CA axes can also be constructed; they also maximize the dispersion of the species scores but subject to the constraint of being uncorrelated with previous CA axes. The constraint is intended to ensure that new information is expressed on the later axes. In practice, we want only a few axes in the hope that they represent most of the variation in the species data.

So we do not need environmental data to apply CA. CA 'extracts' the ordination

axes from the species data alone. CA can be applied not only to presence–absence data, but also to abundance data; for the species scores, we then simply take a weighted average of the values of the sites (Equation 3.28).

5.2.2 Two-way weighted averaging algorithm

Hill (1973) introduced CA into ecology by the algorithm of reciprocal averaging. This algorithm shows once more that CA is an extension of the method of weighted averaging.

If we have measured an environmental variable and recorded the species composition, we can estimate for each species its optimum or indicator value by averaging the values of the environmental variable over the sites in which the species occurs, and can use the averages so obtained to rearrange the species (Table 3.9). If the species show bell-shaped curves against the environmental variable, the rearranged table will have a diagonal structure, at least if the optima of the curves differ between the species (Table 3.9). Conversely, if the indicator values of species are known, the environmental variable at a site can be estimated from the species that it contains, by averaging the indicator values of these species (Section 4.3) and sites can be arranged in order of these averages. But, these methods are only helpful in showing a clear structure in the data if we know in advance which environmental variable determines the occurrences of the species. If this is not known in advance, the idea of Hill (1973) was to discover the 'underlying environmental gradient' by applying this averaging process both ways in an iterative fashion, starting from arbitrary initial values for sites or from arbitrary initial (indicator) values for species. It can be shown mathematically that this iteration process eventually converges to a set of values for sites and species that do not depend on the initial values. These values are the site and species scores of the first CA axis.

We illustrate now the process of reciprocal averaging. For abundance data, it is rather a process of two-way weighted averaging. Table 5.1a shows the Dune Meadow Data (Table 0.1), arranged in arbitrary order. We take as initial values for the sites the numbers 1 to 20, as printed vertically below Table 5.1a. As before, we shall use the word 'score', instead of 'value'. From the site scores, we derive species scores by calculating the weighted average of the site scores for each species. If we denote the abundance of species k at site i by y_{ki}, the score of site i by x_i and the score of species k by u_k, then the score of species k becomes the weighted average of site scores (Section 3.7)

$$u_k = \Sigma_{i=1}^{n} y_{ki}\, x_i / \Sigma_{i=1}^{n} y_{ki} \qquad \text{Equation 5.1}$$

For *Achillea millefolium* in Table 5.1a, we obtain $u_1 = (1 \times 1 + 3 \times 2 + 2 \times 5 + 2 \times 6 + 2 \times 7 + 4 \times 10 + 2 \times 17)/(1 + 3 + 2 + 2 + 2 + 4 + 2) = 117/16 = 7.31$. The species scores thus obtained are also shown in Table 5.1a. From these species scores, we derive new site scores by calculating for each site the weighted average of the species scores, i.e.

Table 5.1a

Species k		Sites (i) 00000000011111111112 12345678901234567890	u_k
1	Ach mil	13 222 4 2	7.31
2	Agr sto	48 43 45447 5	11.33
3	Air pra	2 3	18.20
4	Alo gen	272 53 85 4	9.03
5	Ant odo	432 4 4 4	11.24
6	Bel per	3222 2 2	6.62
7	Bro hor	4 32 2 4	5.60
8	Che alb	1	13.00
9	Cir arv	2	4.00
10	Ele pal	4 458 4	14.84
11	Ely rep	44444 6	4.38
12	Emp nig	2	19.00
13	Hyp rad	2 2 5	16.78
14	Jun art	44 33 4	13.39
15	Jun buf	2 4 43	10.54
16	Leo aut	52233332352222 2562	10.94
17	Lol per	75652664267 2	6.31
18	Pla lan	555 33 23	9.27
19	Poa pra	44542344444 2 13	7.25
20	Poa tri	2765645454 49 2	7.25
21	Pot pal	22	14.50
22	Ran fla	2 2222 4	15.14
23	Rum ace	563 2 2	6.89
24	Sag pro	5 22 242 3	10.35
25	Sal rep	335	19.18
26	Tri pra	252	6.00
27	Tri rep	52125223633261 22	9.47
28	Vic lat	12 1	12.50
29	Bra rut	2226222244 44 634	12.02
30	Cal cus	4 3 3	16.40
x_i		11111111112 12345678901234567890	

Table 5.1 Two-way weighted averaging algorithm of CA applied to the Dune Meadow Data presented in a preliminary section of this book. The site numbers and site scores are printed vertically. a: Original data table with at the bottom the initial site scores. b: Species and sites rearranged in order of their scores obtained after one cycle of two-way weighted averaging. c: Species and sites arranged in order of their final scores (CA scores). Note the minus signs in the site scores; for example, the score of Site 17 is -1.46.

Table 5.1b

Species	Sites (i)	
	00000010011101111112	
k	12534706931288764590	u_k
9 Cir arv	2	4.00
11 Ely rep	44444 6	4.38
7 Bro hor	42 324	5.60
26 Tri pra	2 2 5	6.00
17 Lol per	752656662 7 42	6.31
6 Bel per	3222 2 2	6.62
23 Rum ace	5 3 62 2	6.89
19 Poa pra	44254443424 431	7.25
20 Poa tri	2766554459 44 2	7.25
1 Ach mil	132 242 2	7.31
4 Alo gen	2 72 35 85 4	9.03
18 Pla lan	5 535 3 32	9.27
27 Tri rep	5221265323322 612	9.47
24 Sag pro	5 22242 3	10.35
15 Jun buf	2 43 4	10.54
16 Leo aut	53223332252352 2262	10.94
5 Ant odo	4 243 4 4	11.24
2 Agr sto	48 35 44 744 5	11.33
29 Bra rut	2222262 4426 4 434	12.02
28 Vic lat	1 2 1	12.50
8 Che alb	1	13.00
14 Jun art	4 4 3 3 4	13.39
21 Pot pal	22	14.50
10 Ele pal	4 845 4	14.84
22 Ran fla	2 2 222 4	15.14
30 Cal cus	34 3	16.40
13 Hyp rad	2 2 5	16.78
3 Air pra	2 3	18.20
12 Emp nig	2	19.00
25 Sal rep	3 35	19.18
	11111111	
x_i	67788888899900122234	
	.....	
	24901134867737868983	
	56348878043849124796	

Table 5.1c

Species	Sites (i)	
	10100011010001101121	
k	75076191283492385406	u_k
3 Air pra	2 3	-0.99
5 Ant odo	44423 4	-0.96
1 Ach mil	224221 3	-0.91
26 Tri pra	2 25	-0.88
13 Hyp rad	2 52	-0.84
18 Pla lan	25355 3 3	-0.84
12 Emp nig	2	-0.67
7 Bro hor	242 4 3	-0.66
23 Rum ace	5 36 22	-0.65
28 Vic lat	1 2 1	-0.62
6 Bel per	22 3222	-0.50
17 Lol per	26667 752652 4	-0.50
19 Poa pra	124434 443544 24	-0.39
11 Ely rep	4 4 4 446	-0.37
16 Leo aut	23333 6555222223222	-0.19
20 Poa tri	64542 7 655494 2	-0.18
27 Tri rep	2625 235221332216	-0.08
9 Cir arv	2	-0.06
24 Sag pro	32 52422	0.00
15 Jun buf	2 443	0.08
29 Bra rut	2226 34 62224 24 44	0.18
4 Alo gen	2 723855 4	0.40
8 Che alb	1	0.42
25 Sal rep	3 3 5	0.62
2 Agr sto	4834544457	0.93
14 Jun art	4 43 43	1.28
22 Ran fla	222242	1.56
10 Ele pal	45448	1.77
21 Pot pal	22	1.92
30 Cal cus	433	1.96
	----------	
x_i	10000000000000001112	
	.....	
	49888866631002479990	
	65876284411698262250	

$$x_i = \Sigma_{k=1}^{m} y_{ki}\, u_k / \Sigma_{k=1}^{m} y_{ki} \qquad \text{Equation 5.2}$$

For Site 1 in Table 5.1a, we obtain $x_1 = (1 \times 7.31 + 4 \times 4.38 + 7 \times 6.31 + 4 \times 7.25 + 2 \times 7.25)/(1 + 4 + 7 + 4 + 2) = 112.5/18 = 6.25$. In Table 5.1b, the species and sites are arranged in order of the scores obtained so far. The new site scores are also printed vertically underneath. There is already some diagonal structure, i.e. the occurrences of each species tend to come together along the rows. We can improve upon this structure by calculating new species scores from the site scores that we have just calculated, and so on.

A practical numerical problem with this technique is that, by taking averages, the range of the scores gets smaller and smaller. For example, we started off with a range of 19 (site scores from 1 to 20) and after one cycle the site scores have a range of 14.36 - 6.25 = 8.11 (Table 5.1b). To avoid this, either the site scores or the species scores must be rescaled. Here the site scores have been rescaled. There are several ways of doing so. A simple way is to rescale to a range from 0 to 100 by giving the site with the lowest score the value 0 and the site with the highest score the value 100 and by calculating values for the remaining sites in proportion to their scores; in the example, the rescaled scores would be obtained with the formula $(x_i - 6.25)/0.0811$.

We shall use another way in which the site scores are standardized to (weighted)

Table 5.2 Two-way weighted averaging algorithm of CA.

a: Iteration process

Step 1. Take arbitrary, but unequal, initial site scores (x_i).
Step 2. Calculate new species scores (u_k) by weighted averaging of the site scores (Equation 5.1).
Step 3. Calculate new site scores (x_i) by weighted averaging of the species scores (Equation 5.2).
Step 4. For the first axis, go to Step 5. For second and higher axes, make the site scores (x_i) uncorrelated with the previous axes by the orthogonalization procedure described below.
Step 5. Standardize the site scores (x_i). See below for the standardization procedure.
Step 6. Stop on convergence, i.e. when the new site scores are sufficiently close to the site scores of the previous cycle of the iteration; ELSE go to Step 2.

b: Orthogonalization procedure

Step 4.1. Denote the site scores of the previous axis by f_i and the trial scores of the present axis by x_i.
Step 4.2. Calculate $v = \Sigma_{i=1}^{n} y_{+i}\, x_i f_i / y_{++}$
where $y_{+i} = \Sigma_{k=1}^{m} y_{ki}$
and $y_{++} = \Sigma_{i=1}^{n} y_{+i}$.
Step 4.3 Calculate $x_{i,\text{new}} = x_{i,\text{old}} - v f_i$.
Step 4.4 Repeat Steps 4.1-4.3 for all previous axes.

c: Standardization procedure

Step 5.1 Calculate the centroid, z, of site scores (x_i) $z = \Sigma_{i=1}^{n} y_{+i}\, x_i / y_{++}$.
Step 5.2 Calculate the dispersion of the site scores $s^2 = \Sigma_{i=1}^{n} y_{+i}\, (x_i - z)^2 / y_{++}$.
Step 5.3 Calculate $x_{i,\text{new}} = (x_{i,\text{old}} - z)/s$.
Note that, upon convergence, s equals the eigenvalue.

mean 0 and variance 1 as described in Table 5.2c. If the site scores are so standardized, the dispersion of the species scores can be written as

$$\delta = \Sigma_{k=1}^{m} y_{k+} u_k^2 / y_{++} \qquad \text{Equation 5.3}$$

where
y_{k+} is the total abundance of species k
y_{++} the overall total.

The dispersion will steadily increase in each iteration cycle until, after about 10 cycles, the dispersion approaches its maximum value. At the same time, the site and species scores stabilize. The resulting scores have maximum dispersion and thus constitute the first CA axis.

If we had started from a different set of initial site scores or from a set of arbitrary species scores, the iteration process would still have resulted in the same ordination axis. In Table 5.1c, the species and sites are rearranged in order of their scores on the first CA axis and show a clear diagonal structure.

A second ordination axis can also be extracted from the species data. The need for a second axis may be illustrated in Table 5.1c; Site 1 and Site 19 lie close together along the first axis and yet differ a great deal in species composition. This difference can be expressed on a second axis. The second axis is extracted by the same iteration process, with one extra step in which the trial scores for the second axis are made uncorrelated with the scores of the first axis. This can be done by plotting in each cycle the trial site scores for the second axis against the site scores of the first axis and fitting a straight line by a (weighted) least-squares regression (the weights are y_{+i}/y_{++}). The residuals from this regression (i.e. the vertical deviations from the fitted line: Figure 3.1) are the new trial scores. They can be obtained more quickly by the orthogonalization procedure described in Table 5.2b. The iteration process would lead to the first axis again without the extra step. The intention is thus to extract information from the species data in addition to the information extracted by the first axis. In Figure 5.4, the final site scores of the second axis are plotted against those of the first axis. Site 1 and Site 19 lie far apart on the second axis, which reflects their difference in species composition. A third axis can be derived in the same way by making the scores uncorrelated with the scores of the first two axes, and so on. Table 5.2a summarizes the algorithm of two-way weighted averaging. A worked example is given in Exercise 5.1 and its solution.

In mathematics, the ordination axes of CA are termed eigenvectors (a vector is a set of values, commonly denoting a point in a multidimensional space and 'eigen' is German for 'self'). If we carry out an extra iteration cycle, the scores (values) remain the same, so the vector is transformed into itself, hence, the term eigenvector. Each eigenvector has a corresponding eigenvalue, often denoted by λ (the term is explained in Exercise 5.1.3). The eigenvalue is actually equal to the (maximized) dispersion of the species scores on the ordination axis, and is thus a measure of importance of the ordination axis. The first ordination axis has the largest eigenvalue (λ_1), the second axis the second largest eigenvalue (λ_2),

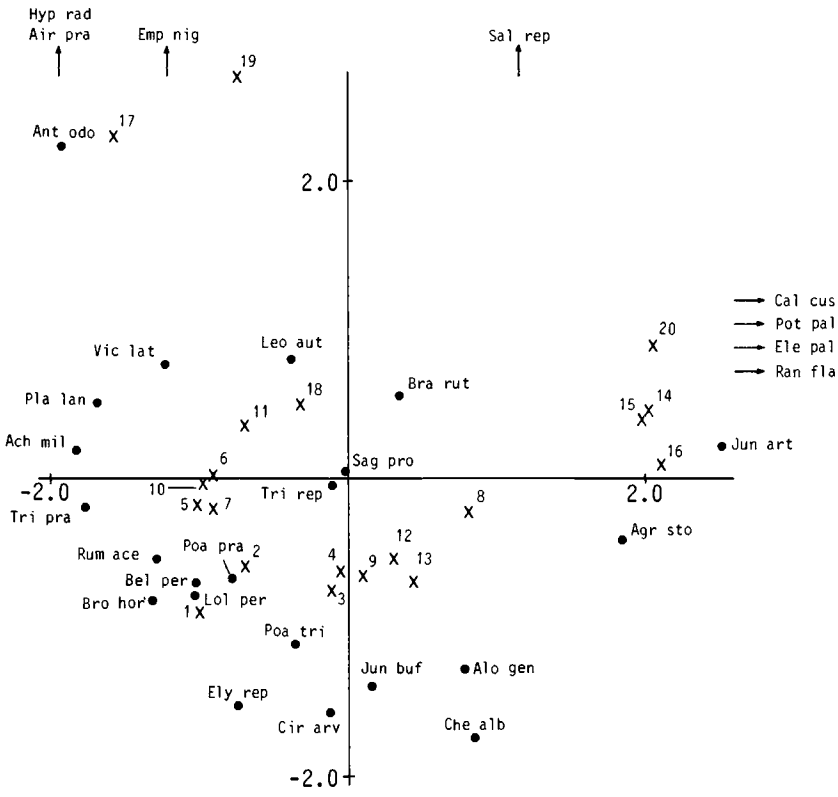


Figure 5.4 CA ordination diagram of the Dune Meadow Data in Hill's scaling. In this and the following ordination diagrams, the first axis is horizontal and the second axis vertical; the sites are represented by crosses and labelled by their number in Table 5.1; species names are abbreviated as in Table 0.1.

and so on. The eigenvalues of CA all lie between 0 and 1. Values over 0.5 often denote a good separation of the species along the axis. For the Dune Meadow Data, $\lambda_1 = 0.53$; $\lambda_2 = 0.40$; $\lambda_3 = 0.26$; $\lambda_4 = 0.17$. As λ_3 is small compared to λ_1 and λ_2, we ignore the third and higher numbered ordination axes, and expect the first two ordination axes to display the biologically relevant information (Figure 5.4).

When preparing an ordination diagram, we plot the site scores and the species scores of one ordination axis against those of another. Because ordination axes differ in importance, one would wish the scores to be spread out most along the most important axis. But our site scores do not do so, because we standardized them to variance 1 for convenience in the algorithm (Table 5.2). An attractive

standardization is obtained by requiring that the average width of the species curves is the same for each axis. As is clear from Figure 5.3b, the width of the curve for Species D is reflected in the spread among its presences along the axis. Therefore, the average curve width along an axis can be estimated from the data. For example, Hill (1979) proposed to calculate, for each species, the variance of the scores of the sites containing the species and to take the (weighted) average of the variances so obtained, i.e. Hill proposed to calculate

$$\Sigma_k \, y_{k+} \, [\Sigma_i \, y_{ki} \, (x_i - u_k)^2 / y_{k+}] / y_{++}.$$

To equalize the average curve width among different axes, we must therefore divide all scores of an axis by its average curve width (i.e. by the square root of the value obtained above). This method of standardization is used in the computer program DECORANA (Hill 1979a). Other than in Table 5.2, the program further uses the convention that site scores are weighted averages of species scores; so we must iterate Step 3 of our algorithm once more, before applying the standardization procedure just described. This scaling has already been used in preparing Figure 5.4 and we shall refer to it as Hill's scaling. A short cut to obtain Hill's scaling from the scores obtained from our algorithm is to divide the site scores after convergence by $\sqrt{(1 - \lambda)/\lambda}$ and the species scores by $\sqrt{\lambda(1 - \lambda)}$. The scores so obtained are expressed in multiples of one standard deviation (s.d.) and have the interpretation that sites that differ by 4 s.d. in score tend to have few species in common (Figure 5.3b). This use of s.d. will be discussed further in Subsection 5.2.4.

CA cannot be applied on data that contain negative values. So the data should not be centred or standardized (Subsection 2.4.4). If the abundance data of each species have a highly skew distribution with many small values and a few extremely large values, we recommend transforming them by taking logarithms: $\log_e (y_{ki} + 1)$, as in Subsection 3.3.1. By doing so, we prevent a few high values from unduly influencing the analysis. In CA, a species is implicitly weighted by its relative total abundance y_{k+}/y_{++} and, similarly, a site is weighted by y_{+i}/y_{++}. If we want to give a particular species, for example, triple its weight, we must multiply all its abundance values by 3. Sites can also be given greater or smaller weight by multiplying their abundance values by constants (ter Braak 1987b).

5.2.3 Diagonal structures: properties and faults of correspondence analysis

Table 5.3a shows artificial data in which the occurrences of species across sites appear rather chaotic and Table 5.3b shows the same data after arranging the species and sites in order of their score on the first CA axis. The data are rearranged into a perfectly diagonal table, also termed a two-way Petrie matrix. (A Petrie matrix is an incidence matrix that has a block of consecutive ones in every row; the matrix is two-way Petrie if the matrix also has a block of consecutive ones in every column, the block in the first column starting in the first row and the block of the last column ending in the last row.) For any table that permits such a rearrangement, we can discover the correct order of species and sites from

the scores of the first axis of CA. This property of CA can be generalized to quantitative data (Gifi 1981) and to (one-way) Petrie matrices (Heiser 1981; 1986). For two-way Petrie matrices with many species and sites and with about equal numbers of occurrences per species and per site, the first eigenvalue is close to 1; e.g. for Table 5.3, $\lambda_1 = 0.87$.

Note that CA does not reveal the diagonal structure if the ones and zeros are interchanged. Their role is asymmetrical, as is clear from the reciprocal averaging algorithm. The ones are important; the zeros are disregarded. Many ecologists feel the same sort of asymmetry between presences and absences of species.

The ordination of Table 5.3 illustrates two 'faults' of CA (Figure 5.5). First, the change in species composition between consecutive sites in Table 5.3, Column b is constant (one species appears; one disappears) and one would therefore wish that this constant change were reflected in equal distances between scores of neighbouring sites along the first axis. But the site scores at the ends of the first axis are closer together than those in the middle of the axis (Figure 5.5b). Secondly, the species composition is explained perfectly by the ordering of the sites and species along the first axis (Table 5.3, Column b) and the importance of the second axis should therefore be zero. However $\lambda_2 = 0.57$ and the site scores on the second axis show a quadratic relation with those on the first axis (Figure 5.5a). This fault is termed the arch effect. The term 'horseshoe' is also in use but is less appropriate, as the ends do not fold inwards in CA.

Table 5.3 CA applied to artificial data (- denotes absence). Column a: The table looks chaotic. Column b: After rearrangement of species and sites in order of their scores on the first CA axis (u_k and x_i), a two-way Petrie matrix appears: $\lambda_1 = 0.87$.

Column a		Column b		u_k
Species	Sites 1 2 3 4 5 6 7	Species	Sites 1 7 2 4 6 5 3	
A	1 - - - - - -	A	1 - - - - - -	-1.40
B	1 - - - - - 1	B	1 1 - - - - -	-1.24
C	1 1 - - - - 1	C	1 1 1 - - - -	-1.03
D	- - - 1 1 1 -	E	- 1 1 1 - - -	-0.56
E	- 1 - 1 - - 1	F	- - 1 1 1 - -	0.00
F	- 1 - 1 - 1 -	D	- - - 1 1 1 -	0.56
G	- - 1 - 1 1 -	G	- - - - 1 1 1	1.03
H	- - 1 - 1 - -	H	- - - - - 1 1	1.24
I	- - 1 - - - -	I	- - - - - - 1	1.40
			- - -	
			1 1 0 0 0 1 1	
			. . . . .	
		x_i	4 0 6 0 6 0 4	
			0 8 0 0 0 8 0	

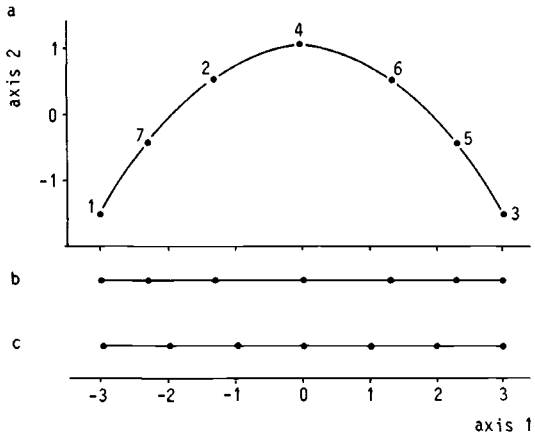


Figure 5.5 Ordination by CA of the two-way Petrie matrix of Table 5.3. a: Arch effect in the ordination diagram (Hill's scaling; sites labelled as in Table 5.3; species not shown). b: One-dimensional CA ordination (the first axis scores of Figure a, showing that sites at the ends of the axis are closer together than sites near the middle of the axis. c: One-dimensional DCA ordination, obtained by nonlinearly rescaling the first CA axis. The sites would not show variation on the second axis of DCA.

Let us now give a qualitative explanation of the arch effect. Recall that the first CA axis maximally separates the species curves by maximizing the dispersion (Equation 5.3) and that the second CA axis also tries to do so but subject to the constraint of being uncorrelated with the first axis (Subsection 5.2.1). If the first axis fully explains the species data in the way of Figure 5.3b, then a possible second axis is obtained by folding the first axis in the middle and bringing the ends together (Figure 5.3c). This folded axis has no linear correlation with the first axis. The axis so obtained separates the species curves, at least Species C from Species B and D, and these from Species A and E, and is thus a strong candidate for the second axis of CA. Commonly CA will modify this folded axis somewhat, to maximize its dispersion, but the order of the site and species scores on the second CA axis will essentially be the same as that of the folded axis. Even if there is a true second underlying gradient, CA will not take it to be the second axis if its dispersion is less than that of the modified folded first axis. The intention in constructing the second CA axis is to express new information, but CA does not succeed in doing so if the arch effect appears.

5.2.4 Detrended correspondence analysis (DCA)

Hill & Gauch (1980) developed detrended correspondence analysis (DCA) as a heuristic modification of CA, designed to correct its two major 'faults': (1) that the ends of the axes are often compressed relative to the axes middle; (2) that

the second axis frequently shows a systematic, often quadratic relation with the first axis (Figure 5.5). The major of these is the arch effect.

The arch effect is 'a mathematical artifact, corresponding to no real structure in the data' (Hill & Gauch 1980). They eliminate it by 'detrending'. Detrending is intended to ensure that, at any point along the first axis, the mean value of the site scores on the subsequent axes is about zero. To this end, the first axis is divided into a number of segments and within each segment the site scores on Axis 2 are adjusted by subtracting their mean (Figure 5.6). In the computer program DECORANA (Hill 1979a), running segments are used for this purpose. This process of detrending is built into the two-way weighted averaging algorithm, and replaces the usual orthogonalization procedure (Table 5.2). Subsequent axes are derived similarly by detrending with respect to each of the existing axes. Detrending applied to Table 5.3 gives a second eigenvalue of 0, as required.

The other fault of CA is that the site scores at the end of the first axis are often closer together than those in the middle of the axis (Figure 5.5b). Through this fault, the species curves tend to be narrower near the ends of the axis than in the middle. Hill & Gauch (1980) remedied this fault by nonlinearly rescaling the axis in such a way that the curve widths were practically equal. Hill & Gauch (1980) based their method on the tolerances of Gaussian response curves for the species, using the term standard deviation (s.d.) instead of tolerance. They noted that the variance of the optima of species present at a site (the 'within-site variance') is an estimate of the average squared tolerance of those species. Rescaling must therefore equalize the within-site variances as nearly as possible. For rescaling, the ordination axis is divided into small segments; the species ordination is expanded in segments with sites with small within-site variance and contracted in segments with sites with high within-site variance. Subsequently, the site scores are calculated by taking weighted averages of the species scores and the scores of sites and species are standardized such that the within-site variance equals 1. The tolerances of the curves of species will therefore approach 1. Hill & Gauch (1980) further define the length of the ordination axis to be the range of the site scores. This length is expressed in multiples of the standard deviation, abbreviated as s.d.

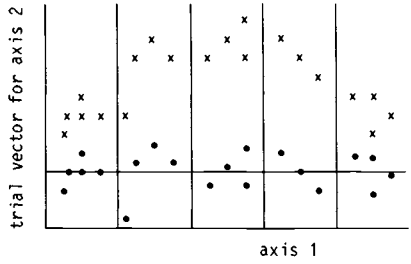


Figure 5.6 Method of detrending by segments (simplified). The crosses indicate site scores before detrending; the dots are site scores after detrending. The dots are obtained by subtracting, within each of the five segments, the mean of the trial scores of the second axis (after Hill & Gauch 1980).

The use of s.d. is attractive: a Gaussian response curve with tolerance 1 rises and falls over an interval of about 4 s.d. (Figure 3.6). Because of the rescaling, most species will about have this tolerance. Sites that differ 4 s.d. in scores can therefore be expected to have no species in common. Rescaling of the CA axis of Table 5.3 results in the desired equal spacing of the site scores (Figure 5.5c); the length of the axis is 6 s.d.

DCA applied to the Dune Meadow Data gives, as always, the same first eigenvalue (0.53) as CA and a lower second eigenvalue (0.29 compared to 0.40 in CA). The lengths of the first two axes are estimated as 3.7 and 3.1 s.d., respectively. Because the first axis length is close to 4 s.d., we predict that sites at opposite ends of the first axis have hardly any species in common. This prediction can be verified in Table 5.1c (the order of DCA scores on the first axis is identical to that of CA); Site 17 and Site 16 have no species in common, but closer sites have one or more species in common. The DCA ordination diagram (Figure 5.7) shows the same overall pattern as the CA diagram of Figure 5.4. There are, however,

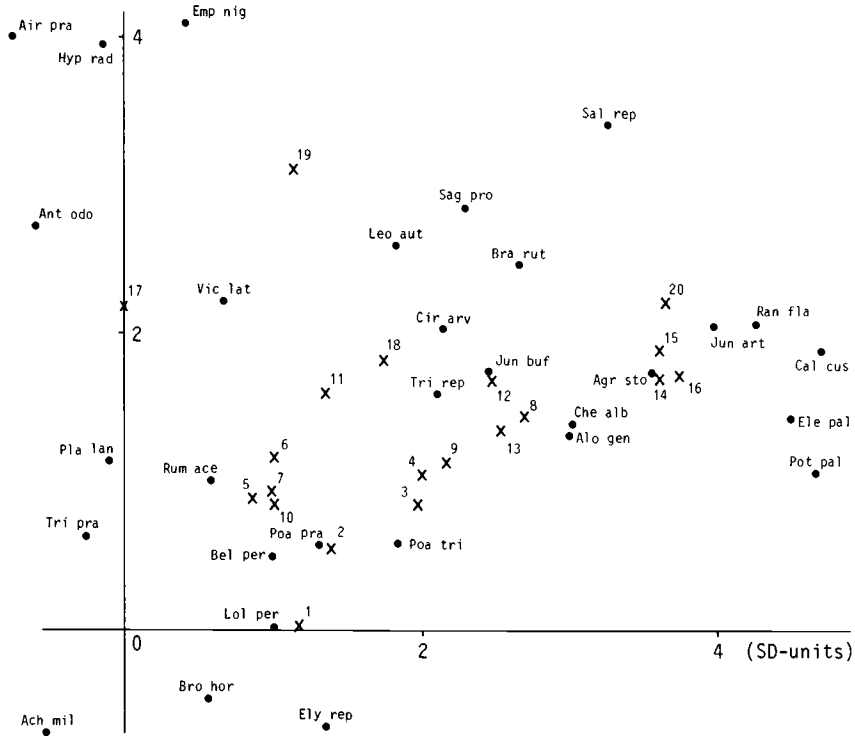


Figure 5.7 DCA ordination diagram of the Dune Meadow Data. The scale marks are in multiples of the standard deviation (s.d.).

differences in details. The arch seen in Figure 5.4 is less conspicuous, the position of Sites 17 and 19 is less aberrant. Further, *Achillea millefolium* is moved from a position close to Sites 2, 5, 6, 7 and 10 to the bottom left of Figure 5.7 and is then closest to Site 1; this move is unwanted, as this species is most abundant in the former group of sites (Table 5.1).

In an extentive simulation study, Minchin (1987) found that DCA, as available in the program DECORANA, can flatten out some of the variation associated with one of the underlying gradients. He ascribed this loss of information to an instability in either, or both, detrending and rescaling. Pielou (1984, p. 197) warned that DCA is 'overzealous' in correcting the 'defects' in CA and that it 'may sometimes lead to the unwitting destruction of ecologically meaningful information'.

DCA is popular among practical field ecologists, presumably because it provides an effective approximate solution to the ordination problem for a unimodal response model in two or more dimensions – given that the data are reasonably representative of sections of the major underlying environmental gradients. Two modifications might increase its robustness with respect to the problems identified by Minchin (1987). First, nonlinear rescaling aggravates these problems; since the edge effect is not too serious, we advise against the routine use of nonlinear rescaling. Second, the arch effect needs to be removed, but this can be done by a more stable, less 'zealous' method of detrending, which was also briefly mentioned by Hill & Gauch (1980): detrending-by-polynomials. The arch is caused by the folding of the first axis (Figure 5.3c), so that the second CA axis is about a quadratic function of the first axis, the third CA axis a cubic function of the first axis, and so on (Hill 1974). The arch is therefore most simply removed by requiring that the second axis is not only uncorrelated with the first axis (x_i), but also uncorrelated with its square (x_i^2) and, to prevent more folding, its cube (x_i^3). In contrast with 'detrending-by-segments', the method of detrending-by-polynomials removes only specific defects of CA that are now theoretically understood. Detrending by polynomials can be incorporated into the two-way weighted averaging algorithm (Table 5.2) by extending Step 4 such that the trial scores are not only made uncorrelated with the previous axes, but also with polynomials of previous axes. The computer program CANOCO (ter Braak 1987b) allows detrending by up to fourth-order polynomials.

5.2.5 Joint plot of species and sites

An ordination diagram mirrors the species data (although often with some distortion), so we can make inferences about the species data from the diagram. With Hill's scaling (Subsection 5.2.2), site scores are weighted averages of the species scores. Site points then lie in the ordination diagram at the centroid of the points of species that occur in them. Sites that lie close to the point of a species are therefore likely to have a high abundance of that species or, for presence–absence data, are likely to contain that species. Also, in so far as CA and DCA are a good approximation to fitting bell-shaped response surfaces to the species data (Subsection 5.2.1 and Section 5.7), the species points are close

to the optima of these surfaces; hence, the expected abundance or probability of occurrence of a species decreases with distance from its position in the plot (Figure 3.14).

Using these rules to interpret DCA diagrams, we predict as an example the rank order of species abundance for three species from Figure 5.7 and compare the order with the data in Table 5.1. The predicted rank order for *Juncus bufonius* is Sites 12, 8, 13, 9, 18 and 4; in the data *Juncus bufonius* is present at four sites, in order of abundance Sites 9, 12, 13 and 7. The predicted rank order for *Rumex acetosa* is Sites 5, 7, 6, 10, 2 and 11; in the data *R. acetosa* occurs in five sites, in order of abundance Sites 6, 5, 7, 9 and 12. *Ranunculus flammula* is predicted to be most abundant at Sites 20, 14, 15, 16 and less abundant, if present at all, at Sites 8, 12 and 13; in the data, *R. flammula* is present in six sites, in order of abundance Sites 20, 14, 15, 16, 8 and 13. We see some agreement between observations and predictions but also some disagreement. What is called for is a measure of goodness of fit of the ordination diagram. Such a measure is, however, not normally available in CA and DCA.

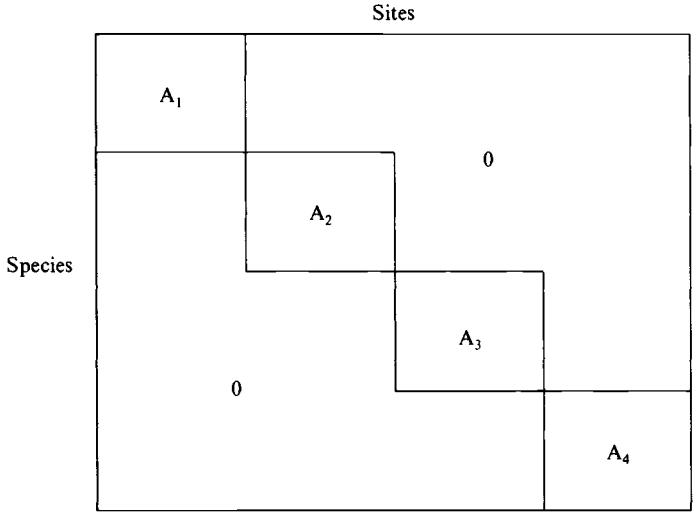
In interpreting ordination diagrams of CA and DCA, one should be aware of the following aspects. Species points on the edge of the diagram are often rare species, lying there either because they prefer extreme (environmental) conditions or because their few occurrences by chance happen to be at sites with extreme conditions. One can only decide between these two possibilities by additional external knowledge. Such species have little influence on the analysis; if one wants to enlarge the remainder of the diagram, it may be convenient not to display them at all. Further, because of the shortcomings of the method of weighted averaging, species at the centre of the diagram may either be unimodal with optima at the centre, or bimodal, or unrelated to the ordination axes. Which possibility is most likely can be decided upon by table rearrangement as in Table 5.1c or by plotting the abundance of a species against the axes. Species that lie between the centre and the outer edge are most likely to show a clear relation with the axes.

5.2.6 Block structures and sensitivity to rare species

CA has attractive properties in the search for block structures. A table is said to have block structure if its sites and species can be divided into clusters, with each cluster of species occurring in a single cluster of sites (Table 5.4). For any table that allows such a clustering, CA will discover it without fail. With the four blocks in Table 5.4, the first three eigenvalues of CA equal 1 and sites from the same cluster have equal scores on the three corresponding axes. An eigenvalue close to 1 can therefore point to an almost perfect block structure or to a diagonal structure in the data (Subsection 5.2.3). The search for block structures or 'near-block structures' by CA forms the basis of the cluster-analysis program TWINSPAN (Chapter 6).

This property of CA is, however, a disadvantage in ordination. If a table contains two disjoint blocks, one of which consists of a single species and a single site, then the first axis of CA finds this questionably uninteresting block. For a similar

Table 5.4 Data table with block structure. Outside the Sub-tables A_1, A_2, A_3 and A_4, there are no presences, so that there are four clusters of sites that have no species in common ($\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 1$).



reason, CA is sensitive to species that occur only in a few species-poor sites. In the 'down-weighting' option of the program DECORANA (Hill 1979a), species that occur in a few sites are given a low weight, so minimizing their influence, but this does not fully cure CA's sensitivity to rare species at species-poor sites.

5.2.7 Gaussian ordination and its relation with CA and DCA

In the introduction to CA (Subsection 5.2.1), we assumed that species show unimodal response curves to environmental variables, intuitively took the dispersion of the species scores as a plausible measure of how well an environmental variable explains the species data, and subsequently defined CA to be the technique that constructs a theoretical variable that explains the species data best in the sense of maximizing the dispersion. Because of the shortcomings of CA noted in the subsequent sections, the dispersion of the species scores is not ideal to measure the fit to the species data. We now take a similar approach but with a better measure of fit and assume particular unimodal response curves. We will introduce ordination techniques that are based on the maximum likelihood principle (Subsections 3.3.2 and 4.2.1), in particular Gaussian ordination, which is a theoretically sound but computationally demanding technique of ordination. We also show that the simpler techniques of CA and DCA give about the same result if particular additional conditions hold true. This subsection may now be skipped at first reading; it requires a working knowledge of Chapters 3 and 4.

One dimension

In maximum likelihood ordination, a particular response model (Subsection 3.1.2) is fitted to the species data by using the maximum likelihood principle. In this approach, the fit is measured by the deviance (Subsection 3.3.2) between the data and the fitted curves. Recall that the deviance is inversely related to the likelihood, namely deviance = $-2 \log_e$ (likelihood). If we fit Gaussian (logit) curves (Figure 3.9) to the data, we obtain Gaussian ordination. In Subsection 3.3.3, we fitted a Gaussian logit curve of pH to the presence–absence data of a particular species (Figure 3.10). In principle, we can fit a separate curve for each species under consideration. A measure of how badly pH explains the species data is then the deviance (Table 3.6) summed over all species. Gaussian ordination of presence–absence data is then the technique that constructs the theoretical variable that best explains the species data by Gaussian logit curves, i.e. that minimizes the deviance between the data and the fitted curves.

A similar approach can be used for abundance data by fitting Gaussian curves to the data, as in Section 3.4, with the assumption that the abundance data follow a Poisson distribution. A Gaussian curve for a particular species has three parameters: optimum, tolerance and maximum (Figure 3.6), for species k denoted by u_k, t_k and c_k, respectively. In line with Equation 3.8, the Gaussian curves can now be written as

$$Ey_{ki} = c_k \exp\left[-0.5(x_i - u_k)^2 / t_k^2\right] \qquad \text{Equation 5.4}$$

where x_i is the score of site i on the ordination axis (the value of the theoretical variable at site i).

To fit this response model to data we can use an algorithm akin to that to obtain the ordination axis in CA (Table 5.2).

Step 1: Start from initial site scores x_i.

Step 2: Calculate new species scores by (log-linear) regression of the species data on the site scores (Section 3.4). For each species, we so obtain new values for u_k, t_k and c_k.

Step 3: Calculate new site scores by maximum likelihood calibration (Subsection 4.2.1).

Step 4: Standardize the site scores and check whether they have changed and, if so, go back to Step 2, otherwise stop.

In this algorithm, the ordination problem is solved by solving the regression problem (Chapter 3) and the calibration problem (Chapter 4) in an iterative fashion so as to maximize the likelihood. In contrast to the algorithm for CA, this algorithm may give different results for different initial site scores because of local maxima in the likelihood function for Equation 5.4. It is therefore not guaranteed that the algorithm actually leads to the (overall) maximum likelihood estimates; hence, we must supply 'good' initial scores, which are also needed to reduce the computational burden. Even for modern computers, the algorithm requires heavy computation. In the following, we show that a good choice for initial scores are

the scores obtained by CA.

The CA algorithm can be thought of a simplification of the maximum likelihood algorithm. In CA, the regression and calibration problems are both solved by weighted averaging. Recall that in CA the species score (u_k) is a position on the ordination axis x indicating the value most preferred by that particular species (its optimum) and that the site score (x_i) is the position of that particular site on the axis.

We saw in Section 3.7 that the optimum or score of a species (u_k) can be estimated efficiently by weighted averaging of site scores provided that (Figure 3.18b):

A1. the site scores are homogeneously distributed over the whole range of occurrence of the species along the axis x.

In Section 4.3, we saw that the score (x_i) of a site is estimated efficiently by weighted averaging of species optima provided the species packing model holds, i.e. provided (Figure 4.1):

A2. the species' optima (scores) are homogeneously distributed over a large interval around x_i.

A3. the tolerances of species t_k are equal (or at least independent of the optima, ter Braak 1985).

A4. the maxima of species c_k are equal (or at least independent of the optima; ter Braak 1985).

Under these four conditions the scores obtained by CA approximate the maximum likelihood estimates of the optima of species and the site values in Gaussian ordination (ter Braak 1985). For presence–absence data, CA approximates similarly the maximum likelihood estimates of the Gaussian logit model (Subsection 3.3.3). CA does not, however, provide estimates for the maximum and tolerance of a species.

A problem is that assumptions A1 and A2 cannot be satisfied simultaneously for all sites and species: the first assumption requires that the range of the species optima is amply contained in the range of the site scores whereas the second assumption requires the reverse. So CA scores show the edge effect of compression of the end of the first axis relative to the axis middle (Subsection 5.2.3). In practice, the ranges may coincide or may only partly overlap. CA does not give any clue about which possibility is likely to be true. The algorithm in Table 5.2 results in species scores that are weighted averages of the site scores and, consequently, the range of the species scores is contained in the range of the site scores. But it is equally valid mathematically to stop at Step 3 of the algorithm, so that the site scores are weighted averages of the species scores and thus all lie within the range of the species scores; this is done in the computer program DECORANA (Hill 1979). The choice between these alternatives is arbitrary. It may help interpretation of CA results to go one step further in the direction of the maximum likelihood estimates by one regression step in which the data of each species are regressed on the site scores of CA by using the Gaussian response model. This can be done by methods discussed in Chapter 3. The result is new species scores (optima) as well as estimates for the tolerances and maxima. As an example, Figure 5.8 shows Gaussian response curves along the first CA axis fitted to the

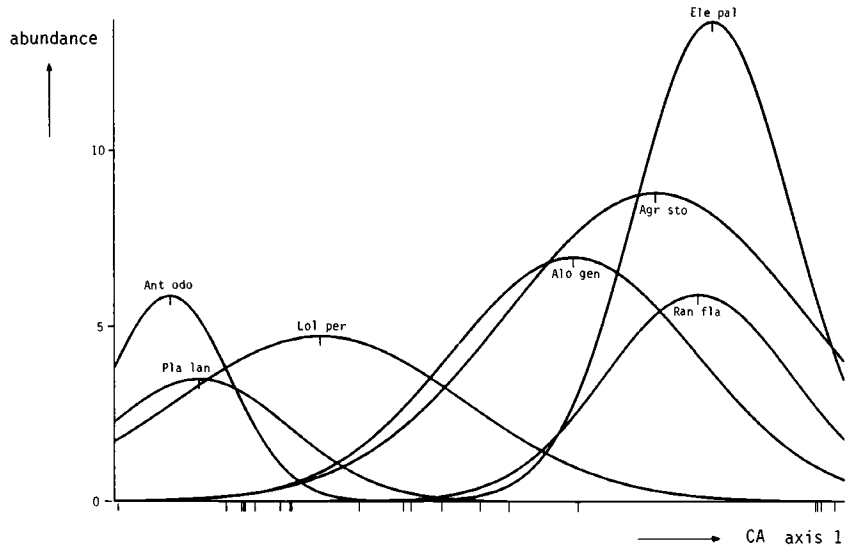


Figure 5.8 Gaussian response curves for some Dune Meadow species, fitted by log-linear regression of the abundances of species (Table 5.1) on the first CA axis. The sites are shown as small vertical lines below the horizontal axis.

Dune Meadow Data in Table 5.1. The curve of a particular species was obtained by a log-linear regression (Section 3.4) of the data of the species on the site scores of the first CA axis by using $b_0 + b_1 x + b_2 x^2$ in the linear predictor (Equation 3.18).

Two dimensions

In two dimensions, Gaussian ordination means fitting the bivariate Gaussian surfaces (Figure 3.14)

$$Ey_{ki} = c_k \exp(-0.5[(x_{i1} - u_{k1})^2 + (x_{i2} - u_{k2})^2]/t_k^2) \qquad \text{Equation 5.5}$$

where
(u_{k1}, u_{k2}) are the coordinates of the optimum of species k in the ordination diagram
c_k is the maximum of the surface
t_k is the tolerance
(x_{i1}, x_{i2}) are the coordinates of site i in the diagram.

These Gaussian surfaces look like that of Figure 3.14, but have circular contours because the tolerances are taken to be the same in both dimensions.

One cannot hope for more than that the two-axis solution of CA provides an approximation to the fitting of Equation 5.5 if the sampling distribution of

the abundance data is Poisson and if:
A1. site points are homogeneously distributed over a rectangular region in the ordination diagram with sides that are long compared to the tolerances of the species,
A2. optima of species are homogeneously distributed over the same region,
A3. the tolerances of species are equal (or at least independent of the optima),
A4. the maxima of species are equal (or at least independent of the optima).

However as soon as the sides of the rectangular region differ in length, the arch effect (Subsection 5.2.3) crops up and the approximation is bad. Figure 5.9b shows the site ordination diagram obtained by applying CA to artificial species data (40 species and 50 sites) simulated from Equation 5.5 with $c_k = 5$ and $t_k = 1$ for each k. The true site points were completely randomly distributed over

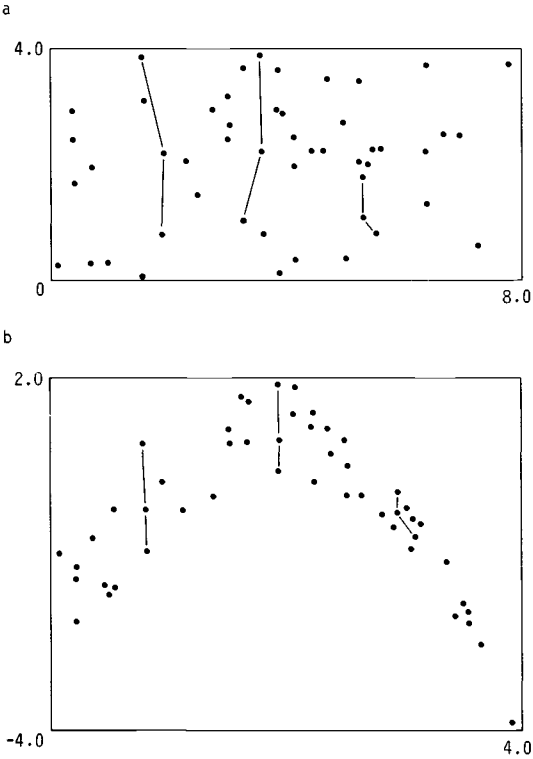


Figure 5.9 CA applied to simulated species data. a: True configuration of sites (•). b: Configuration of sites obtained by CA, showing the arch effect. The data were obtained from the Gaussian model of Equation 5.5 with Poisson error, $c_k = 5$, $t_k = 1$ and optima that were randomly distributed in the rectangle $[-1,9] \times [-0.5,4.5]$. The vertical lines in Figures a and b connect identical sites.

a rectangular region with sides of 8 and 4 s.d. (Figure 5.9a). The CA ordination diagram is dominated by the arch effect, although the actual position of sites within the arch still reflects their position on the second axis in Figure 5.9a. The configuration of site scores obtained by DCA was much closer to the true configuration. DCA forcibly imposes Conditions A1, A2 and A3 upon the solution, the first one by detrending and the second and third one by rescaling of the axes.

We also may improve the ordination diagram of DCA by going one step further in the direction of maximum likelihood ordination by one extra regression step. We did so for the DCA ordination (Figure 5.7) of Dune Meadow Data in Table 0.1. For each species with more than 4 presences, we carried out a log-linear regression of the data of the species on the first two DCA axes using the response model

$$\log_e \mathrm{E}y_{ki} = b_0 + b_{1k}\, x_{i1} + b_{3k}\, x_{i2} + b_{4k}\, (x_{i1}^2 + x_{i2}^2) \qquad \text{Equation 5.6}$$

where x_{i1} and x_{i2} are the scores of site i on the DCA axes 1 and 2, respectively.

If $b_{4k} < 0$, this model is equivalent to Equation 5.5 (as in Subsection 3.3.3). The new species scores are then obtained from the estimated parameters in Equation 5.6 by $u_{k1} = -b_{1k}/(2b_{4k})$, $u_{k2} = -b_2/(2b_{4k})$ and $t_k = 1/\sqrt{(-2b_{4k})}$.

If $b_{4k} > 0$, the fitted surface shows a minimum and we have just plotted the DCA scores of the species. Figure 5.10a shows how the species points obtained by DCA change by applying this regression method to the 20 species with four or more presences. A notable feature is that *Achillea millefolium* moves towards its position in the CA diagram (Figure 5.4). In Figure 5.10b, circles are drawn with centres at the estimated species points and with radius t_k. The circles are contours where the expected abundance is 60% of the maximum expected abundance c_k. Note that exp (-0.5) = 0.60.

From Figure 5.10b, we see, for example, that *Trifolium repens* has a high tolerance (a large circle, thus a wide ecological amplitude) whereas *Bromus hordaceus* has a low tolerance (a small circle, thus a narrow ecological amplitude). With regression, the joint plot of DCA can be interpreted with more confidence. This approach also leads to a measure of goodness of fit. A convenient measure of goodness of fit is here

$$V = 1 - (\Sigma_k\, D_{k1})/\,(\Sigma_k\, D_{k0}) \qquad \text{Equation 5.7}$$

where D_{k0} and D_{k1} are the residual deviances of the kth species for the null model (the model without explanatory variables) and the model depicted in the diagram (Equation 5.6), respectively. These deviances are obtained from the regressions (as in Table 3.7). We propose to term V the fraction of deviance accounted for by the diagram. For the two-axis ordination (only partially displayed in Figure 5.10b) $V = (1 - 360/987) = 0.64$. For comparison, $V = 0.51$ for the one-axis ordination (partially displayed in Figure 5.8).

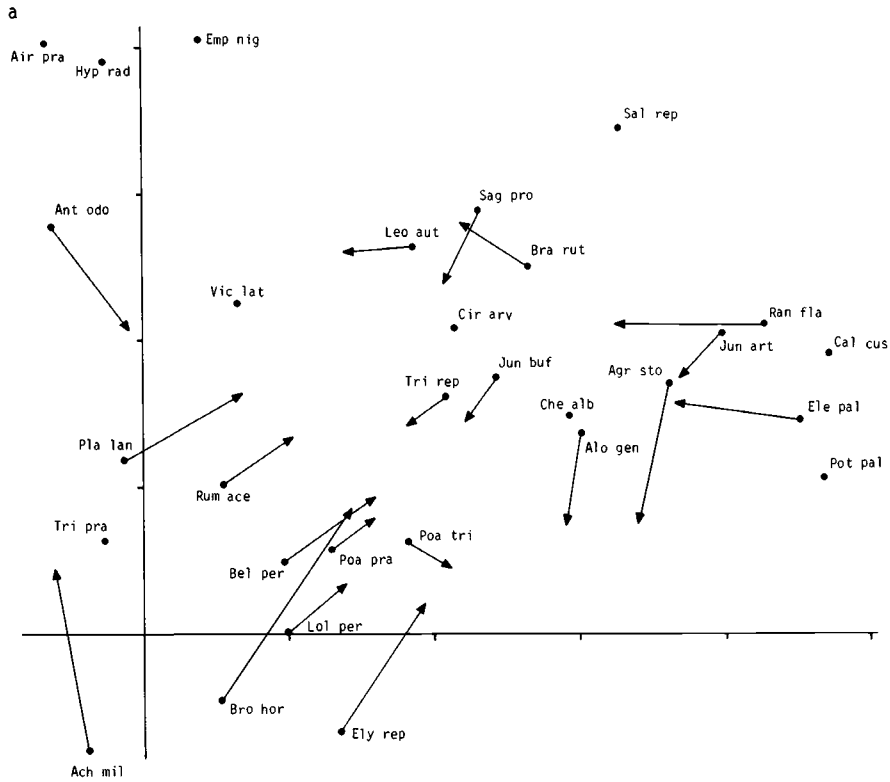


Figure 5.10 Gaussian response surfaces for several Dune Meadow species fitted by log-linear regression of the abundances of species on the site scores of the first two DCA axes (Figure 5.7). a: Arrows for species running from their DCA scores (Figure 5.7) to their fitted optimum. b: Optima and contours for some of the species. The contour indicates where the abundance of a species is 60% of the abundance at its optimum.

The regression approach can of course be extended to more complicated surfaces (e.g. Equation 3.24), but this will often be impractical, because these surfaces are more difficult to represent graphically.

5.3 Principal components analysis (PCA)

5.3.1 From least-squares regression to principal components analysis

Principal components analysis (PCA) can be considered to be an extension of fitting straight lines and planes by least-squares regression. We will introduce

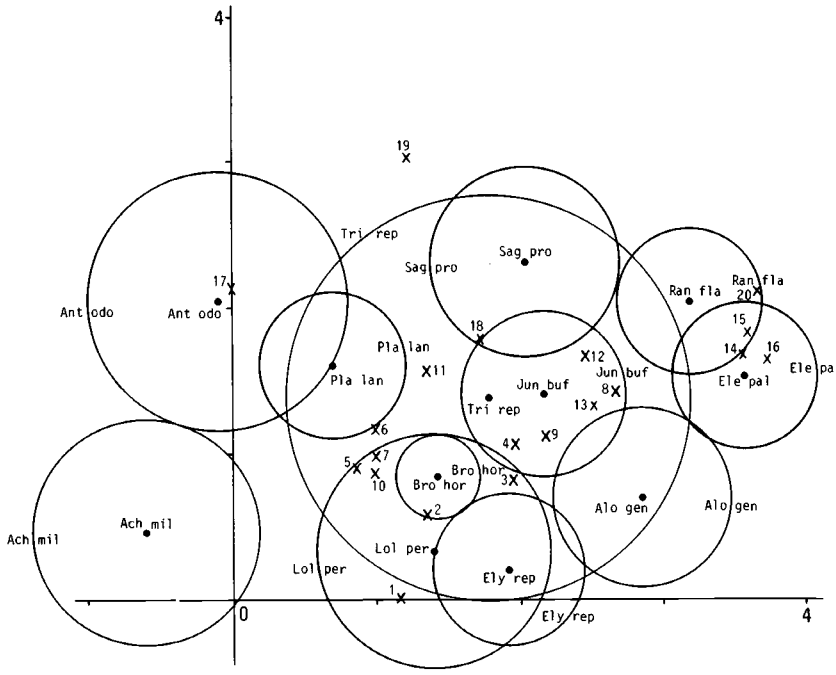


Figure 5.10b

PCA, assuming the species data to be quantitative abundance values.

Suppose we want to explain the abundance values of several species by a particular environmental variable, say moisture, and suppose we attempt to do so by fitting straight lines to the data. Then, for each species, we have to carry out a least-squares regression of its abundance values on the moisture values and obtain, among other things, the residual sum of squares, i.e. the sum of squared vertical distances between the observed abundance values and the fitted line (Figure 3.1; Subsection 3.2.2). This is a measure of how badly moisture explains the data of a single species. To measure how badly moisture explains the data of all species, we now use the total of the separate residual sums of squares over all species, abbreviated the total residual sum of squares. If the total residual sum of squares is small, moisture can explain the species data well.

Now, suppose that, among a set of environmental variables, moisture is the variable that best explains the species data in the sense of giving the least total residual sum of squares. As in all ordination techniques, we now wish to construct a theoretical variable that explains the species data still better. PCA is the ordination technique that constructs the theoretical variable that minimizes the total residual sum of squares after fitting straight lines to the species data. PCA does so by

choosing best values for the sites, the site scores. This is illustrated in Figure 5.11 for the Dune Meadow Data. The site scores are indicated by ticks below the horizontal axis. The fitted lines are shown for six of the 30 species and the observed abundance values and residuals for one of them. Any other choice of site scores would result in a larger sum of squared residuals. Note that Figure 5.11 shows only 20 out of all $20 \times 30 = 600$ residuals involved. The horizontal axis in Figure 5.11 is the first PCA axis, or first principal component. The score of a species in PCA is actually the slope of the line fitted for the species against the PCA axis. A positive species score thus means that the abundance increases along the axis (e.g. *Agrostis stolonifera* in Figure 5.11); a negative score means that the abundance decreases along the axis (e.g *Lolium perenne* in Figure 5.11) and a score near 0 that the abundance is poorly (linearly) related to the axis (e.g. *Sagina procumbens* in Figure 5.11).

If a single variable cannot explain the species data sufficiently well, we may attempt to explain the data with two variables by fitting planes (Subsection 3.5.2). Then, for each species we have to carry out a least-squares regression of its abundance values on two explanatory variables (Figure 3.11), obtain its residual sum of squares and, by addition over species, the total residual sum of squares. The first two axes of PCA are now the theoretical variables minimizing the total residual sum of squares among all possible choices of two explanatory variables. Analogously, the first three PCA axes minimize the total residual sum of squares by fitting the data to hyperplanes, and so on. PCA is thus a multi-species extension of multiple (least-squares) regression. The difference is that in multiple regression

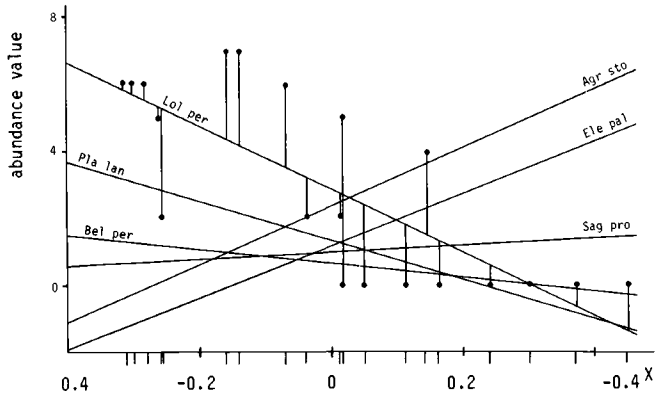


Figure 5.11 Straight lines for several Dune Meadow species, fitted by PCA to the species abundances of Table 5.1. Also shown are the abundances of *Lolium perenne* and their deviations from the fitted straight line. The horizontal axis is the first principal component. Fitting straight lines by least-squares regression of the abundances of species on the site scores of the first PCA axis gives the same results. The slope equals the species score of the first axis. The site scores are shown by small vertical lines below the horizontal axis.

the explanatory variables are supplied environmental variables whereas in PCA the explanatory variable are theoretical variables estimated from the species data alone. It can be shown (e.g. Rao 1973) that the same result as above is obtained by defining the PCA axes sequentially as follows. The first PCA axis is the variable that explains the species data best, and second and later axes also explain the species data best but subject to the constraint of being uncorrelated with previous PCA axes. In practice, we ignore higher numbered PCA axes that explain only a small proportion of variance in the species data.

5.3.2 Two-way weighted summation algorithm

We now describe an algorithm that has much in common with that of CA and that gives the ordination axes of PCA. The algorithm also shows PCA to be a natural extension of straight-line regression.

If the relation between the abundance of a species and an environmental variable is rectilinear, we can summarize the relation by the intercept and slope of a straight line. The error part of the model is taken to consist of independent and normally distributed errors with a constant variance. The parameters (intercept and slope) are then estimated by least-squares regression of the species abundances on the values of the environmental variable (Subsection 3.2.2). Conversely, when the intercepts and slopes are known, we can estimate the value of the environmental variable from the species abundances at a site by calibration (Subsection 4.2.3). If it is not known in advance which environmental variable determines the abundances of the species, the idea is as in CA (Subsection 5.2.2) to discover the 'underlying environmental gradient' by applying straight-line regression and calibration alternately in an iterative fashion, starting from arbitrary initial values for sites or from arbitrary initial values for the intercepts and slopes of species. As in CA, the iteration process eventually converges to a set of values for species and sites that does not depend on the initial values.

The iteration process reduces to simple calculations when we first centre the abundances of each species to mean 0 and standardize the site scores to $\bar{x} = 0$ and $\Sigma_i\ (x_i - \bar{x})^2 = 1$. Then, the equations to estimate the intercept and the slope of a straight line (Equations 3.6a,b) reduce to $b_0 = 0$ and $b_1 = \Sigma_i\ y_i\ x_i$, because in the notation of Subsection 3.2.2 $\bar{y} = 0$, $\bar{x} = 0$ and $\Sigma_i\ (x_i - \bar{x})^2 = 1$. Hence we ignore the intercepts and concentrate on the slope parameters. From now on, b_k will denote the slope parameter for species k and y_{ki} the centred abundance of species k at site i (i.e. $y_{k+} = 0$). In this notation, the slope parameter of species k is calculated by

$$b_k = \Sigma_{i=1}^{n}\ y_{ki}\ x_i \qquad \text{Equation 5.8}$$

As an example, Table 5.5a shows the Dune Meadow Data used before with an extra column of species means and, as arbitrary initial scores for the sites, values obtained by standardizing the numbers 1 to 20 (bottom row). For *Achillea millefolium*, the mean abundance is 0.80 and we obtain

Table 5.5a

Species k		Sites (i) 00000000011111111112 12345678901234567890	mean	b_k
1	Ach mil	13 222 4 2	0.80	-1.98
2	Agr sto	48 43 45447 5	2.40	1.55
3	Air pra	2 3	0.25	1.49
4	Alo gen	272 53 85 4	1.80	-2.06
5	Ant odo	432 4 4 4	1.05	0.60
6	Bel pre	3222 2 2	0.65	-1.96
7	Bro hor	4 32 2 4	0.75	-2.85
8	Che alb	1	0.05	0.10
9	Cir arv	2	0.10	-0.50
10	Ele pal	4 458 4	1.25	4.21
11	Ely rep	44444 6	1.30	-6.17
12	Emp nig	2	0.10	0.66
13	Hyp rad	2 2 5	0.45	2.19
14	Jun art	44 33 4	0.90	2.02
15	Jun buf	2 4 43	0.65	0.02
16	Leo aut	52233332352222 2562	2.70	0.93
17	Lol per	75652664267 2	2.90	-9.42
18	Pla lan	555 33 23	1.30	-1.24
19	Poa pra	44542344444 2 13	2.40	-6.05
20	Poa tri	2765645454 49 2	3.15	-7.93
21	Pot pal	22	0.20	0.62
22	Ran fla	2 2222 4	0.70	2.52
23	Rum ace	563 2 2	0.90	-2.52
24	Sag pro	5 22 242 3	1.00	-0.12
25	Sal rep	335	0.55	3.70
26	Tri pra	252	0.45	-1.57
27	Tri rep	52125223633261 22	2.35	-1.88
28	Vic lat	12 1	0.20	0.31
29	Bra rut	2226222244 44 634	2.45	2.89
30	Cal cus	4 3 3	0.50	2.29
x_i		---------- 00000000000000000000 .... 33222111000011122233 73951740622604715937		

Table 5.5 Two-way weighted summation algorithm of PCA applied to the Dune Meadow Data. a: The original data table with at the bottom the initial site scores. b: The species and sites rearranged in order of their scores obtained after one cycle of two-way weighted summation. c: The species arranged in order of their final scores (PCA scores).

Table 5.5b

Species	Sites (i)	b_k
	00010000011011111112	
k	23704156931828749650	
17 Lol per	566657262 74 2	-9.42
20 Poa tri	7654526459 44 2	-7.93
11 Ely rep	44 444 6	-6.17
19 Poa pra	454444234244 31	-6.05
7 Bro hor	4 243 2	-2.85
23 Rum ace	3 562 2	-2.52
4 Alo gen	27 2 35 58 4	-2.06
1 Ach mil	3 24 122 2	-1.98
6 Bel per	32 22 2 2	-1.96
27 Tri rep	52261 25323232 62 1	-1.88
26 Tri pra	2 25	-1.57
18 Pla lan	53 55 3 32	-1.24
9 Cir arv	2	-0.50
24 Sag pro	5 22224 3	-0.12
15 Jun buf	2 43 4	0.02
8 Che alb	1	0.10
28 Vic lat	1 2 1	0.31
5 Ant odo	24 43 4 4	0.60
21 Pot pal	2 2	0.62
12 Emp nig	2	0.66
16 Leo aut	52332 33225325226 22	0.93
3 Air pra	2 3	1.49
2 Agr sto	4 8 35 44 4 745	1.55
14 Jun art	4 4 334	2.02
13 Hyp rad	2 2 5	2.19
30 Cal cus	4 3 3	2.29
22 Ran fla	2 2 2 224	2.52
29 Bra rut	2222 262 4246 3444	2.89
25 Sal rep	3 3 5	3.70
10 Ele pal	4 4 854	4.21
	---------	
x_i	00000000000000000000	
	.....	
	32221111100001122334	
	38109987454174869220	

Table 5.5c

Species	Sites (i)	b_k
	01000100100111011121	
k	60725113894793824506	
17 Lol per	66652776225 4	-9.21
18 Pla lan	535 53 3 2	-5.77
19 Poa pra	344424453441 24	-5.69
20 Poa tri	44576 26 55 944 2	-4.80
1 Ach mil	24232 1 2	-3.81
23 Rum ace	6 3 5 2 2	-3.68
27 Tri rep	562523 2231 222361	-3.67
5 Ant odo	342 4 44	-3.52
7 Bro hor	4242 3	-3.31
16 Leo aut	333535 252226232222	-2.86
11 Ely rep	44 44 64	-2.86
26 Tri pra	5 2 2	-2.63
6 Bel per	2 32 22 2	-2.11
28 Vic lat	1 2 1	-0.67
13 Hyp rad	2 25	-0.08
9 Cir arv	2	0.01
12 Emp nig	2	0.09
8 Che alb	1	0.11
3 Air pra	23	0.15
15 Jun buf	2 4 3 4	0.40
29 Bra rut	622 24 2622 3 24 444	0.94
24 Sag pro	2 25 3224	0.98
21 Pot pal	22	1.07
25 Sal rep	3 3 5	1.86
4 Alo gen	2 7 32 558 4	3.33
30 Cal cus	4 33	3.40
22 Ran fla	22 2242	3.95
14 Jun art	4 4 343	4.29
10 Ele pal	4 4548	8.08
2 Agr sto	4 38 5444457	8.67
	---------	
x_i	00000000000000000000	
	.....	
	33222110000001112334	
	10866647401151464075	

$b_1 = (1 - 0.80) \times (-0.37) + (3 - 0.80) \times (-0.33) + (0 - 0.80) \times (-0.29) + \ldots + (0 - 0.80) \times (0.37) = -1.98$.

From the slopes thus obtained (Table 5.5a, last column), we derive new site scores by least-squares calibration (Equation 4.2 with $a_k = 0$). The site scores so obtained are proportional to

$$x_i = \Sigma_{k=1}^{m} y_{ki} b_k \qquad \text{Equation 5.9}$$

because the denominator of Equation 4.1 has the same value for each site. This denominator is unimportant in PCA, because the next step in the algorithm is to standardize the site scores, as shown in Table 5.6c. For Site 1 in Table 5.5a, we get from Equation 5.9 the site score $x_1 = (1 - 0.80) \times (-1.98) + (0 - 2.40) \times (1.55) + (0 - 0.25) \times (1.49) + \ldots + (0 - 0.50) \times (2.29) = -0.19$. Note that the species mean abundance is subtracted each time from the abundance value. In Table 5.5b, the species and sites are arranged in order of the scores obtained so far, in which the slopes (b_k) form the species scores. The abundance of the species in the top row (*Lolium perenne*) has the tendency to decrease along the row, whereas the abundance of the species in the bottom row (*Eleocharis palustris*) has the tendency to increase across the row. The next cycle of the iteration is to calculate new species scores (b_k), then new site scores, and so on. As in CA, the scores stabilize after several iterations and the resulting scores (Table 5.5c) constitute the first ordination axis of PCA. In Table 5.5c, the species and sites are arranged in order of their scores on the first axis. Going from top row to bottom row, we see first a decreasing trend in abundance across the columns (e.g. for *Lolium perenne*), then hardly any trend (e.g. for *Sagina procumbens*) and finally an increasing trend (e.g. for *Agrostis stolonifera*). A graphical display of the trends has already been shown in Figure 5.11. The order of species in Table 5.5c is quite different from the order in the table arranged by CA (Table 5.1c), but the difference in ordering of the sites is more subtle.

In the above iteration algorithm of PCA (Table 5.6), weighted sums (Equations 5.8 and 5.9) replace the weighted averages in CA (Table 5.2; Equations 5.1 and 5.2). For this analogy to hold, let us consider the data y_{ki} as weights (which can be negative in PCA), so that the species scores are a weighted sum of the site scores and, conversely, the site scores are a weighted sum of the species scores (Table 5.6). The standard terminology used in mathematics is that x_i is a linear combination of the variables (species) and that b_k is the loading of species k.

After the first axis, a second axis can be extracted as in CA, and so on. (There is a subtle difference in the orthogonalization procedure, which need not concern us here.) The axes are also eigenvectors to which correspond eigenvalues as in CA (Subsection 5.2.2). The meaning of the eigenvalues in PCA is given below. The axes are also termed principal components.

So PCA decomposes the observed values into fitted values and residuals (Equations 3.1 and 3.2). In one dimension, we have the decomposition

$$y_{ki} = b_k x_i + \text{residual} \qquad \text{Equation 5.10}$$

Table 5.6 Two-way weighted summation algorithm of PCA.

a: Iteration process

Step 1. Take arbitrary initial site scores (x_i), not all equal to zero.
Step 2. Calculate new species scores (b_k) by weighted summation of the site scores (Equation 5.8).
Step 3. Calculate new site scores (x_i) by weighted summation of the species scores (Equation 5.9).
Step 4. For the first axis go to Step 5. For second and higher axes, make the site scores (x_i) uncorrelated with the previous axes by the orthogonalization procedure described below.
Step 5. Standardize the site scores (x_i). See below for the standardization procedure.
Step 6. Stop on convergence, i.e. when the new site scores are sufficiently close to the site scores of the previous cycle of the iteration; ELSE go to Step 2.

b: Orthogonalization procedure

Step 4.1. Denote the site scores of the previous axis by f_i and the trial scores of the present axis by x_i.
Step 4.2. Calculate $v = \Sigma_{i=1}^{n}\ x_i f_i$.
Step 4.3 Calculate $x_{i,\text{new}} = x_{i,\text{old}} - v f_i$.
Step 4.4 Repeat Steps 4.1-4.3 for all previous axes.

c: Standardization procedure

Step 5.1 Calculate the sum of squares of the site scores $s^2 = \Sigma_{i=1}^{n}\ x_i^2$.
Step 5.2 Calculate $x_{i,\text{new}} = x_{i,\text{old}}/s$.
Note that, upon convergence, s equals the eigenvalue.

where y_{ki} is the (mean corrected) observed value and $b_k\ x_i$ the fitted value.

As an example, the values fitted by the first PCA axis (Table 5.5c) for the centred abundances of *Agrostis stolonifera* (b_2 = 8.67) at Site 6 (x_6 = -0.31) and Site 16 (x_{16} = 0.45) are: 8.67 × (-0.31) = -2.75 and 8.67 × 0.45 = 3.99, respectively. Adding the mean value of *A. stolonifera* (2.40), we obtain the values -0.35 and 6.39, respectively, which are close to the observed abundance values of 0 and 7 at Site 6 and Site 16. In PCA, the sum of squared residuals in Equation 5.10 is minimized (Subsection 5.3.1). Analogously, one can say that PCA maximizes the sum of squares of fitted values and the maximum is the eigenvalue of the first axis. In two dimensions (Figure 5.12), we have the decomposition

$$y_{ki} = (b_{k1}\ x_{i1} + b_{k2}\ x_{i2}) + \text{residual} \qquad \text{Equation 5.11}$$

where
b_{k1} and b_{k2} are the scores of species k
x_{i1} and x_{i2} are the scores of site i on Axis 1 and Axis 2, respectively.

On the second axis, the score of *A. stolonifera* is 6.10 and the scores of Sites 6 and 16 are -0.17 and 0.033 (Figure 5.12), so that the fitted values become 8.67 × (-0.31) + 6.10 × (-0.17) = -3.72 and 8.67 × 0.45 + 6.10 × 0.033 = 4.10.

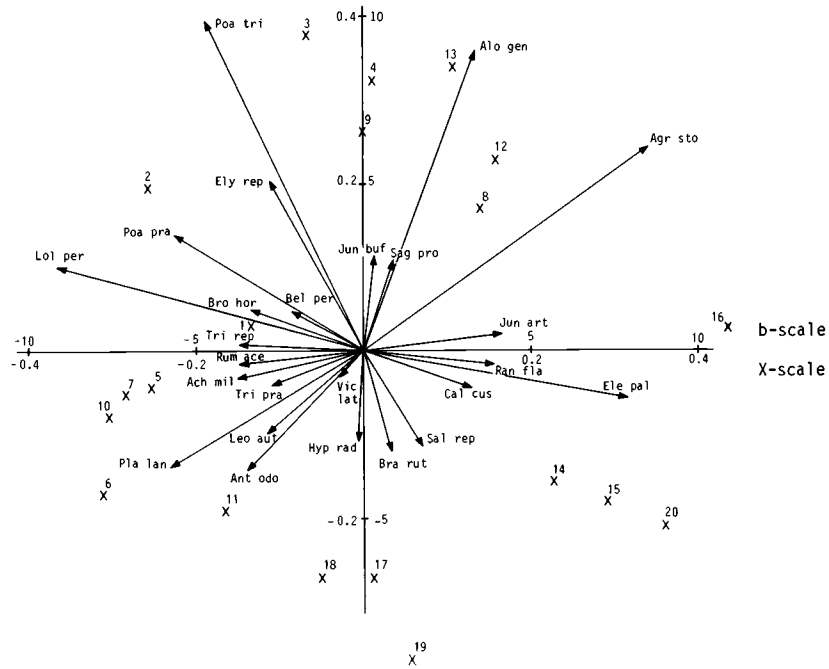


Figure 5.12 PCA-ordination diagram of the Dune Meadow Data in covariance biplot scaling with species represented by arrows. The *b* scale applies to species, the *x* scale to sites. Species not represented in the diagram lie close to the origin (0,0).

The first two PCA axes thus give approximate abundance values of $-3.72 + 2.40 = -1.3$ and $4.10 + 2.40 = 6.5$, slightly worse values than those obtained from the first axis, but most of the remaining abundance values in the data table will be approximated better with two axes than with a single axis. The sum of squares of fitted values now equals $\lambda_1 + \lambda_2$. Further, the total sum of squares ($\Sigma_k \Sigma_i y_{ki}^2$) equals the sum of all eigenvalues. (This equality means that we can reconstruct the observed values exactly from the scores of species and sites on all eigenvectors and the mean abundance values.) The fraction of variance accounted for (explained) by the first two axes is therefore $(\lambda_1 + \lambda_2)/(\text{sum of all eigenvalues})$. This measure is the equivalent of R^2 in Section 3.2. For the Dune Meadow Data, $\lambda_1 = 471$, $\lambda_2 = 344$ and the total sum of squares = 1598. So the two-axes solution explains $(471 + 344)/1598 = 51\%$ of the variance. The first axis actually explains $471/1598 = 29\%$ of the variance and the second axis $344/1598 = 22\%$.

5.3.3 *Best lines and planes in m-dimensional space*

Here we present a geometric approach to PCA. In this approach, the aim of PCA is seen as being to summarize multivariate data in a graphical way. The approach is best illustrated with data on two species only. Figure 5.13a displays the abundances of Species A and B at 25 sites in the form of a scatter diagram, with axes labelled by the species names. The simplest summary of data is by the mean abundances of A (25) and B (15). Knowing the means, we may shift the axes to the centroid of the data points, i.e. to the point with the coordinates (25,15), provided we remember that the origin (0,0) of the new coordinate system is the point (25,15) in the old coordinate system. Next we draw a line through the new origin in the direction of maximum variance in the plot. This line is the first principal component (PC1), or first PCA axis, and perpendicularly we draw PC2. Next we rotate the plot, so that PC1 is horizontal (Figure 5.13b). Figure 5.13b is an ordination diagram with arrows representing the species. These arrows are the shifted and rotated axes of the species in the original diagram. PC2 shows so much less variation than PC1 that PC2 can possibly be neglected. This is done in Figure 5.13c showing a one-dimensional ordination; the points in Figure 5.13c were obtained from Figure 5.13b by drawing perpendicular lines from each point on the horizontal axis (projection onto PC1). In this way, the first coordinate of the points in Figure 5.13b is retained in Figure 5.13c; this coordinate is the site score on PC1. The first coordinate of the arrows in Figure 5.13b is the species loading on PC1, which is also represented by an arrow in Figure 5.13c. These arrows indicate the direction in which Species A and Species B increase in abundance; hence Figure 5.13c still shows which sites have high abundances of Species A and of Species B (those on the right side) and which sites have low abundance (those on the left side).

The example is, of course, artificial. Usually there are many species ($m \geqslant 3$), so that we need an m-dimensional coordinate system, and we want to derive a two-dimensional or three-dimensional ordination diagram. Yet the principle remains the same: PCA searches for the direction of maximum variance; this is PC1, the best line through the data points. It is the best line in the sense that it minimizes the sum of squares of perpendicular distances between the data points and the line (as is illustrated in Figure 5.13a for $m = 2$). So the first component in Figure 5.13a is neither the regression line of Species B on Species A nor that of Species A on Species B, because regression minimizes the sum of squares of vertical distances (Figure 3.1). But, as we have seen in Subsection 5.3.1, PCA does give the best regression of Species A on PC1 and of Species B on PC1 (Figure 5.11). After the first component, PCA seeks the direction of maximum variance that is perpendicular onto the first axis; that is PC2, which with PC1 forms the best plane through the data points, and so on. In general, the site scores are obtained by projecting each data point from the m-dimensional space onto the PCA axes and the species scores are obtained by projecting the unit vectors: for the first species (1,0,0,...); for the second species (0,1,0,0,...), etc., onto the PCA axes (Figure 5.13).

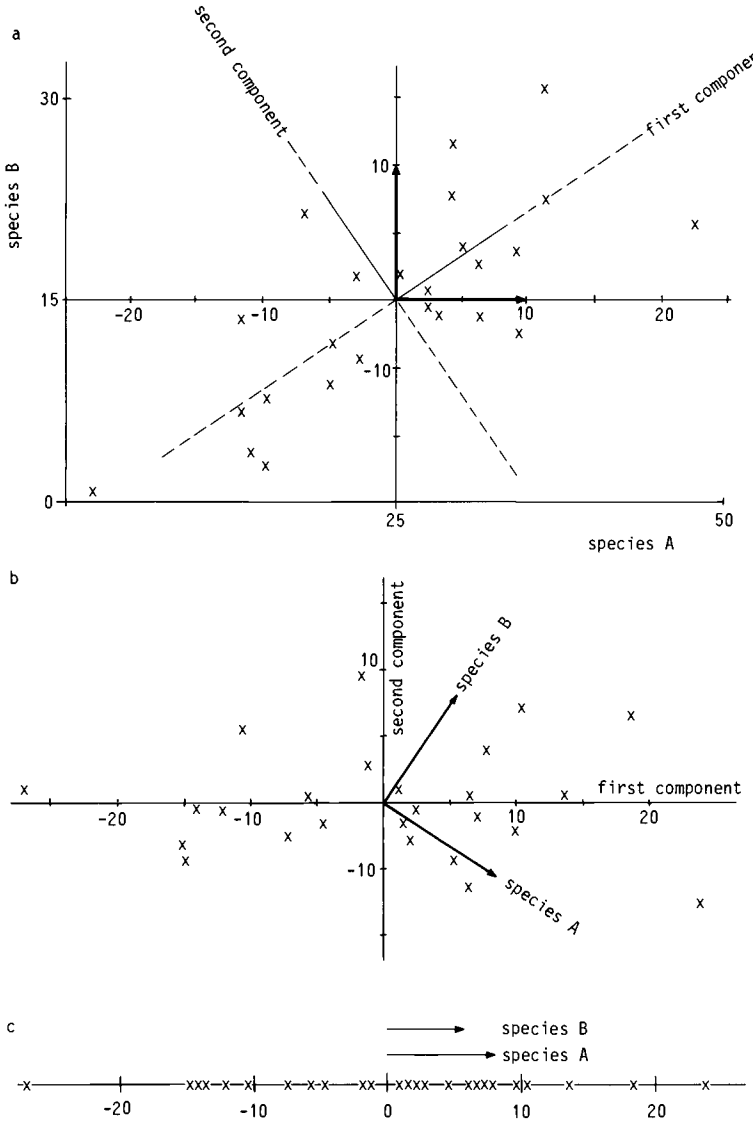


Figure 5.13 Artificial abundance data for two species A and B at 25 sites. a: First principal component, running through the centroid of the sites in the direction of the greatest variance (at 34° of the axis of Species A). b: Rotated version of Figure a with the first principal component horizontally. c: One-dimensional PCA ordination with species represented by arrows. The scores are simply those of the first axis of b.

5.3.4 *Biplot of species and site scores*

The scores obtained from a PCA for species and sites can be used to prepare a biplot (Gabriel 1971). The biplot serves the same function as the joint plot in CA (Subsection 5.2.5), but the rules to interpret the biplot are rather different. We limit the discussion to the two-dimensional biplot as it is more difficult to visualize three-dimensional or higher ones. The prefix 'bi' in biplot refers to the joint representation of sites and species, and not to the dimension of the plot; for example, Figure 5.13c shows a one-dimensional biplot.

The ranges of the scores for sites and for species (scores and loadings) in PCA are often of a different order of magnitude. For example in Table 5.3c, the range of the species scores is 17.9 whereas the range of the site scores is 0.8. A biplot is therefore constructed most easily by drawing separate plots of sites and of species on transparent paper, each one with its own scaling. In each of the plots, the scale unit along the vertical axis must have the same physical length as the scale unit along the horizontal axis, as in CA. A biplot is obtained by superimposing the plots with the axes aligned. A biplot may therefore have different scale units for the sites (x scale) and species (b scale). Figures 5.12 and 5.15 provide examples for the Dune Meadow Data.

In Subsection 5.3.1, we showed that for each species PCA fits a straight line in one dimension to the (centred) abundances of the species (Figure 5.11; Equation 5.10) and in two dimensions a plane with respect to the PCA axes (Figure 3.11; Equation 5.11). The abundance of a species as fitted by PCA thus changes linearly across the biplot. We represent the fitted planes in a biplot by arrows as shown in Figure 5.12. The direction of the arrow indicates the direction of steepest ascent of the plane, i.e. the direction in which the abundance of the corresponding species increases most, and the length of the arrow equals the rate of change in that direction. In the perpendicular direction, the fitted abundance is constant. The arrows are obtained by drawing lines that join the species points to the origin, the point with coordinates (0,0).

The fitted abundances of a species can be read from the biplot in very much the same way as from a scatter diagram, i.e. by projecting each site onto the axis of the species. (This is clear from Figure 5.13a.) The axis of a species in a biplot is in general, however, not the horizontal axis or the vertical axis, as in Figure 5.13a, but an oblique axis, the direction of which is given by the arrow of the species. As an example of how to interpret Figure 5.12, some of the site points are projected onto the axis of *Agrostis stolonifera* in Figure 5.14. Without doing any calculations, we can see the ranking of the fitted abundances of *A. stolonifera* among the sites from the order of the projection points of the sites along the axis of that species. From Figure 5.14, we thus infer that the abundance of *A. stolonifera* is highest at Site 16, second highest at Site 13, and so on to Site 6, which has the lowest inferred abundance. The inferred ranking is not perfect when compared with the observed ranking, but not bad either.

Another useful rule to interpret a biplot is that the fitted value is positive if the projection point of a site lies, along the species' axis, on the same side of the origin as the species point does, and negative if the origin lies between the

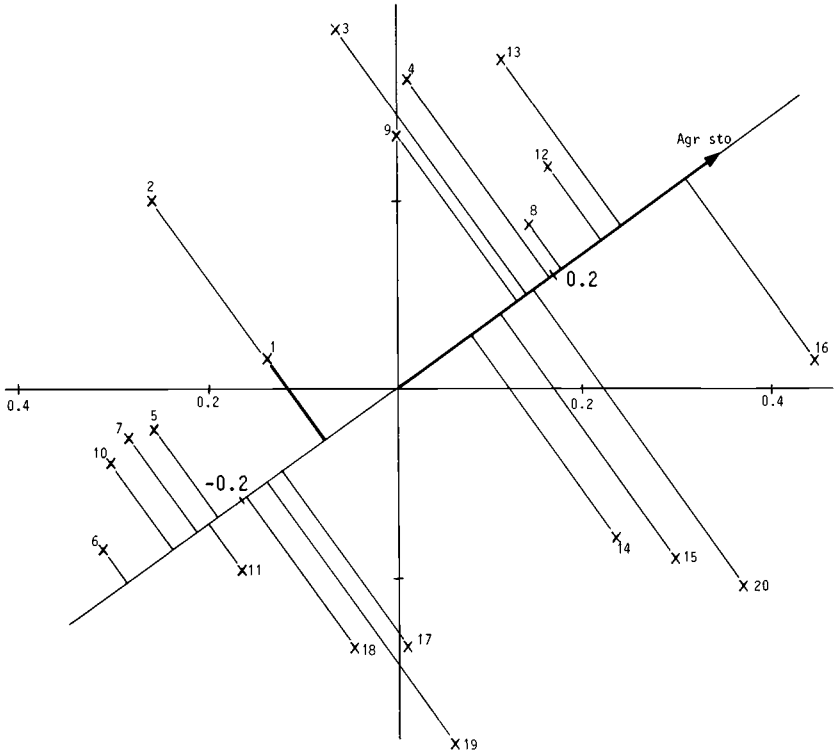


Figure 5.14 Biplot interpretation of Figure 5.12 for *Agrostis stolonifera*. For explanation, see text.

projection point and the species point. As we have centred the abundance data, the fitted abundance is higher than the species mean in the former case and lower than the species mean in the latter case. For example, Site 3 and Site 20 are inferred to have a higher than average abundance of *A. stolonifera*, whereas Sites 2 and 19 are inferred to have a lower than average abundance of this species. These inferences are correct, as can be seen from Table 5.5. One can also obtain quantitative values for the abundances as represented in the biplot, either algebraically with Equation 5.11 or geometrically as follows (ter Braak 1983). For this, we need the distance of the species point from the origin. In Figure 5.12, we see from the b scale that *A. stolonifera* lies at a distance of about 10 from the origin. We need further the projection points of sites onto the species' axis (Figure 5.14). From the x scale, we see that, for example, the projection point of Site 20 lies a distance of about 0.2 from the origin. The fitted value is now about $10 \times 0.2 = 2$. Adding the mean of *A. stolonifera* (2.4), we obtain

4.4 as the fitted abundance for *A. stolonifera* at Site 20; the observed value is 5. This biplot accounts in this way for 51% of the variance in abundance values of all species. This value was computed at the end of Subsection 5.3.2. Note, however, that the fraction of variance accounted for usually differs among species. In general, the abundances of species that are far from the origin are better represented in the biplot than the abundances of species near the origin. For example, the fractions accounted for are 80% for *Agrostis stolonifera*, 78% for *Poa trivialis*, 25% for *Bromus hordaceus*, 4% for *Brachythecium rutabulum* and 3% for *Empetrum nigrum*.

The scaling of the species and site scores in the biplot requires attention. From Equation 5.11, we deduce that scaling is rather arbitrary; for example, the fitted values remain the same if we jointly plot the species points $(3b_{k1}, 5b_{k2})$ and the site points $(x_{i1}/3, x_{i2}/5)$. Yet, there are two types of scaling that have special appeal.

In the first type of scaling, the site scores are standardized to unit sum of squares and the species scores are weighted sums of the site scores (Table 5.6). The sum of squared scores of species is then equal to the eigenvalue of the axis. In this scaling, the angle between arrows of each pair of species (Figure 5.12) provides an approximation of their pair-wise correlation, i.e.

$$r \approx \cos \theta$$

with r the correlation coefficient and θ the angle.

Consequently, arrows that point in the same direction indicate positively correlated species, perpendicular arrows indicate lack of correlation and arrows pointing in the opposite direction indicate negatively correlated species. This biplot is termed the covariance biplot and is considered in detail by Corsten & Gabriel (1976).

In the second type of scaling, the species scores are standardized to unit sum of squares and the site scores are standardized, so that their sum of squares equals the eigenvalue of each axis. Then, the site scores are the weighted sum of the species scores. This scaling was used implicitly in Subsection 5.3.3 and is intended to preserve Euclidean Distances between sites (Equation 5.16), i.e. the length of the line segment joining two sites in the biplot then approximates the length of the line segment joining the sites in m-dimensional space, the axes of which are formed by the species. When scaled in this way, the biplot is termed a Euclidean Distance biplot (ter Braak 1983). Figure 5.15 shows this biplot for the Dune Meadow Data.

The Euclidean Distance biplot is obtained from the covariance biplot by simple rescaling of species and site scores. Species k with coordinates (b_{k1},b_{k2}) in the covariance biplot gets the coordinates $(b_{k1}/\sqrt{\lambda_1},b_{k2}/\sqrt{\lambda_2})$ in the Euclidean Distance biplot, and site i with coordinates (x_{i1},x_{i2}) gets coordinates $(x_{i1}\sqrt{\lambda_1},x_{i2}\sqrt{\lambda_2})$ in the Euclidean Distance biplot. Figure 5.15 does not look very different from Figure 5.12, because the ratio of $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$ is close to 1.

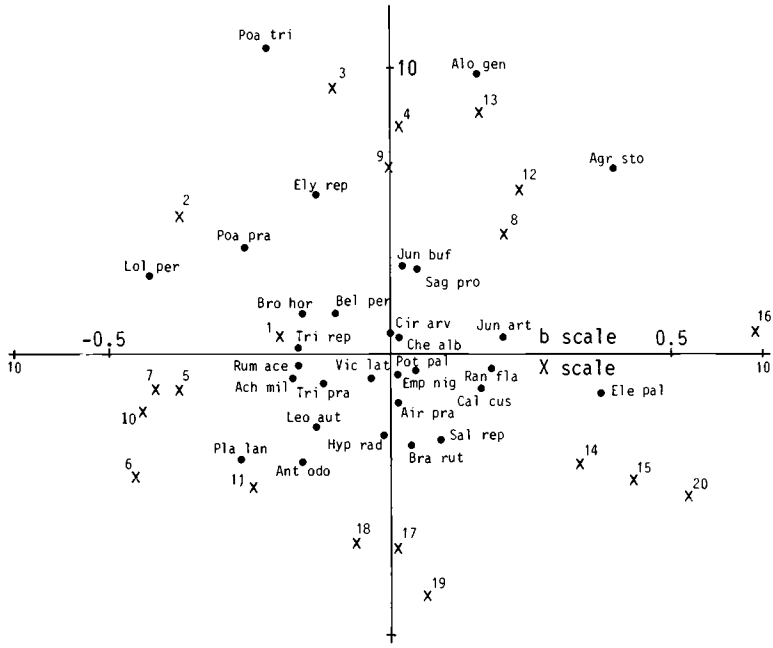


Figure 5.15 Euclidean Distance biplot of Dune Meadow Data.

5.3.5 Data transformation

We have so far described the standard form of PCA as treated in statistical textbooks (e.g. Morrison 1967). In ecology, this form is known as 'species-centred PCA'. In a variant of this, 'standardized PCA', abundances of each species are also divided by its standard deviation. In species-centred PCA, each species is implicitly weighted by the variance of its abundance values. Species with high variance, often the abundant ones, therefore dominate the PCA solution, whereas species with low variance, often the rare ones, have only minor influence on the solution. This may be reason to apply standardized PCA, in which all species receive equal weight. However the rare species then unduly influence the analysis if there are a lot of them, and chance can dominate the results. We therefore recommend species-centred PCA, unless there is strong reason to use standardized PCA. Standardization is necessary if we are analysing variables that are measured in different units, for example quantitative environmental variables such as pH, mass fraction of organic matter or ion concentrations. Noy Meir et al. (1975) fully discuss the virtues and vices of various data transformations in PCA.

The fraction of variance accounted for by the first few axes is not a measure

of the appropriateness of a particular data transformation. By multiplying the abundances of a single species by a million, the first axis of a species-centred PCA will in general account for nearly all the variance, just because nearly all the variance after this transformation is due to this species and the first axis almost perfectly represents its abundances.

If some environmental variables are known to influence the species data strongly, the axes of a PCA will probably show what is already known. To detect unknown variation, one can for each species first apply a regression on the known environmental variables, collect the residuals from these regressions in a two-way table and apply PCA to this table of residuals. This analysis is called partial PCA and is standardly available in the computer program CANOCO (ter Braak 1987b). The analysis is particularly simple if, before sampling, groups of sites are recognized. Then, the deviations of the group means should be analysed instead of the deviations from the general mean. An example is the analysis of vegetation change in permanent plots by Swaine & Greig-Smith (1980).

5.3.6 R-mode and Q-mode algorithms

The iteration algorithm in Table 5.6 is a general-purpose algorithm to extract eigenvectors and eigenvalues from an $m \times n$ matrix **Y** with elements y_{ki}. The algorithm is used in the computer program CANOCO (ter Braak 1987b) to obtain the solution to species-centred PCA if the rows are centred and to standardized PCA if the rows are standardized, but also to non-centred PCA (Noy Meir 1973) if the data are neither centred nor standardized. However many computer programs for PCA use other algorithms, most of which implicitly transform the data. Centring by variables is done implicitly when PCA is carried out on the matrix of covariances between the variables. Also, standardization by variables is implicit in an analysis of the correlation matrix. The role of species in our discussion therefore corresponds to the role of variables in a general-purpose computer program for PCA. The rest of Subsection 5.3.6 may be skipped at a first reading.

Algorithms that are based on the covariance matrix or correlation matrix are termed R-mode algorithms. More generally, R-mode algorithms extract eigenvectors from the species-by-species cross-product matrix **A** with elements

$$a_{kl} = \Sigma_i \, y_{ki} \, y_{li} \quad (k, l = 1, ..., m)$$

where, as before, y_{ki} is the data after transformation.

By contrast, Q-mode algorithms extract eigenvectors from the site-by-site cross-product matrix **C** with elements

$$c_{ij} = \Sigma_k \, y_{ki} \, y_{kj} \quad (i, j = 1, ..., n).$$

A particular Q-mode algorithm is obtained from Table 5.6 by inserting Equation 5.8 in Equation 5.9. In this way, Steps 2 and 3 are combined into a single step, in which

new $x_i = \Sigma_{j=1}^{n} c_{ij} x_j$.

It can be shown that the eigenvalues of the Matrix **A** equal those of the Matrix **C**, and further that the eigenvectors of **C** can be obtained from those of **A** by applying Equation 5.9 to each eigenvector and, conversely, that the eigenvectors of **A** can be obtained from those of **C** by applying Equation 5.8 to each eigenvector of **C**. The terms R-mode and Q-mode therefore refer to different algorithms and not to different methods. If the number of species is smaller than the number of sites, R-mode algorithms are more efficient than Q-mode algorithms, and conversely.

5.4 Interpretation of ordination with external data

Once data on species composition have been summarized in an ordination diagram, the diagram is typically interpreted with help of external knowledge on sites and species. Here we discuss methods that facilitate interpretation when data on environmental variables are collected at different sites. Analogous methods exist when there is external data on the species, for example growth form of plant species or indicator values for environmental variables from previous studies or from the literature (Table 5.7).

Simple interpretative aids include:

- writing the values of an environmental variable in the order of site scores of an ordination axis below the arranged species data table (Table 5.7)
- writing the values of an environmental variable near the site points in the ordination diagram (Figure 5.16)
- plotting the site scores of an ordination axis against the values of an environmental variable (Figure 5.17)
- calculating (rank) correlation coefficients between each of the quantitative environmental variables and each of the ordination axes (Table 5.8)
- calculating mean values and standard deviations of ordination scores for each class of a nominal environmental variable (ANOVA, Subsection 3.2.1) and plotting these in the ordination diagram (Figure 5.16).

An ordination technique that is suited for the species composition data extracts theoretical environmental gradients from these data. We therefore expect straight line (or at least monotonic) relations between ordination axes and quantitative environmental variables that influence species. Correlation coefficients are therefore often adequate summaries of scatter plots of environmental variables against ordination axes.

Three of these simple interpretative aids are directed to the interpretation of axes instead of to the interpretation of the diagram as a whole. But the ordination axes do not have a special meaning. Interpretation of other directions in the diagram is equally valid. A useful idea is to determine the direction in the diagram that has maximum correlation with a particular environmental variable (Dargie 1984). For the jth environmental variable, z_j, that direction can be found by multiple (least-squares) regression of z_j on the site scores of the first ordination axis (x_1) and the second ordination axis (x_2), i.e. by estimating the parameters b_1 and

Table 5.7 Values of environmental variables and Ellenberg's indicator values of species written alongside the ordered data table of the Dune Meadow Data, in which species and sites are arranged in order of their scores on the second DCA axis. A1: thickness of A1 horizon (cm), 9 meaning 9 cm or more; moisture: moistness in five classes from 1 = dry to 5 = wet; use: type of agricultural use, 1 = hayfield, 2 = a mixture of pasture and hayfield, 3 = pasture; manure: amount of manure applied in five classes from 0 = no manure to 5 = heavy use of manure. The meadows are classified by type of management: SF, standard farming; BF, biological farming; HF, hobby farming; NM, nature management; F, R, N refer to Ellenberg's indicator values for moisture, acidity and nutrients, respectively.

Species	Site (i)	F	R	N
	00001000010111111121			
k	12350749638162485709			
1 Ach mil	13 242 2 2	4		5
11 Ely rep	4444 46	5		8
7 Bro hor	4 2423			3
17 Lol per	756266526 47 2	5		7
6 Bel per	3222 2 2			5
19 Poa pra	445244443244 3 1	5		
20 Poa tri	27664555494 24	7		7
26 Tri pra	2 2 5			
23 Rum ace	5 3 26 2			5
21 Pot pal	2 2	10	3	2
18 Pla lan	535 5 3	3		2
4 Alo gen	27 23 55 48	9	7	7
8 Che alb	1	4		7
10 Ele pal	4 8 4 5 4	10		
27 Tri rep	52262135223 3621 2			7
2 Agr sto	4 83 54 744 4 5	6		5
15 Jun buf	2 4 3 4	7	3	
30 Cal cus	3 4 3			
14 Jun art	4 4 3 3 4	8		2
9 Cir arv	2			7
22 Ran fla	22 2 2 2 4	9	3	2
28 Vic lat	1 2 1	2	3	2
29 Bra rut	2222226 2444 64 43			
16 Leo aut	52333223235 2252226	5		5
5 Ant odo	442 3 4 4		5	
24 Sag pro	52 222 4 3	6	7	6
25 Sal rep	3 53			
13 Hyp rad	2 2 5	5	4	3
3 Air pra	2 3	2	3	1
12 Emp nig	2	6	4	
j				
1 A1	34463344464466959444			
2 moisture	11212124155154515255			
3 use	22211321223332312111			
4 manure	42421341233132000000			
5 SF	10100010010011000000			
6 BF	01001000000100000000			
7 HF	00010101101000000000			
8 NM	00000000000000111111			

Table 5.8 Correlation coefficients ($100 \times r$) of environmental variables with the first four DCA axes for the Dune Meadow Data.

Variable	Axes			
	1	2	3	4
1 A1	58	24	7	9
2 moisture	76	57	7	-7
3 use	35	-21	-3	-5
4 manure	6	-68	-7	-64
5 SF	22	-29	5	-60
6 BF	-28	-24	39	22
7 HF	-22	-26	-55	-14
8 NM	21	73	17	56
Eigenvalue	0.54	0.29	0.08	0.05

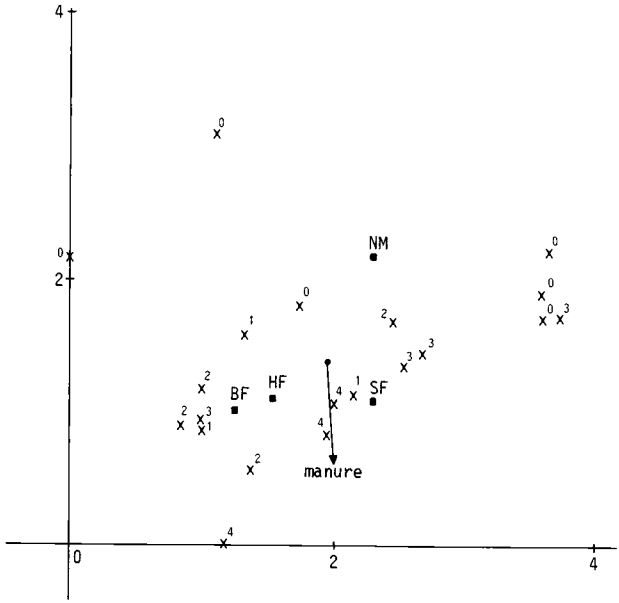


Figure 5.16 The amount of manure written on the DCA ordination of Figure 5.7. The trend in the amount across the diagram is shown by an arrow, obtained by a multiple regression of manure on the site scores of the DCA axes. Also shown are the mean scores for the four types of management, which indicate, for example, that the nature reserves (NM) tend to lie at the top of the diagram.

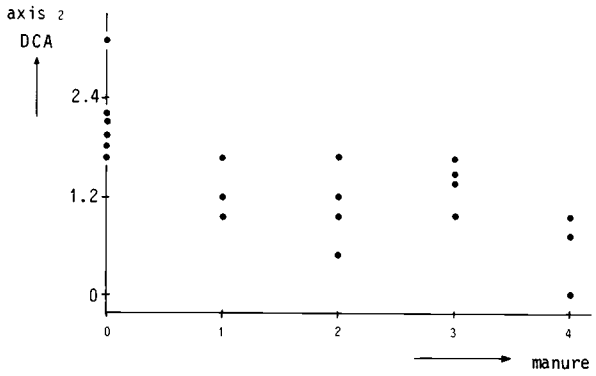


Figure 5.17 Site scores of the second DCA axis plotted against the amount of manure.

b_2 of the regression equation (as in Subsection 3.5.2)

$$Ez_j = b_0 + b_1 x_1 + b_2 x_2 \qquad \text{Equation 5.12}$$

The direction of maximum correlation makes an angle of θ with the first axis where θ = arctan (b_2/b_1) and the maximum correlation equals the multiple correlation coefficient (Subsection 3.2.1). This direction can be indicated in the ordination diagram by an arrow running from the centroid of the plot, for instance with coordinates (0,0), to the point with coordinates (b_1,b_2), as illustrated for manure in Figure 5.16. This is an application of the biplot idea; the environmental variable is represented in the diagram by an arrow that points in the direction of maximum change (Subsection 5.3.4). Several environmental variables can be accommodated in this way in a single ordination diagram.

In Chapter 3, presence and abundance of a single species represented the response variable to be explained by the environmental variables. By applying an ordination technique to the abundances of many species, we have reduced many response variables to a few ordination axes. It is therefore natural to consider the ordination axes as new derived response variables and to attempt to explain each of them by use of multiple regression analysis. For example, we can fit for the first axis (x_1) the response model

$$Ex_1 = c_0 + c_1 z_1 + c_2 z_2 + \ldots + c_q z_q \qquad \text{Equation 5.13}$$

where z_j is the jth (out of q) environmental variables and c_j is the corresponding regression coefficient. The multiple correlation coefficient and the fraction of variance accounted for by the regression (Subsection 3.2.1) indicate whether the environmental variables are sufficient to predict the variation in species composition that is represented by the first ordination axis. Table 5.9 shows an example.

Table 5.9 Multiple regression of the first CA axis on four environmental variables of the dune meadow data, which shows that moisture contributes significantly to the explanation of the first axis, whereas the other variables do not.

Term	Parameter		Estimate	s.e.	t
constant	c_0		-2.32	0.50	-4.62
A1	c_1		0.14	0.08	1.71
moisture	c_2		0.38	0.09	4.08
use	c_3		0.31	0.22	1.37
manure	c_4		-0.00	0.12	-0.01
ANOVA table					
	d.f.	s.s.	m.s.	F	
Regression	4	17.0	4.25	10.6	
Residual	15	6.2	0.41		
Total	19	23.2	1.22		
$R^2 = 0.75$					
$R^2_{adj} = 0.66$					

There are good reasons not to include the environmental variables in the ordination analysis itself, nor to reverse the procedure by applying ordination to the environmental data first and by adding the species data afterwards: the main variation in the environmental data is then sought, and this may well not be the major variation in species composition. For example, if a single environmental variable is important for the species and many more variables are included in the analysis, the first few axes of the environmental ordination mainly represent the relations among the unimportant variables and the relation of the important variable with the species' data would not be discovered. It is therefore better to search for the largest variation in the species' data first and to find out afterwards which of the environmental variables is influential.

5.5 Canonical ordination

5.5.1 Introduction

Suppose we are interested in the effect on species composition of a particular set of environmental variables. What can then be inferred from an indirect gradient analysis (ordination followed by environmental gradient interpretation)? If the ordination of the species data can be readily interpreted with these variables, the environmental variables are apparently sufficient to explain the main variation in the species' composition. But, if the environmental variables cannot explain the main variation, they may still explain some of the remaining variation, which

can be substantial, especially in large data sets. For example, a strong relation of the environmental variables with the fifth ordination axis will go unnoticed, when only four ordination axes are extracted, as in some of the computer programs in common use. This limitation can only be overcome by canonical ordination.

Canonical ordination techniques are designed to detect the patterns of variation in the species data that can be explained 'best' by the observed environmental variables. The resulting ordination diagram expresses not only a pattern of variation in species composition but also the main relations between the species and each of the environmental variables. Canonical ordination thus combines aspects of regular ordination with aspects of regression.

We introduce, consecutively, the canonical form of CA, the canonical form of PCA (redundancy analysis) and two other linear canonical techniques, namely canonical correlation analysis and canonical variate analysis. After introducing these particular techniques, we discuss how to interpret canonical ordination axes and the possible effect of data transformations.

5.5.2 Canonical correspondence analysis (CCA)

To introduce canonical correspondence analysis (CCA), we consider again the artificial example by which we have introduced CA (Subsection 5.2.1). In this example (reproduced in Figure 5.18a), five species each preferred a slightly different moisture value. The species score was defined to be the value most preferred and was calculated by averaging the moisture values of the sites in which the species is present. Environmental variables were standardized to mean 0 and variance 1 (Table 5.2c) and the dispersion of the species scores after standardization was taken to express how well a variable explains the species data.

Now suppose, as before, that moisture is the best single variable among the environmental variables measured. In Subsection 5.2.1, we proceeded by constructing the theoretical variable that best explains the species data and, in Section 5.4, we attempted to explain the variable so obtained by a combination of measured environmental variables (Equation 5.13). But, as discussed in Subsection 5.5.1, such attempts may fail, even if we measure environmental variables influencing the species. So why not consider combinations of environmental variables from the beginning? In the example, someone might suggest considering a combination of moisture and phosphate, and Figure 5.18b actually shows that, after standardization, the combination ($3 \times$ moisture $+ 2 \times$ phosphate) gives a larger dispersion than moisture alone. So it can be worthwhile to consider not only the environmental variables singly but also all possible linear combinations of them, i.e. all weighted sums of the form

$$x_i = c_0 + c_1 z_{1i} + c_2 z_{2i} + \ldots + c_q z_{qi} \qquad \text{Equation 5.14}$$

where
z_{ji} is the value of environmental variable j at site i
c_j is the weight (not necessary positive) belonging to that variable
x_i is the value of the resulting compound environmental variable at site i.

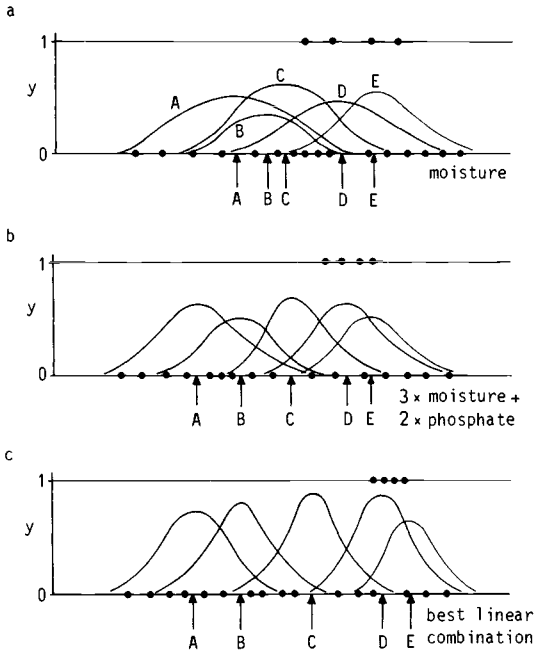


Figure 5.18 Artificial example of unimodal response curves of five species (A-E) with respect to standardized environmental variables showing different degrees of separation of the species curves. a: Moisture. b: Linear combination of moisture and phosphate, chosen a priori. c: Best linear combination of environmental variables, chosen by CCA. Sites are shown as dots, at $y = 1$ if Species D is present and at $y = 0$ if Species D is absent.

CCA is now the technique that selects the linear combination of environmental variables that maximizes the dispersion of the species scores (Figure 5.18c; ter Braak 1987a). In other words, CCA chooses the best weights (c_j) for the environmental variables. This gives the first CCA axis.

The second and further CCA axes also select linear combinations of environmental variables that maximize the dispersion of the species scores, but subject to the constraint of being uncorrelated with previous CCA axes (Subsection 5.2.1). As many axes can be extracted as there are environmental variables.

CA also maximizes the dispersion of the species scores, though irrespective of any environmental variable; that is, CA assigns scores (x_i) to sites such that the dispersion is absolutely maximum (Subsection 5.2.1). CCA is therefore 'restricted correspondence analysis' in the sense that the site scores are restricted to be a linear combination of measured environmental variables (Equation 5.14). By incorporating this restriction in the two-way weighted averaging algorithm

of CA (Table 5.2), we obtain an algorithm for CCA. More precisely, in each iteration cycle, a multiple regression must be carried out of the site scores obtained in Step 3 on the environmental variables (for technical reasons with y_{+i}/y_{++} as site weights). The fitted values of this regression are by definition a linear combination of the environmental variables (Equation 5.14) and are thus the new site scores to continue with in Step 4 of Table 5.2a. As in CA, the scores stabilize after several iterations and the resulting scores constitute an ordination axis of CCA. The corresponding eigenvalue actually equals the (maximized) dispersion of the species scores along the axis. The eigenvalues in CCA are usually smaller than those in CA because of the restrictions imposed on the site scores in CCA.

The parameters of the final regression in the iteration process are the best weights, also called canonical coefficients, and the multiple correlation of this regression is called the species–environment correlation. This is the correlation between the site scores that are weighted averages of the species scores and the site scores that are a linear combination of the environmental variables. The species–environment correlation is a measure of the association between species and environment, but not an ideal one; axes with small eigenvalues may have misleadingly high species–environment correlations. The importance of the association is expressed better by the eigenvalue because the eigenvalue measures how much variation in the species data is explained by the axis and, hence, by the environmental variables.

CCA is restricted correspondence analysis but the restrictions become less strict the more environmental variables are included in the analysis. If $q \geqslant n - 1$, then there are actually no restrictions any more; CCA is then simply CA. The arch effect may therefore crop up in CCA, as it does in CA (Gauch 1982). The method of detrending (Hill & Gauch 1980) can be used to remove the arch and is available in the computer program CANOCO (ter Braak 1987b). But in CCA, the arch can be removed more elegantly by dropping superfluous environmental variables. Variables that are highly correlated with the arched axis (often the second axis) are most likely to be superfluous. So a CCA with the superfluous variables excluded does not need detrending.

In Subsection 5.2.7, we saw that CA approximated the maximum likelihood solution of Gaussian ordination when Conditions A1-A4 hold true. If we change the Gaussian ordination model by stating that the site scores must be a linear combination of the environmental variables, the maximum likelihood solution of the model so obtained is again approximated by CCA when these conditions hold true (ter Braak 1986a). The data on species composition are thus explained by CCA through a Gaussian response model in which the explanatory variable is a linear combination of the environmental variables. Furthermore, tests of real data showed that CCA is extremely robust when these assumptions do not hold. The vital assumption is that the response model is unimodal. For a simpler model where relations are monotonic, the results can still be expected to be adequate in a qualitative sense, but for more complex models the method breaks down.

As an example, we use the Dune Meadow Data, which concerns the impact of agricultural use on vegetation in dune meadows on the Island of Terschelling (the Netherlands). The data set consists of 20 relevés, 30 plant species (Table

0.1) and 5 environmental variables (Table 0.2), one of which is the nominal variable 'type of management' consisting of four classes. CCA can accommodate nominal explanatory variables by defining dummy variables as in multiple regression (Subsection 3.5.5). For instance, the dummy variable 'nature management' (Table 5.7) indicates that meadows received that type of management. The first eigenvalue of CCA is somewhat lower than that of CA (0.46 compared to 0.54). Multiple regression of the site scores of the first CA axis on the environmental variables, as we proposed in Section 5.4, resulted in a multiple correlation of 0.87. If the multiple regression is carried out within the iteration algorithm, as in CCA, the multiple correlation increases to 0.96, which is the species-environment correlation. The CCA scores for species and sites look similar to those of CA: not surprisingly, since the multiple correlation obtained with CA is already high. We conclude that, in this example, the measured environmental variables account for the main variation in the species composition. This is true for the second axis also. The second eigenvalue of CCA is 0.29, compared to 0.40 for CA and the second species-environment correlation is 0.89, compared to a multiple correlation of 0.83 in CA. Table 5.10 shows the canonical coefficients that define the first two axes and the correlations of the environmental variables with these axes. These correlations are termed intra-set correlations to distinguish them from the inter-set correlations, which are the correlations between the environmental variables and the site scores that are derived from the species scores. (The inter-set correlation is R times the intra-set correlation; R is the species-environment correlation of the axis). From the correlations in Table 5.10, we infer that the first axis is a moisture gradient and that the second axis is a manuring axis, separating the meadows managed as a nature reserve from the standardly farmed meadows. This can be seen also from the CCA ordination diagram (Figure 5.19a).

Table 5.10 Canonical correspondence analysis: canonical coefficients ($100 \times c$) and intra-set correlations ($100 \times r$) of environmental variables with the first two axes of CCA for the Dune Meadow Data. The environmental variables were standardized first to make the canonical coefficients of different environmental variables comparable. The class SF of the nominal variable 'type of management' was used as reference class in the analysis (Subsection 3.5.5).

Variable	Coefficients		Correlations	
	Axis 1	Axis 2	Axis 1	Axis2
A1	9	-37	57	-17
moisture	71	-29	93	-14
use	25	5	21	-41
manure	-7	-27	-30	-79
SF	-	-	16	-70
BF	-9	16	-37	15
HF	18	19	-36	-12
NM	20	92	56	76

The species and sites are positioned as points in the CCA diagram as in CA and their joint interpretation is also as in CA; sites with a high value of a species tend to be close to the point for that species (Subsection 5.2.5). The environmental variables are represented by arrows and can be interpreted in conjunction with the species points as follows. Each arrow determines an axis in the diagram and the species points must be projected onto this axis. As an example, the points of a few species are projected on to the axis for manuring in Figure 5.19b. The order of the projection points now corresponds approximately to the ranking of the weighted averages of the species with respect to amount of manure. The weighted average indicates the 'position' of a species curve along an environmental variable (Figure 5.18a) and thus the projection point of a species also indicates this position, though approximately. Thus *Cirsium arvense*, *Alopecurus geniculatus*, *Elymus repens* and *Poa trivialis* mainly occur in these data in the highly manured meadows, *Agrostis stolonifera* and *Trifolium repens* in moderately manured meadows and *Ranunculus flammula* and *Anthoxanthum odoratum* in meadows with little manuring. One can interpret the other arrows in a similar way. From Figure 5.19a, one can see at a glance which species occur mainly in wetter conditions (those on the right of the diagram) and which prefer drier conditions (those on the left of the diagram).

The joint plot of species points and environmental arrows is actually a biplot that approximates the weighted averages of each of the species with respect to each of the environmental variables. The rules for quantitative interpretation of the CCA biplot are the same as for the PCA biplot described in Subsection 5.3.4. In the diagram, the weighted averages are approximated as deviations from the grand mean of each environmental variable; the grand mean is represented by the origin (centroid) of the plot. A second useful rule to interpret the diagram is therefore that the inferred weighted average is higher than average if the projection point lies on the same side of the origin as the head of an arrow and is lower than average if the origin lies between the projection point and the head of an arrow. As in Subsection 5.3.2, a measure of goodness of fit is $(\lambda_1 + \lambda_2)/(\text{sum of all eigenvalues})$, which expresses the fraction of variance of the weighted averages accounted for by the diagram. In the example, Figure 5.19a accounts for 65% of the variance of the weighted averages. (The sum of all canonical eigenvalues is 1.177.)

The positions of the heads of the arrows depend on the eigenvalues and on the intra-set correlations. In Hill's scaling (Subsection 5.2.2), the coordinate of the head of the arrow for an environmental variable on axis s is $r_{js}\sqrt{\lambda_s (1 - \lambda_s)}$, with r_{js} the intra-set correlation of environmental variable j with axis s and λ_s is the eigenvalue of axis s. The construction of biplots for detrended canonical correspondence analysis is described by ter Braak (1986a). Environmental variables with long arrows are more strongly correlated with the ordination axes than those with short arrows, and therefore more closely related to the pattern of variation in species composition shown in the ordination diagram.

Classes of nominal environmental variables can also be represented by arrows (ter Braak 1986a). The projection of a species on such an arrow approximates the fraction of the total abundance of that species that is achieved at sites of

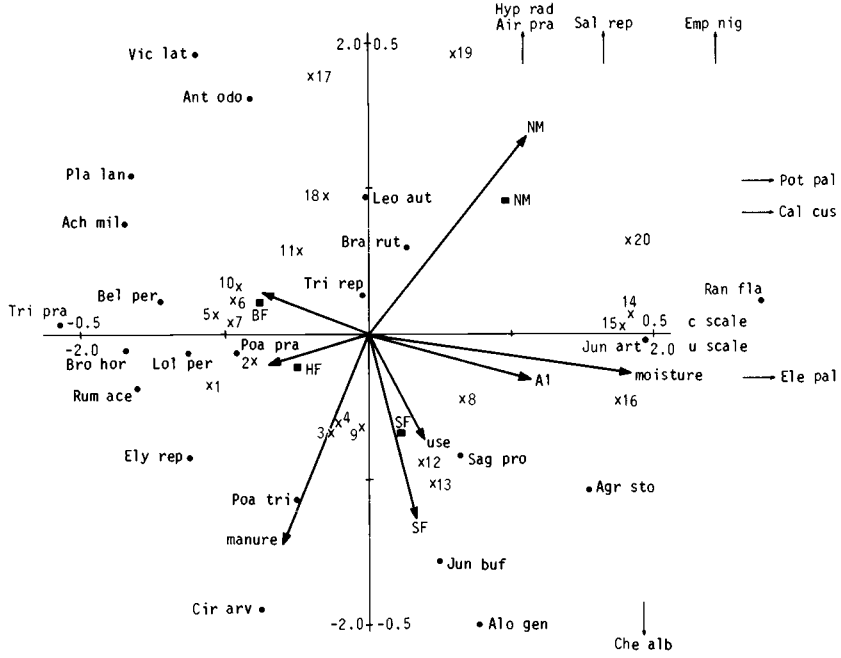


Figure 5.19 CCA of the Dune Meadow Data. a: Ordination diagram with environmental variables represented by arrows. The *c* scale applies to environmental variables, the *u* scale to species and sites. The types of management are also shown by closed squares at the centroids of the meadows of the corresponding types of management. b: Inferred ranking of the species along the variable amount of manure, based on the biplot interpretation of Part a of this figure.

that class. However it is sometimes more natural to represent each class of a nominal variable by a point at the centroid (the weighted average) of the sites belonging to that class (Figure 5.19a). Classes consisting of sites with high values for a species are then positioned close to the point of that species. In Figure 5.19a, the meadows managed as a nature reserve are seen to lie at the top-right of the diagram; the meadows of standard farms lie at the bottom.

A second example (from ter Braak 1986a) concerns the presence or absence of 133 macrophytic species in 125 freshwater ditches in the Netherlands. The first four axes of detrended correspondence analyses (DCA) were poorly related (multiple correlation $R < 0.60$) to the measured environmental variables, which were: electrical conductivity (κ), orthophosphate concentration (PHOSPHATE), both transformed to logarithms, chloride ratio (CHLORIDE, the share of chloride ions in κ) and soil type (clay, peaty soil, sand). By choosing the axes in the

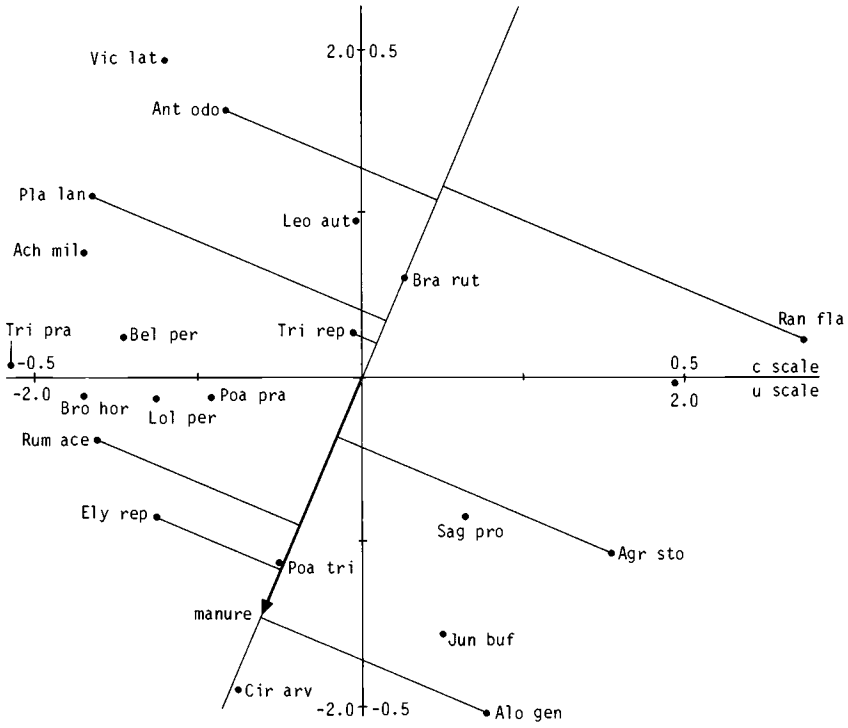


Figure 5.19b

light of these environmental variables by means of CCA, the multiple correlations increased considerably, R being 0.82 and 0.81 for the first two axes. The eigenvalues dropped somewhat - for the first two axes, from 0.34 and 0.25 in DCA to 0.20 and 0.13 in CCA. Apparently, the environmental variables are not sufficient to predict the main variation in species composition extracted by DCA, but they do predict a substantial part of the remaining variation. From the CCA ordination diagram (Figure 5.20), it can be seen that κ and PHOSPHATE are strongly correlated (> 0.8) with the first CCA axis. Species with a high positive score on that axis are therefore almost restricted to ditches with high κ and PHOSPHATE, and species with a large negative score to ditches with low κ and PHOSPHATE. Species with intermediate scores are either unaffected by κ and PHOSPHATE or restricted to intermediate values of κ and PHOSPHATE. The second CCA axis is strongly correlated ($r = 0.9$) with CHLORIDE. The arrow for PEAT shows that species whose distribution is the most restricted to peaty soils lie in the top-left corner of the diagram. The arrows for SAND and CLAY are to be interpreted analogously.

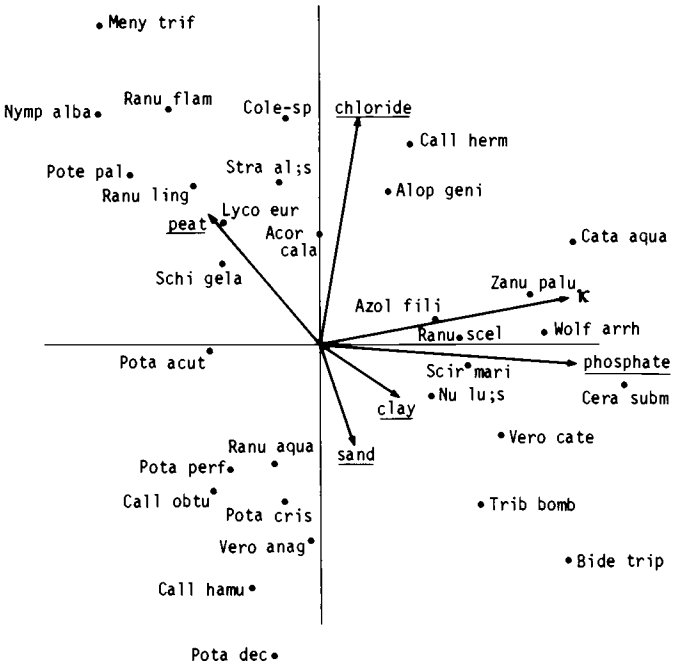


Figure 5.20 CCA ordination diagram of the ditch vegetation data (sites are not shown).

5.5.3 Redundancy analysis (RDA)

Redundancy analysis (RDA) is the canonical form of PCA and was invented by Rao (1964). RDA has so far been neglected by ecologists, but appears attractive when used in combination with PCA.

As in PCA (Subsection 5.3.1), we attempt to explain the data of all species by fitting a separate straight line to the data of each species. As a measure of how badly a particular environmental variable explains the species data, we take the total residual sum of squares, as in PCA (Figure 5.11). The best environmental variable is then the one that gives the smallest total residual sum of squares. From this, we can derive a canonical ordination technique, as in Subsection 5.5.2, by considering also linear combinations of environmental variables. RDA is the technique selecting the linear combination of environmental variables that gives the smallest total residual sum of squares.

PCA also minimizes the total residual sum of squares, but it does so without looking at the environmental variables (Subsection 5.3.1). We can obtain the RDA axes by extending the algorithm of PCA (Table 5.6) in a similar fashion to how

we modified the CA algorithm in Subsection 5.5.2; in each iteration cycle, the site scores calculated in Step 3 are regressed on the environmental variables with Equation 5.13 and the fitted values of the regression are taken as the new site scores to continue in Step 4 of the algorithm. (In contrast to CCA, we must now use equal site weights in the regression.) So the site scores are restricted to a linear combination of the environmental variables and RDA is simply PCA with a restriction on the site scores. The species–environment correlation is obtained in the same way as for CCA; but, in RDA, this correlation equals the correlation between the site scores that are weighted sums of the species scores and the site scores that are a linear combination of the environmental variables.

We illustrate RDA with the Dune Meadow Data, using the same environmental variables as in Subsection 5.5.2. The first two axes of PCA explained 29% and 21% of the total variance in the species data, respectively. RDA restricts the axes to linear combinations of the environmental variables and the RDA axes explain therefore less, namely 26% and 17% of the total variance. The first two species–environment correlations are 0.95 and 0.89, both a little higher than the multiple correlations resulting from regressing the first two PCA axes on the environmental variables. We conclude, as with CCA, that the environmental variables account for the main variation in the species composition. From the canonical coefficients and intra-set correlations (Table 5.11), we draw the same conclusions as with CCA, namely that the first axis is mainly a moisture gradient and the second axis a manuring gradient.

The RDA ordination diagram (Figure 5.21) can be interpreted as a biplot (Subsection 5.3.4). The species points and site points jointly approximate the species abundance data as in PCA, and the species points and environmental arrows

Table 5.11 Redundancy analysis: canonical coefficients (100 $\times$ c) and intra-set correlations (100 $\times$ r) of environmental variables with the first two axes of RDA for the Dune Meadow Data. The environmental variables were standardized first to make the canonical coefficients of different environmental variables comparable. The class SF of the nominal variable 'type of management' was used as reference class in the analysis (as in Table 5.10).

Variable	Coefficients		Correlations	
	Axis 1	Axis 2	Axis 1	Axis 2
A1	-1	-5	54	-6
moisture	15	9	92	12
use	5	-6	15	29
manure	-8	16	-26	86
SF	-	-	25	76
BF	-10	0	-48	-11
HF	-10	-2	-40	13
NM	-4	-13	51	-79

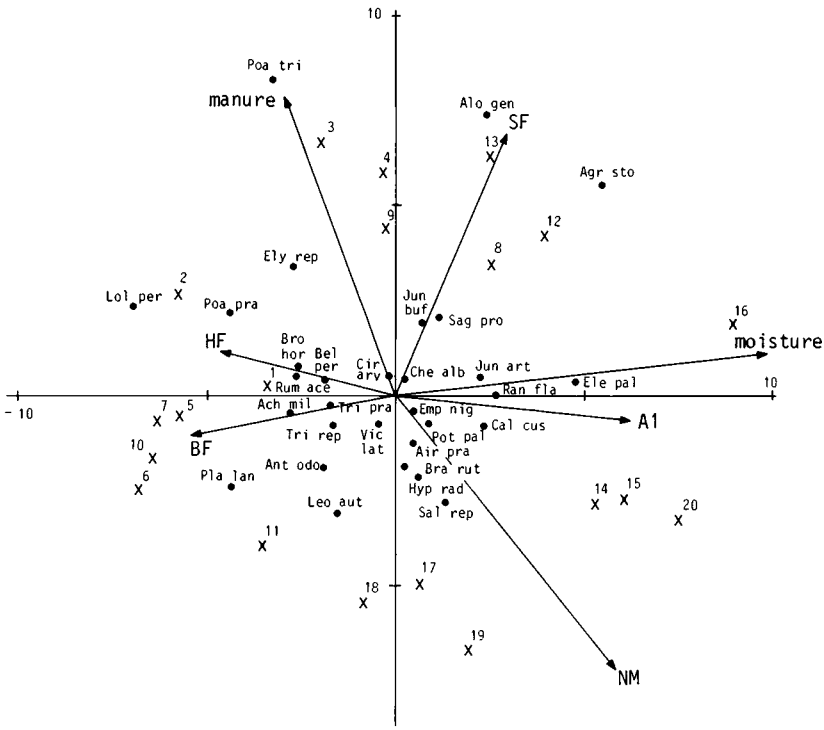


Figure 5.21 RDA ordination diagram of the Dune Meadow Data with environmental variables represented by arrows. The scale of the diagram is: 1 unit in the plot corresponds to 1 unit for the sites, to 0.067 units for the species and to 0.4 units for the environmental variables.

jointly approximate the covariances between species and environmental variables. If species are represented by arrows as well (a natural representation in a PCA biplot), the cosine of the angle between the arrows of a species and an environmental variable is an approximation of the correlation coefficient between the species and the environmental variable. One gets a qualitative idea of such correlations from the plot by noting that arrows pointing in roughly the same direction indicate a high positive correlation, that arrows crossing at right angles indicate near-zero correlation, and that arrows pointing in roughly opposite directions indicate a high negative correlation. If arrows are drawn for *Poa trivialis*, *Elymus repens* and *Cirsium arvense* in Figure 5.21, they make sharp angles with the arrow for manuring; hence, the abundances of these species are inferred to be positively correlated with the amount of manure. We can be more confident about this inference for *Poa trivialis* than for *Cirsium arvense* because the former species lies much further from the centre of the diagram than the latter species. As in

PCA, species at the centre of the diagram are often not very well represented and inferences from the diagram about their abundances and correlations are imprecise. From Figure 5.21, we infer also that, for instance, *Salix repens*, *Hypochaeris radicata* and *Aira praecox* are negatively correlated with the amount of manure.

A measure of goodness of fit of the biplot of species and environmental variables is $(\lambda_1 + \lambda_2)/(\text{sum of all eigenvalues})$, which expresses the fraction of variance of all covariances between species and environment accounted for by the diagram. For the example, Figure 5.21 accounts for 71% of this variance.

The scaling of Figure 5.21 conforms to that of the Euclidean distance biplot (Subsection 5.3.4): the sum of squares of the species scores is unity and site points are obtained by weighted summation of species scores. The positions of the heads of arrows of the environmental variables depend on the intra-set correlations (Table 5.11) and the eigenvalues. With this scaling, the coordinate of the head of the arrow for an environmental variable on axis s must be $r_{js}\sqrt{(\lambda_s/n)}$ where r_{js} is the intra-set correlation of environmental variable j with axis s, n is the number of sites, and λ_s the eigenvalue of axis s. The diagram scaled in this way gives not only a least-squares approximation of the covariances between species and environment, but also approximations of the (centred) abundances values, of the Euclidean Distances among the sites as based on the species data (Equation 5.15), and of covariances among the environmental variables, though the latter two approximations are not least-squares approximations. Other types of scaling are possible (ter Braak 1987b).

5.5.4 Canonical correlation analysis (COR)

The species–environment correlation was a by-product in CCA and RDA, but is central in canonical correlation analysis (COR). The idea of COR is to choose coefficients (scores) for species and coefficients for environmental variables so as to maximize the species–environment correlation. In COR, the species–environment correlation is defined as in RDA, as the correlation between site scores (x_i^*) that are weighted sums of species scores: ($x_i^* = \Sigma_k\ b_k\ y_{ki}$) and site scores ($x_i$) that are a linear combination of the environmental variables ($x_i = c_0 + \Sigma_j\ c_j\ z_{ji}$). An algorithm to obtain the COR axes is given in Table 5.12. The resulting species–environment correlation is termed the canonical correlation, and is actually the squareroot of the first eigenvalue of COR. Step 2 of the algorithm makes the difference from RDA: in RDA, the species scores are simply a weighted sum of the site scores, whereas in COR the species scores are parameters estimated by a multiple regression of the site scores on the species variables. This regression has the practical consequence that, in COR, the number of species must be smaller than the number of sites. It can be shown that the restriction on the number of species is even stronger than that: the number of species *plus* the number of environmental variables must be smaller than the number of sites. This requirement is not met in our Dune Meadow Data and is generally a nuisance in ecological research. By contrast, RDA and CCA set no upper limit to the number of species that can be analysed. Examples of COR can be found in Gittins

Table 5.12 An iteration algorithm for canonical correlation analysis (COR).

Step 1. Start with arbitrary initial site scores (x_i), not all equal to zero.
Step 2. Calculate species scores by multiple regression of the site scores on the species variables. The species scores (b_k) are the parameter estimates of this regression.
Step 3. Calculate new site scores (x_i^*) by weighted summation of the species scores(Equation 5.9). The site scores in fact equal the fitted values of the multiple regression of Step 2.
Step 4. Calculate coefficients for the environmental variables by multiple regression of the site scores (x_i^*) on the environmental variables. The coefficients (c_j) are the parameter estimates of this regression.
Step 5. Calculate new site scores (x_i) by weighted summation of the coefficients of the environmental variables, i.e. by $x_i = \Sigma_{j=1}^{q} c_j z_{ji}$. The site scores in fact equal the fitted values of the multiple regression of Step 4.
Step 6. For second and higher axes, orthogonalize the site scores (x_i) as in Table 5.6.
Step 7. Standardize the site scores (x_i) as in Table 5.6.
Step 8. Stop on convergence, i.e. when the new site scores are sufficiently close to the site scores of the previous cycle of the iteration process; ELSE go to Step 2.

(1985). COR allows a biplot to be made, from which the approximate covariances between species and environmental variables can be derived in the same way as in RDA (Subsection 5.5.3). The construction of the COR biplot is given in Subsection 5.9.3.

In our introduction to COR, species and environmental variables enter the analysis in a symmetric way (Table 5.12). Tso (1981) presented an asymmetric approach in which the environmental variables explain the species data. In this approach COR is very similar to RDA, but differs from it in the assumptions about the error part of the model (Equations 5.10 and 5.14): uncorrelated errors with equal variance in RDA and correlated normal errors in COR. The residual correlations between errors are therefore additional parameters in COR. When the number of species is large, there are so many of them that they cannot be estimated reliably from data from few sites. This causes practical problems with COR that are absent in RDA and CCA.

5.5.5 Canonical variate analysis (CVA)

Canonical variate analysis (CVA) belongs to the classical linear multivariate techniques along with PCA and COR. CVA is also termed linear discriminant analysis.

If sites are classified into classes or clusters, we may wish to know how the species composition differs among sites of different classes. If we have recorded the abundance values of a single species only, this question reduces to how much the abundance of the species differs between classes, a question studied in Subsection 3.2.1 by analysis of variance. If there are more species, we may wish to combine the abundance values of the species to make the differences between classes clearer than is possible on the basis of the abundance values of a single species. CVA does so by seeking a weighted sum of the species abundances; however not one

that maximizes the total variance along the first ordination axis, as PCA does, but one that maximizes the ratio of the between-class sum of squares and the within-class sum of squares of the site scores along the first ordination axis. (These sums of squares are the regression sum of squares and residual sum of squares, respectively, in an ANOVA of the site scores, cf. Subsection 3.2.1.).

Formally, CVA is a special case of COR in which the set of environmental variables consists of a single nominal variable defining the classes. So from Subsection 3.5.5, the algorithm of Table 5.12 can be used to obtain the CVA axes. We deduce that use of CVA makes sense only if the number of sites is much greater than the number of species and the number of classes (Schaafsma & van Vark 1979; Varmuza 1980). Consequently, many ecological data sets cannot be analysed by CVA without dropping many species. Examples of CVA can be found in Green (1979), Pielou (1984) and Gittins (1985).

In contrast to CVA, CCA and RDA can be used to display differences in species composition between classes without having to drop species from the analysis.

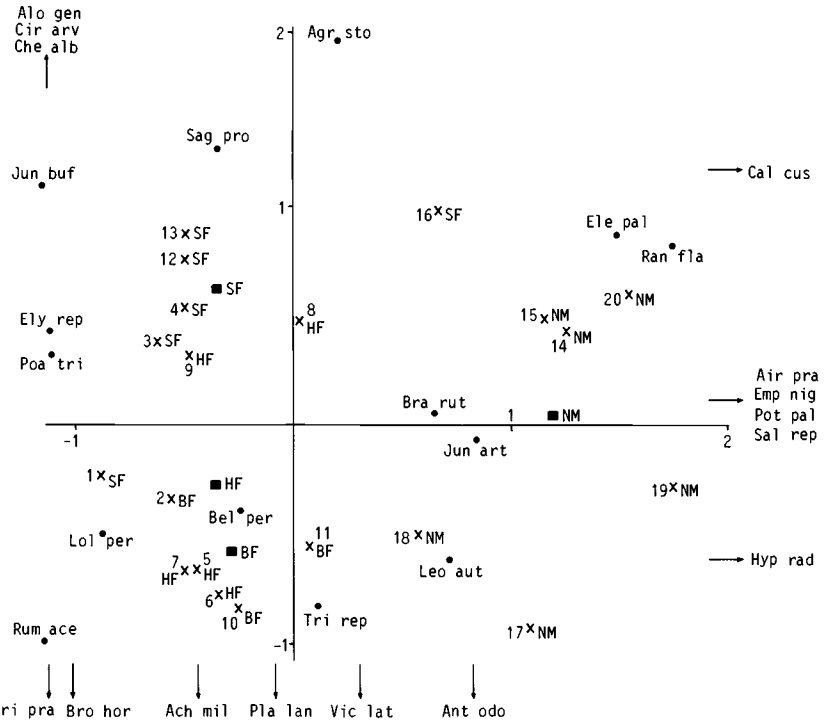


Figure 5.22 CCA ordination diagram of the Dune Meadow Data optimally displaying differences in species composition among different types of management.

For this, we must code classes as dummy environmental variables, as in Subsection 3.5.5. Such an analysis by CCA is equivalent to the analysis of concentration proposed by Feoli & Orlóci (1979). As an example, Figure 5.22 displays the differences in vegetation composition between the meadows receiving different types of management in our Dune Meadow Data. The first axis ($\lambda_1 = 0.32$) is seen to separate the meadows receiving nature management from the remaining meadows and second axis ($\lambda_2 = 0.18$) separates the meadows managed by standard farming from those managed by hobby farming and biological farming, although the separations are not perfect. The species displayed on the right side of the diagram occur mainly in the meadows receiving nature management and those on the upper left in the meadows managed by standard farming, and so on. Figure 5.22 displays almost the same information as Figure 5.19a, as can be seen by joining Site Points 16 and 18 in both diagrams. Moisture and manuring are presumably the major factors bringing about vegetation differences between types of management.

5.5.6 Interpreting canonical axes

To interpret the ordination axes, one can use the canonical coefficients and the intra-set correlations. The canonical coefficients define the ordination axes as linear combinations of the environmental variables by means of Equation 5.14 and the intra-set correlations are the correlation coefficients between the environmental variables and these ordination axes. As before, we assume that the environmental variables have been standardized to a mean of 0 and a variance of 1 before the analysis. This standardization removes arbitrariness in the units of measurement of the environmental variables and makes the canonical coefficients comparable among each other, but does not influence other aspects of the analysis.

By looking at the signs and relative magnitudes of the intra-set correlations and of the canonical coefficients standardized in this way, we may infer the relative importance of each environmental variable for prediction of species composition. The canonical coefficients give the same information as the intra-set correlations, if the environmental variables are mutually uncorrelated, but may provide rather different information if the environmental variables are correlated among one another, as they usually are in field data. Both a canonical coefficient and an intra-set correlation relate to the rate of change in species composition by changing the corresponding environmental variable. However it is assumed that other environmental variables are being held constant in the former case, whereas the other environmental variables are assumed to covary with that one environmental variable in the particular way they do in the data set in the latter case. If the environmental variables are strongly correlated with one another, for example simply because the number of environmental variables approaches the number of sites, the effects of different environmental variables on the species composition cannot be singled out and, consequently, the canonical coefficients will be unstable. This is the multicollinearity problem discussed in the context of multiple regression in Subsection 3.5.3. The algorithms to obtain the canonical axes show that canonical coefficients are actually coefficients of a multiple regression (Subsection 5.5.2),

so both suffer identical problems. If the multicollinearity problem arises (the program CANOCO (ter Braak 1987b) provides statistics to help detecting it), one should abstain from attempts to interpret the canonical coefficients. But the intra-set correlations do not suffer from this problem. They are more like simple correlation coefficients. They can still be interpreted. One can also remove environmental variables from the analysis, keeping at least one variable per set of strongly correlated environmental variables. Then, the eigenvalues and species–environment correlations will usually only decrease slightly. If the eigenvalues and species–environment correlations drop considerably, one has removed too many (or the wrong) variables.

Their algorithms indicate that COR and CVA are hampered also by strong correlations among species, whereas CCA and RDA are not. So in CCA and RDA, the number of species is allowed to exceed the number of sites.

5.5.7 Data transformation

As in CA and PCA, any kind of transformation of the species abundances may affect the results of CCA and RDA. We refer to Subsections 5.2.2 and 5.3.5 for recommendations about this. The results of COR and CVA are affected by non-linear transformations of the species data, but not by linear transformations. Canonical ordination techniques are not influenced by linear transformations of the environmental variables, but non-linear transformation of environmental variables can be considered if there is some reason to do so. Prior knowledge about the possible impact of the environmental variables on species composition may suggest particular non-linear transformations and particular non-linear combinations, i.e. environmental scalars in the sense of Loucks (1962) and Austin et al. (1984). The use of environmental scalars can also circumvent the multicollinearity problem described in Subsection 5.5.6.

5.6 Multidimensional scaling

In Section 5.1, ordination was defined as a method that arranges site points in the best possible way in a continuum such that points that are close together correspond to sites that are similar in species composition, and points which are far apart correspond to sites that are dissimilar. A particular ordination technique is obtained by further specifying what 'similar' means and what 'best' is. The definition suggests that we choose a measure of (dis)similarity between sites (Subsection 6.2.2), replace the original species composition data by a matrix of dissimilarity values between sites and work further from the dissimilarity matrix to obtain an ordination diagram. This final step is termed multidimensional scaling.

In general, it is not possible to arrange sites such that the mutual distances between the sites in the ordination diagram are equal to the calculated dissimilarity values. Therefore a measure is needed that expresses in a single number how well or how badly the distances in the ordination diagram correspond to the dissimilarity values. Such a measure is termed a loss function or a stress function. In metric ordination techniques such as CA and PCA, the loss function depends

on the actual numerical values of the dissimilarities, whereas, in non-metric techniques, the loss function depends only on the rank order of the dissimilarities.

In CA and PCA, one need not calculate a matrix of dissimilarity values first, yet those techniques use particular measures of dissimilarity. In CA, the implied measure of dissimilarity is the chi-squared distance and, in PCA, the Euclidean Distance, as follows immediately from Subsection 5.3.3. The chi-square distance δ_{ij} between site i and site j is defined as

$$\delta_{ij}^2 = y_{++} \Sigma_{k=1}^{m} (y_{ki}/y_{+i} - y_{kj}/y_{+j})^2 / y_{k+} \qquad \text{Equation 5.15}$$

and the Euclidean Distance δ_{ij} between these sites is

$$\delta_{ij}^2 = \Sigma_{k=1}^{m} (y_{ki} - y_{kj})^2 \qquad \text{Equation 5.16}$$

The chi-squared distance involves proportional differences in abundances of species between sites, whereas the Euclidean Distance involves absolute differences. Differences in site and species totals are therefore less influential in CA than in PCA, unless a data transformation is used in PCA to correct for this effect.

A simple metric technique for multidimensional scaling is principal coordinate analysis (PCO), also called classical scaling (Gower 1966; Pielou 1977, p.290-395). PCO is based on PCA, but is more general than PCA, in that other measures of dissimilarity may be used than Euclidean Distance. In PCO, the dissimilarity values δ_{ij} are transformed into similarity values by the equation

$$c_{ij} = -0.5 \, (\delta_{ij}^2 - \delta_{i+}^2/n - \delta_{+j}^2/n + \delta_{++}^2/n^2) \qquad \text{Equation 5.17}$$

where the index + denotes a sum of squared dissimilarities. The matrix with elements c_{ij} is then subjected to the Q-mode algorithm of PCA (Subsection 5.3.6). If the original dissimilarities were computed as Euclidean Distances, PCO is identical to species-centred PCA calculated by the Q-mode algorithm.

In most techniques for (non-metric) multidimensional scaling, we must specify a priori the number of ordination axes and supply an initial ordination of sites. The technique then attempts to modify the ordination iteratively to minimize the stress. In contrast to the iterative algorithms for CA, PCA and PCO, different initial ordinations may lead to different results, because of local minima in the stress function (Subsection 5.2.7); hence, we must supply a 'good' initial ordination or try a series of initial ordinations. From such trials, we then select the ordination with minimum stress.

The best known technique for non-metric multidimensional scaling is ascribed to Shepard (1962) and Kruskal (1964). The stress function, which is minimized in their technique, is based on the Shepard diagram. This is a scatter diagram of the dissimilarities (δ_{ij}) calculated from the species data against the distances d_{ij} between the sites in the ordination diagram.

The ordination fits perfectly (stress = 0), if the dissimilarities are monotonic with the distances, i.e. if the points in the Shepard plot lie on a monotonically increasing curve. If they do not, we can fit a monotonic curve through the points

by least-squares. This is called monotonic or isotonic regression (Barlow et al. 1972). We then use as stress, a function of the residual sum of squares (for example, Kruskal's stress formula 1, which is the residual sum of squares divided by the total sum of squared distances). The algorithm to seek the ordination that minimizes the stress proceeds further as described above. Note that the method can work equally well with similarities, the only modifications being that a monotonically decreasing curve is fitted in the Shepard diagram. There are two methods to deal with equal dissimilarity values (ties). In the primary approach to ties, the corresponding fitted distances need not be equal, whereas they must be equal in the secondary approach. The primary approach to ties is recommended, because equal dissimilarity values do not necessarily imply equal habitat differences, in particular if the equalities arise between pairs of sites that have no species in common (Prentice 1977).

The Shepard–Kruskal method is based on the rank order of all dissimilarities. But calculated dissimilarities may not be comparable in different parts of a gradient, for example if there is a trend in species richness. This potential problem can be overcome by making a separate Shepard diagram for each site, in which we plot the dissimilarities and distances between the particular site and all remaining sites. Each Shepard diagram the distances leads to a stress value and the total stress is taken to be a combination of the separate stress values. This is the local non-metric technique proposed by Sibson (1972). Prentice (1977; 1980) advocated a particular similarity coefficient for use in Sibson's technique. This coefficient is

$$s_{ij} = \Sigma_k \min(y_{ki}, y_{kj}) \qquad \text{Equation 5.18}$$

Kendall (1971) proved that this coefficient contains all the information required to reconstruct the order of sites when abundances of species follow arbitrary unimodal response curves.

5.7 Evaluation of direct gradient and indirect gradient analysis techniques

Table 5.13 summarizes the techniques described in Chapters 3, 4 and 5 by type of response model and types of variables. We can classify response models as linear and non-linear. Each linear technique (from multiple regression to COR) has non-linear counterparts. A non-linear model that has special relevance in community ecology is the unimodal model. In principle, unimodal models can be fitted to data by the general methods used for non-linear models (in particular by maximum likelihood methods). For regression analysis, these methods are available (GLM, Chapter 3) but, in ordination, they are not so readily available and tend to require excessive computing. Therefore we have also introduced much simpler methods for analysing data for which unimodal models are appropriate. These simple methods start from the idea that the optima of species response curves can be estimated roughly by weighted averaging and we have shown (Section 3.7) that under particular conditions the estimates are actually quite good. This idea resulted in CA, DCA and CCA.

Multidimensional scaling is left out of Table 5.13, because it is unclear what

Table 5.13 Summary of gradient analysis techniques classified by type of response model and types of variables involved. MR, normal multiple regression; IR, inverse regression; PCA, principal components analysis; RDA, redundancy analysis; COR, canonical correlation analysis; WAE, weighted averaging of environmental values; GLM, generalized linear modelling; ML, maximum likelihood; WAI, weighted averaging of indicator values; CA, correspondence analysis; DCA, detrended correspondence analysis; CCA, canonical correspondence analysis; DCCA, detrended canonical correspondence analysis; env.vars, environmental variables; comp. gradients, composite gradients of environmental variables, either measured or theoretical.

	Response model		Number of variables		
	linear	unimodal	response (species)	explanatory (env. vars)	composite (comp.gradients)
Regression	MR	WAE, GLM, ML	one at a time	$\geqslant 1$*	one per species
Calibration	IR	WAI, ML	$\geqslant 1$*	rarely >1	none
Ordination	PCA	CA, DCA, ML	many	none	a few for all species
Canonical ordination	RDA	CCA, DCCA, ML	many	many*	a few for all species
	COR	variants of CCA,ML	many*	many*	a few

* less than number of sites, except for WAE, WAI and some applications of ML.

response models multidimensional scaling can cope with. Whether (non-metric) multidimensional scaling may detect a particular underlying data structure depends in an unknown way on the chosen dissimilarity coefficient and on the initial ordinations supplied. Non-metric multidimensional scaling could sometimes give better ordinations than DCA does, but the question is whether the improvements are worth the extra effort in computing power and manpower (Clymo 1980; Gauch et al. 1981).

Unimodal models are more general than monotonic ones (Figure 3.3), so it makes sense to start by using unimodal models and to decide afterwards whether one could simplify the model to a monotonic one. Statistical tests can help in this decision (Subsection 3.2.3). In ordination, we might therefore start by using CA, DCA or CCA. This initial analysis will provide a check on how unimodal the data are. If the lengths of the ordination axes are less than about 2 s.d., most of the response curves (or surfaces) will be monotonic, and we can consider using PCA or RDA. The advantage of using PCA and RDA is that in their biplot they provide more quantitative information than CA, DCA and (D)CCA in their joint plot, but this advantage would be outweighed by disadvantages when the data are strongly non-linear (ordination lengths greater than about 4 s.d.).

As illustrated by the Dune Meadow Data whose ordination lengths are about 3 s.d., DCA and PCA may result in similar configurations of site points (Figures 5.7 and 5.15). That they result in dissimilar configurations of species points, even if the ordination lengths are small, is simply due to the difference in meaning

of the species scores in DCA and PCA (Subsections 5.2.5 and 5.3.5).

Table 5.13 also shows the types of variables involved in regression, calibration, ordination and canonical ordination. We distinguish between response variables, explanatory variables and 'composite' variables, which in community ecology typically correspond to species presences or abundances, measured environmental variables and 'composite gradients', respectively. A composite gradient is either a linear combination of measured environmental variables or a theoretical variable. Which technique is the appropriate one to use largely depends on the research purpose and the type of data available. Ordination and cluster analysis (Chapter 6) are the only available techniques when one has no measured environmental data. Calibration must be considered if one wants to make inferences about values of a particular environmental variable from species data and existing knowledge of species–environment relations. Regression and canonical ordination are called for if one wants to build up and extend the knowledge of species–environment relations (Subsections 3.1.1 and 5.1.1).

Whether to use regression or to use canonical ordination depends on whether it is considered advantageous to analyse all species simultaneously or not. In a simultaneous analysis by canonical ordination, one implicitly assumes that all species are reacting to the same composite gradients of environmental variables according to a common response model. The assumption arises because canonical ordination constructs a few composite gradients for all species. By contrast in regression analysis, a separate composite gradient is constructed for each species. Regression may therefore result in more detailed descriptions and more accurate predictions of each particular species, at least if sufficient data are available. However ecological data that are collected over a large range of habitat variation require non-linear models; building good non-linear models by regression is not easy, because it requires construction of composite gradients that are non-linear combinations of environmental variables (Subsection 3.5.4). In CCA, the composite gradients are linear combinations of environmental variables, giving a much simpler analysis, and the non-linearity enters the model through a unimodal model for a few composite gradients, taken care of in CCA by weighted averaging. Canonical ordination is easier to apply and requires less data than regression. It provides a summary of the species–environment relations. The summary may lack the kind of detail that can in principle be provided by regression; on the other hand, the advantages of using regression, with its machinery of statistical tests, may be lost in practice, through the sheer complexity of non-linear model building and through lack of data. Because canonical ordination gives a more global picture than regression, it may be advantageous to apply canonical ordination in the early exploratory phase of the analysis of a particular data set and to apply regression in subsequent phases to selected species and environmental variables.

As already shown in the examples in Subsection 5.5.2, canonical ordination and ordination followed by environmental interpretation can be used fruitfully in combination. If the results do not differ much, then we know that no important environmental variables have been overlooked in the survey. But note that those included could merely be correlated with the functionally important ones. A further proviso is that the number of environmental variables (q) is small compared to

the number of sites (n). If this proviso is not met, the species–environment correlation may yield values close to 1, even if none of the environmental variables affects the species. (Note the remarks about R^2 in Subsection 3.2.1.) In particular, canonical ordination and ordination give identical ordination axes if $q \geqslant n - 1$. If the results of ordination and canonical ordination do differ much, then we may have overlooked major environmental variables, or important non-linear combinations of environmental variables already included in the analysis. But note that the results will also differ if CA or DCA detect a few sites on their first axis that have an aberrant species composition and if these sites are not aberrant in the measured environmental variables. After deleting the aberrant sites, the ordinations provided by (D)CA and CCA may be much more alike.

The question whether we have overlooked major environmental variables can also be studied by combining ordination and canonical ordination in a single analysis. Suppose we believe that two environmental variables govern the species composition in a region. We may then choose two ordination axes as linear combinations of these variables by canonical ordination, and extract further (unrestricted) axes as in CA or PCA, i.e. by the usual iteration process, making the axes unrelated to the previous (canonical) axes in each cycle. The eigenvalues of the extra axes measure residual variation, i.e. variation that cannot be explained by linear combinations of the environmental variables already included in the analysis. Such combined analyses are called partial ordination. Partial PCA (Subsection 5.3.5) is a special case of this.

A further extension of the analytical power of ordination is partial canonical ordination. Suppose the effects of particular environmental variables are to be singled out from 'background' variation imposed by other variables. In an environmental impact study, for example, the effects of impact variables are to be separated from those of other sources of variation, represented by 'covariables'. One may then want to eliminate ('partial out') the effects of the covariables and to relate the residual variation to the impact variables. This is achieved in partial canonical ordination. Technically, partial canonical ordination can be carried out by any computer program for canonical ordination. The usual environmental variables are simply replaced by the residuals obtained by regressing each of the impact variables on the covariables. The theory of partial RDA and partial CCA is described by Davies & Tso (1982) and ter Braak (1988). Partial ordination and partial canonical ordination are available in the computer program CANOCO (ter Braak 1987b). The program also includes a Monte Carlo permutation procedure to investigate the statistical significance of the effects of the impact variables.

5.8 Bibliographic notes

A simple ordination technique of the early days was polar ordination (Bray & Curtis 1957; Gauch 1982), which has been recently reappraised by Beals (1985). PCA was developed early this century by K. Pearson and H. Hotelling (e.g. Mardia et al. 1979) and was introduced in ecology by Goodall (1954). PCA was popularized by Orlóci (1966). CA has been invented independently since 1935 by several authors working with different types of data and with different rationales. Mathematically

CA is the same as reciprocal averaging, canonical analysis of contingency tables, and optimal or dual scaling of nominal variables (Gifi 1981; Gittins 1985; Greenacre 1984; Nishisato 1980). Benzécri et al. (1973) developed CA in a geometric context. Neither of these different approaches to CA is particularly attractive in ecology. Hill (1973) developed an ecological rationale (Subsection 5.2.2). The dispersion of the species scores by which we introduced CA in Subsection 5.2.1 is formally identical to the 'squared correlation ratio' (η^2) used by Torgerson (1958, Section 12.7) and Nishisato (1980, p.23) and also follows from the reciprocal gravity problem in Heiser (1986). RDA is also known under several names (Israëls 1984): PCA of instrumental variables (Rao 1964), PCA of y with respect to x, reduced rank regression (Davies & Tso 1982). Ter Braak (1986a) proposed CCA. COR was derived by H. Hotelling in 1935 (Gittins 1985). Campbell & Atchley (1981) provide a good geometric and algebraic introduction to CVA and Williams (1983) discusses its use in ecology. Methods to obtain the maximum likelihood solutions for Gaussian ordination have been investigated, under the assumption of a normal distribution, a Poisson distribution and a Bernoulli distribution for the species data, by Gauch et al. (1974), Kooijman (1977) and Goodall & Johnson (1982), respectively. However the computational burden of these methods and, hence, the lack of reliable computer programs have so far prevented their use on a routine basis. Ihm & van Groenewoud (1984) and ter Braak (1985) compared Gaussian ordination and CA. Non-metric multidimensional scaling started with the work by Shepard (1962) and Kruskal (1964). Schiffman et al. (1981) provide a clear introduction. They refer to local non-metric scaling as (row) conditional scaling. Meulman & Heiser (1984) describe a canonical form of non-metric multidimensional scaling. Early applications of non-metric multidimensional scaling in ecology were Anderson (1971), Noy-Meir (1974), Austin (1976), Fasham (1977), Clymo (1980) and Prentice (1977; 1980). The simple unfolding model (response models with circular contours) can in principle be fitted by methods of multidimensional scaling (Kruskal & Carroll 1969; Dale 1975; de Sarbo & Rao 1984; Heiser 1987), but Schiffman et al. (1981) warn of practical numerical problems that may reduce the usefulness of this approach. Most of the problems have, however, been circumvented by Heiser (1987).

Many textbooks use matrix algebra to introduce multivariate analysis techniques, because it provides an elegant and concise notation (Gordon 1981; Mardia et al. 1979; Greenacre 1984; Rao 1973; Gittins 1985). For ecologists, the book of Pielou (1984) is particularly recommended. All techniques described in Chapter 5 can be derived from the singular-value decomposition of a matrix, leading to singular vectors and singular values (Section 5.9). The decomposition can be achieved by many numerical methods (e.g. Gourlay & Watson 1973), one of which is the power algorithm (Table 5.6). The power algorithm is used in Chapter 5 because it provides the insight that ordination is simultaneously regression and calibration, and because it does not require advanced mathematics. The power algorithm can easily be programmed on a computer, but is one of the slowest algorithms available to obtain a singular-value decomposition. Hill (1979a) and ter Braak (1987b) use the power algorithm with a device to accelerate the process. The iteration processes of Tables 5.2 and 5.6 are examples of alternating least-

squares methods (Gifi 1981) and are related to the EM algorithm (Everitt 1984). The power algorithm is also a major ingredient of partial least squares (Wold 1982).

Computer programs for PCA, COR and CVA are available in most statistical computer packages. CA and DCA are available in DECORANA (Hill 1979). The program CANOCO (ter Braak 1987b) is an extension of DECORANA and it also includes PCA, PCO, RDA, CCA, CVA and partial variants of these techniques. All these techniques can be specified in terms of matrix algebra (Section 5.9). With the facilities for matrix algebra operations in GENSTAT (Alvey et al. 1977) or SAS (SAS Institute Inc. 1982), one can therefore write one's own programs to analyse small to medium-sized data sets. Schiffman et al. (1981) describe various programs for multidimensional scaling.

Chapter 5 uses response models as a conceptual basis for ordination. Carroll (1972) defined a hierarchy of response models, from the linear model (Equation 5.11), through the model with circular contour lines (Equation 5.5) to the full quadratic model (Equation 3.24) with ellipsoidal contours of varying orientation. He terms these models the vector model, the (simple) unfolding model and the general unfolding model, respectively (also Davison 1983). By taking even more flexible response models, we can define even more general ordination techniques. However the more flexible the model, the greater the computational problems (Prentice 1980). Future research must point out how flexible the model can be to obtain useful practical solutions.

5.9 Ordination methods in terms of matrix algebra

What follows in this section is a short introduction to ordination methods in terms of matrix algebra:

- to facilitate communication between ecologists and the mathematicians they may happen to consult
- to bridge the gap between the approach followed in Chapter 5 and the mainstream of statistical literature on multivariate methods
- to suggest computational methods based on algorithms for singular-value decomposition of a matrix or to extract eigenvalues and eigenvectors from a symmetric matrix.

To start, please read Section 5.8 first.

5.9.1 Principal components analysis (PCA)

Let $\mathbf{Y} = \{y_{ki}\}$ be an $m \times n$ matrix containing the data on m species (rows of the matrix) and n sites (columns of the matrix). In the most familiar form of PCA, species-centred PCA, the data are abundances with the species means already subtracted, so that $y_{k+} = 0$ as in Subsection 5.3.1. PCA is equivalent to the singular-value decomposition (SVD) of $\mathbf{Y}$ (e.g. Rao 1973; Mardia et al. 1979; Greenacre 1984)

$$\mathbf{Y} = \mathbf{P}\,\Lambda^{0.5}\,\mathbf{Q}' \qquad \text{Equation 5.19}$$

where $\mathbf{P}$ and $\mathbf{Q}$ are orthonormal matrices of dimensions $m \times r$ and $n \times r$, respectively, with $r = \min(m,n)$, i.e. $\mathbf{P'P} = \mathbf{I}$ and $\mathbf{Q'Q} = \mathbf{I}$, and Λ is a diagonal matrix with diagonal elements λ_s ($s = 1 \ldots r$), which are arranged in order of decreasing magnitude $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \ldots \geqslant 0$.

The columns of $\mathbf{P}$ and $\mathbf{Q}$ contain the singular vectors of $\mathbf{Y}$, and $\lambda_s^{0.5}$ is the sth singular value of $\mathbf{Y}$. If the sth column of $\mathbf{P}$ is denoted by $\mathbf{p}_s$, an m vector, and the sth column of $\mathbf{Q}$ by $\mathbf{q}_s$, an n vector, Equation 5.19 can be written as

$$\mathbf{Y} = \Sigma_{s=1}^{r} \lambda_s^{0.5} \mathbf{p}_s \mathbf{q}_s' \qquad \text{Equation 5.20}$$

The least-squares approximation of $\mathbf{Y}$ in Equation 5.11 of Subsection 5.3.2 is obtained from Equation 5.20 by retaining only the first two terms of this summation, and by setting $\mathbf{b}_s = \lambda_s^{0.5} \mathbf{p}_s$ and $\mathbf{x}_s = \mathbf{q}_s$ ($s = 1, 2$).
The kth element of $\mathbf{b}_1$ then contains the species score b_{k1}, and the ith element of $\mathbf{x}_1$ contains the site score x_{i1} on the first axis of PCA. Similarly, $\mathbf{b}_2$, and $\mathbf{x}_2$ contain the species and sites scores on the second axis of PCA. The species and sites scores on both axes form the coordinates of the points for species and sites in the biplot (Subsection 5.3.4). The interpretation of the PCA biplot follows from Equation 5.11: inner products between species points and site points provide a least-squares approximation of the elements of the matrix $\mathbf{Y}$ (Gabriel 1971; 1978). Equation 5.20 shows that the total sum of squares $\Sigma_{ki}\ y_{ki}^2$ equals $\lambda_1 + \ldots + \lambda_r$, the sum of all eigenvalues, and that the total residual sum of squares

$$\Sigma_{k,i} [y_{ki} - (b_{k1}\, x_{i1} + b_{k2}\, x_{i2})]^2 = \lambda_3 + \lambda_4 + \ldots + \lambda_r.$$

An appropriate measure of goodness of fit is therefore $(\lambda_1 + \lambda_2)/(\text{sum of all eigenvalues})$. From $\mathbf{P'P} = \mathbf{I}$, $\mathbf{Q'Q} = \mathbf{I}$ and Equation 5.20, we obtain

$$\mathbf{b}_s = \mathbf{Y}\mathbf{x}_s \qquad \text{Equation 5.21}$$

and

$$\lambda_s \mathbf{x}_s = \mathbf{Y'}\mathbf{b}_s. \qquad \text{Equation 5.22}$$

Hence, the species scores are a weighted sum of the site scores and the site scores are proportional to a weighted sum of the species scores (Table 5.6 and Subsection 5.3.2). Equation 5.21 and Equation 5.22 show that $\mathbf{b}_s$ and $\mathbf{x}_s$ are eigenvectors of $\mathbf{YY'}$ and $\mathbf{Y'Y}$, respectively, and that λ_s is their common eigenvalue; whence, the R-mode and Q-mode algorithms of Subsection 5.3.6.

The SVD of the species-by-species cross-product matrix $\mathbf{YY'}$ is $\mathbf{P}\ \Lambda\ \mathbf{P'}$, as follows from Equation 5.19 by noting that $\mathbf{Q'Q} = \mathbf{I}$. A least-squares approximation of the matrix $\mathbf{YY'}$ in two dimensions is therefore given by the matrix $\mathbf{b}_1\mathbf{b}'_1 + \mathbf{b}_2\,\mathbf{b}'_2$. Since $\mathbf{YY'}/(n-1)$ contains covariances between species, the biplot of $\mathbf{x}_s$ and $\mathbf{b}_s$ is termed the covariance biplot (Corsten & Gabriel 1976; ter Braak 1983).

The SVD of the site-by-site cross-product matrix $\mathbf{Y}'\mathbf{Y}$ is $\mathbf{Q}\ \Lambda\ \mathbf{Q}'$. A biplot of $\mathbf{Y}$ and $\mathbf{Y}'\mathbf{Y}$ is therefore obtained by redefining $\mathbf{b}_s$ and $\mathbf{x}_s$ as $\mathbf{b}_s = \mathbf{p}_s$ and $\mathbf{x}_s = \lambda_s^{0.5}\mathbf{q}_s$. The inter-site distances in this biplot approximate the Euclidean Distances between sites as defined by Equation 5.16; hence the name Euclidean Distance biplot. The approximation is, however, indirect namely through Equation 5.17 with c_{ij} the (i, j)th element of $\mathbf{Y}'\mathbf{Y}$. A consequence of this is that the inter-site distances are always smaller than the Euclidean Distances.

5.9.2 Correspondence analysis (CA)

In CA, the species-by-sites matrix $\mathbf{Y}$ contains the abundance values y_{ki} in which $y_{ki} \geqslant 0$. The data is not previously centred in CA. Let $\mathbf{M} = \text{diag}\ (y_{k+})$, an $m \times m$ diagonal matrix containing the row totals of $\mathbf{Y}$, $\mathbf{N} = \text{diag}\ (y_{+i})$, an $n \times n$ diagonal matrix containing the column totals of $\mathbf{Y}$.

As stated in Subsection 5.2.1, CA chooses standardized site scores $\mathbf{x}$ that maximize the dispersion of species scores, which are themselves weighted averages of the site scores (Equation 5.1). In matrix notation, the vector of species scores $\mathbf{u} = (u_k)[k = 1, ..., m]$ is

$$\mathbf{u} = \mathbf{M}^{-1}\mathbf{Y}\mathbf{x} \qquad \text{Equation 5.23}$$

and the dispersion is

$$\delta = \mathbf{u}'\mathbf{M}\mathbf{u}/\mathbf{x}'\mathbf{N}\mathbf{x} = \mathbf{x}'\mathbf{Y}'\mathbf{M}^{-1}\mathbf{Y}\mathbf{x}\ /\ \mathbf{x}'\mathbf{N}\mathbf{x} \qquad \text{Equation 5.24}$$

where the denominator takes account of the standardization of $\mathbf{x}$ (Table 5.2c), provided $\mathbf{x}$ is centred ($\mathbf{1}'\mathbf{N}\mathbf{x} = 0$).

The problem of maximizing δ with respect to $\mathbf{x}$ has as solution the second eigenvector of the eigenvalue equation

$$\mathbf{Y}'\mathbf{M}^{-1}\mathbf{Y}\mathbf{x} = \lambda\ \mathbf{N}\ \mathbf{x} \qquad \text{Equation 5.25}$$

with $\delta = \lambda$ (Rao 1973, Section 1f.2 and p.74; Mardia et al. 1979, Theorem A9.2).

This can be seen by noting that the first eigenvector is a trivial solution ($\mathbf{x} = \mathbf{1}$; $\lambda = 1$); because the second eigenvector is orthogonal to the first eigenvector in the $\mathbf{N}$ metric, the second eigenvector maximizes δ subject to $\mathbf{1}'\mathbf{N}\mathbf{x} = 0$. What is called the first eigenvector of CA in Section 5.2 is thus the second eigenvector of Equation 5.25, i.e. its first non-trivial eigenvector. The second non-trivial eigenvector of Equation 5.25 is similarly seen to maximize δ, subject to being centred and to being orthogonal to the first non-trivial eigenvector, and so on for subsequent axes. Equation 5.25 can be rewritten as

$$\lambda\ \mathbf{x} = \mathbf{N}^{-1}\mathbf{Y}'\mathbf{u} \qquad \text{Equation 5.26}$$

Equations 5.23 and 5.26 form the 'transition equations' of CA. In words: the species scores are weighted averages of the site scores and the site scores are proportional to weighted averages of the species scores (Table 5.2 and Exercise 5.1.3).

The eigenvectors of CA can also be obtained from the SVD

$$\mathbf{M}^{-0.5}\,\mathbf{Y}\,\mathbf{N}^{-0.5} = \mathbf{P}\,\Lambda^{0.5}\mathbf{Q}' \qquad \text{Equation 5.27}$$

by setting $\mathbf{u}_s = \lambda_s^{0.5}\,\mathbf{M}^{-0.5}\,\mathbf{p}_s$ and $\mathbf{x}_s = \mathbf{N}^{-0.5}\,\mathbf{q}_s$, where $\mathbf{p}_s$ and $\mathbf{q}_s$ are the sth columns of $\mathbf{P}$ and $\mathbf{Q}$, respectively ($s = 1, \ldots, r$).

This can be seen by inserting the equations for $\mathbf{u}_s$ and $\mathbf{x}_s$ in Equations 5.23 and 5.26, and rearranging terms. It is argued in Subsection 5.2.7 that it is equally valid to distribute λ_s in other ways among $\mathbf{u}_s$ and $\mathbf{x}_s$, as is done, for example, in Hill's scaling (Subsection 5.2.2).

CA differs from PCA in the particular transformation of $\mathbf{Y}$ in Equation 5.27 and in the particular transformation of the singular vectors described just below that equation.

5.9.3 Canonical correlation analysis (COR)

As in species-centred PCA, let $\mathbf{Y}$ be an $m \times n$ matrix in which the kth row contains the centred abundance values of the kth species (i.e. $y_{k+} = 0$) and let $\mathbf{Z}$ be a $q \times n$ matrix in which the jth row contains the centred values of the jth environmental variable (i.e. $z_{j+} = 0$). Define

$$\mathbf{S}_{12} = \mathbf{YZ}',\ \mathbf{S}_{11} = \mathbf{YY}',\ \mathbf{S}_{22} = \mathbf{ZZ}' \text{ and } \mathbf{S}_{21} = \mathbf{S}'_{12}. \qquad \text{Equation 5.28}$$

The problem of COR is to determine coefficients for the species $\mathbf{b} = (b_k)[k = 1, \ldots, m]$ and for the environmental variables $\mathbf{c} = (c_j)[j = 1, \ldots, q]$ that maximize the correlation between $\mathbf{x}^* = \mathbf{Y}'\mathbf{b}$ and $\mathbf{x} = \mathbf{Z}'\mathbf{c}$. The solution for $\mathbf{b}$ and $\mathbf{c}$ is known to be the first eigenvector of the respective eigenvalue equations

$$\mathbf{S}_{12}\,\mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}\,\mathbf{b} = \lambda\,\mathbf{S}_{11}\,\mathbf{b} \qquad \text{Equation 5.29}$$

$$\mathbf{S}_{21}\,\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{c} = \lambda\,\mathbf{S}_{22}\,\mathbf{c} \qquad \text{Equation 5.30}$$

The eigenvalue λ equals the squared canonical correlation (Rao 1973; Mardia et al. 1979; Gittins 1985).

Note that $\mathbf{b}$ can be derived from a multiple regression of $\mathbf{x}$ on the species, or from $\mathbf{c}$, by

$$\mathbf{b} = (\mathbf{YY}')^{-1}\,\mathbf{Yx} = \mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{c} \qquad \text{Equation 5.31}$$

and, similarly, $\mathbf{c}$ can be derived from a multiple regression of $\mathbf{x}^*$ on the environmental variables, or from $\mathbf{b}$, by

$$\lambda \mathbf{c} = (\mathbf{Z}\mathbf{Z}')^{-1}\,\mathbf{Z}\mathbf{x}^* = \mathbf{S}_{22}^{-1}\,\mathbf{S}_{21}\,\mathbf{b} \qquad \text{Equation 5.32}$$

It can be verified that **b** and **c** from Equations 5.31 and 5.32 satisfy Equations 5.29 and 5.30, by inserting **b** from Equation 5.31 into Equation 5.32 and by inserting **c** from Equation 5.32 into Equation 5.31; but note that we could have distributed λ in other ways among Equations 5.31 and 5.32. Equations 5.31 and 5.32 form the basis of the iteration algorithm of Table 5.12. Step 7 of Table 5.12 takes care of the eigenvalue: at convergence, **x** is divided by λ (Table 5.6c). Once convergence is attained, **c** should be divided by λ to ensure that the final site scores satisfy $\mathbf{x} = \mathbf{Z}'\mathbf{c}$ (Step 5);hence the λ in Equation 5.32. The second and further axes obtained by Table 5.12 also maximize the correlation between **x** and $\mathbf{x}^*$, but subject to being uncorrelated to the site scores of the axes already extracted.

COR can also be derived from the SVD of

$$\mathbf{S}_{11}^{-0.5}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-0.5} = \mathbf{P}\,\Lambda^{0.5}\,\mathbf{Q}' \qquad \text{Equation 5.33}$$

The equivalence of Equation 5.31 with Equation 5.33 can be verified by pre-multiplying Equation 5.33 on both sides with $\mathbf{S}_{11}^{-0.5}$ and post-multiplying Equation 5.33 on both sides with **Q** and by defining

$$\mathbf{B} = \mathbf{S}_{11}^{-0.5}\,\mathbf{P}\,\Lambda^{0.5} \text{ and } \mathbf{C} = \mathbf{S}_{22}^{-0.5}\,\mathbf{Q}. \qquad \text{Equation 5.34}$$

The sth column of **B** and of **C** contain the canonical coefficients on the sth axis of the species and environmental variables, respectively. The equivalence of Equation 5.32 with Equation 5.33 can be shown similarly.

COR allows a biplot to be made in which the correlations between species and environmental variables are approximated. The problem to which the canonical correlation biplot is the solution can be formulated as follows: determine points for species and environmental variables in t-dimensional space in such a way that their inner products give a weighted least-squares approximation to the elements of the covariance matrix $\mathbf{S}_{12}$. In the approximation, the species and the environmental variables are weighted inversely with their covariance matrices $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$, respectively. Let the coordinates of the points for the species be collected in the $m \times t$ matrix **G** and those for the environmental variables in the $q \times t$ matrix **H**. The problem is then to minimize

$$\|\mathbf{S}_{11}^{-0.5}\,(\mathbf{S}_{12} - \mathbf{G}\mathbf{H}')\,\mathbf{S}_{22}^{-0.5}\|^2 = \|\mathbf{S}_{11}^{-0.5}\,\mathbf{S}_{12}\,\mathbf{S}_{22}^{-0.5} - (\mathbf{S}_{11}^{-0.5}\,\mathbf{G})(\mathbf{S}_{22}^{-0.5}\,\mathbf{H})'\|^2 \qquad \text{Equation 5.35}$$

with respect to the matrices **G** and **H**, where $\|\bullet\|$ is the Euclidean matrix norm, e.g. $\|\mathbf{Y}\|^2 = \Sigma_{k,i}\, y_{ki}^2$.

From the properties of an SVD (Subsection 5.9.1), it follows that the minimum is attained when $\mathbf{S}_{11}^{-0.5}\,\mathbf{G}$ and $\mathbf{S}_{22}^{-0.5}\,\mathbf{H}$ correspond to the first t columns of the matrices $\mathbf{P}\Lambda^{0.5}$ and **Q** of Equation 5.33, respectively. The required least-squares

approximation is thus obtained by setting **G** and **H** equal to the first t columns of $\mathbf{S}_{11}^{0.5}\,\mathbf{P}\Lambda^{0.5}$ and $\mathbf{S}_{22}^{0.5}\,\mathbf{Q}$, respectively. Again, Λ can be distributed in other ways among **P** and **Q**. For computational purposes, note that

$$\mathbf{S}_{11}^{0.5}\,\mathbf{P}\,\Lambda^{0.5} = \mathbf{S}_{11}\,\mathbf{S}_{11}^{-0.5}\,\mathbf{P}\,\Lambda^{0.5} = \mathbf{S}_{11}\,\mathbf{B} = \mathbf{Y}\mathbf{Y}'\mathbf{B} = \mathbf{Y}\mathbf{X} \qquad \text{Equation 5.36}$$

and

$$\mathbf{S}_{22}^{0.5}\,\mathbf{Q} = \mathbf{S}_{22}\,\mathbf{S}_{22}^{-0.5}\,\mathbf{Q} = \mathbf{S}_{22}\,\mathbf{C} = \mathbf{Z}\mathbf{X} \qquad \text{Equation 5.37}$$

where $\mathbf{X} = \mathbf{Z}'\mathbf{C}$. Because $\mathbf{X}'\mathbf{X} = \mathbf{I}$, the biplot can thus be constructed from the inter-set correlations of the species and the intra-set correlations of the environmental variables (which are the correlations of the site scores **x** with the species variables and environmental variables, respectively). This construction rule requires the assumption that the species and environmental variables are standardized to unit variance, so that $\mathbf{S}_{12}$ is actually a correlation matrix. The angles between arrows in the biplot are, however, not affected by whether either covariances or correlations between species and environment are approximated in the canonical correlation biplot.

5.9.4 Redundancy analysis (RDA)

RDA is obtained by redefining $\mathbf{S}_{11}$ in subsection 5.9.3 to be the identity matrix (Rao 1973, p.594-595). In the RDA biplot, as described in Subsection 5.5.3, the coordinates of the point for the species and the variables are given in the matrices **P** and $\mathbf{S}_{22}^{0.5}\,\mathbf{Q}\,\Lambda^{0.5}$, respectively.

5.9.5 Canonical correspondence analysis (CCA)

CCA maximizes Equation 5.24 subject to Equation 5.14, provided **x** is centred. If the matrix **Z** is extended with a row of ones, Equation 5.14 becomes $\mathbf{x} = \mathbf{Z}'\mathbf{c}$, with $\mathbf{c} = (c_j)[j = 0, 1, \ldots, q]$. By inserting $\mathbf{x} = \mathbf{Z}'\mathbf{c}$ in Equation 5.24 and (re)defining, with **Y** non-centred,

$$\mathbf{S}_{12} = \mathbf{Y}\mathbf{Z}',\ \mathbf{S}_{11} = \mathbf{M} = \text{diag}\,(y_{k+}) \text{ and } \mathbf{S}_{22} = \mathbf{Z}\mathbf{N}\mathbf{Z}' \qquad \text{Equation 5.38}$$

we obtain

$$\delta = \mathbf{c}'\mathbf{S}_{21}\,\mathbf{S}_{11}^{-1}\,\mathbf{S}_{12}\,\mathbf{c} \,/\, \mathbf{c}'\,\mathbf{S}_{22}\,\mathbf{c} \qquad \text{Equation 5.39}$$

The solutions of CCA can therefore be derived from the eigenvalue Equation 5.30 with $\mathbf{S}_{12}$, $\mathbf{S}_{11}$ and $\mathbf{S}_{22}$ defined as in Equation 5.38. If defined in this way, CCA has a trivial solution $\mathbf{c}' = (1, 0, 0, \ldots, 0)$, $\lambda = 1$, $\mathbf{x} = \mathbf{1}$ and the first non-trivial eigenvector maximizes δ subject to $\mathbf{1}'\mathbf{N}\mathbf{x} = \mathbf{1}'\mathbf{N}\mathbf{Z}'\mathbf{c} = 0$ and the maximum δ equals the eigenvalue. A convenient way to exclude the trivial solution is to subtract from each environmental variable its weighted mean $\bar{z}_j = \Sigma_i\, y_{+i}\, z_{ji} / y_{++}$

(and to remove the added row of ones in the matrix **Z**). Then, the matrix **Z** has weighted row means equal to 0: $\Sigma_i\ y_{+i}\ z_{ji} = 0$. The species scores and the canonical coefficients of the environmental variables can be obtained from Equation 5.33 and Equation 5.34, by using the definitions of Equation 5.38.

As described in Subsection 5.5.2, the solution of CCA can also be obtained by extending the iteration algorithm of Table 5.2. Steps 1, 4, 5 and 6 remain the same as in Table 5.2. In matrix notation, the other steps are

Step 2 $\mathbf{b} = \mathbf{M}^{-1}\,\mathbf{Y}\,\mathbf{x}$ Equation 5.40

Step 3a $\mathbf{x}^* = \mathbf{N}^{-1}\,\mathbf{Y}'\mathbf{b}$ Equation 5.41

Step 3b $\mathbf{c} = (\mathbf{ZNZ}')^{-1}\,\mathbf{ZN}\,\mathbf{x}^*$ Equation 5.42

Step 3c $\mathbf{x} = \mathbf{Z}'\mathbf{c}$ Equation 5.43

with $\mathbf{b} = \mathbf{u}$, the m vector containing the species scores u_k ($k = 1, \ldots, m$).

Once convergence has been attained, to ensure that the final site scores satisfy $\mathbf{x} = \mathbf{Z}'\mathbf{c}$, $\mathbf{c}$ should be divided by λ, as in COR (below Equation 5.32). This amounts to replacing $\mathbf{c}$ in Equation 5.42 by $\lambda\mathbf{c}$ (as in Equation 5.32). To show that the algorithm gives a solution of Equation 5.30, we start with Equation 5.42, modified in this way, insert $\mathbf{x}^*$ of Equation 5.41 in Equation 5.42, next insert $\mathbf{b}$ by using Equation 5.40, next insert $\mathbf{x}$ by using Equation 5.43 and finally use the definitions of $\mathbf{S}_{11}$, $\mathbf{S}_{12}$ and $\mathbf{S}_{22}$ in CCA.

CCA allows a biplot to be made, in which the inner products between points for species and points for environmental variables give a weighted least-squares approximation of the elements of the $m \times q$ matrix

$$\mathbf{W} = \mathbf{M}^{-1}\mathbf{YZ}',$$

the (k,j)th element of which is the weighted average of species k with respect to the (centred) environmental variable j. In the approximation, the species are given weight proportional to their total abundance (y_{k+}) and the environmental variables are weighted inversely with their covariance matrix $\mathbf{S}_{22}$. The possibility for such a biplot arises because

$$\mathbf{M}^{0.5}\,\mathbf{W}\,\mathbf{S}_{22}{}^{-0.5} = \mathbf{S}_{11}{}^{-0.5}\,\mathbf{S}_{12}\,\mathbf{S}_{22}{}^{-0.5} \qquad \text{Equation 5.44}$$

so that, from Equations 5.44 and 5.33, after rearranging terms,

$$\mathbf{W} = (\mathbf{S}_{11}{}^{-0.5}\,\mathbf{P})\,\Lambda^{0.5}\,(\mathbf{S}_{22}{}^{0.5}\,\mathbf{Q})' \qquad \text{Equation 5.45}$$

Apart from particular considerations of scale (Subsection 5.2.2), the coordinates of the points for species and environmental variables in the CCA biplot are thus

given by the first t columns of $\mathbf{S}_{11}^{-0.5}\,\mathbf{P}\,\Lambda^{0.5}$ and $\mathbf{S}_{22}^{0.5}\,\mathbf{Q}$, respectively. The matrix $\mathbf{S}_{11}^{-0.5}\,\mathbf{P}\,\Lambda^{0.5}$ actually contains the species scores, as follows from Equation 5.34. The other matrix required for the biplot can be obtained by

$$\mathbf{S}_{22}^{0.5}\mathbf{Q} = \mathbf{S}_{22}\,\mathbf{S}_{22}^{-0.5}\,\mathbf{Q} = \mathbf{S}_{22}\,\mathbf{C} = \mathbf{ZNZ'C} = \mathbf{ZNX} \qquad \text{Equation 5.46}$$

5.10 Exercises

Exercise 5.1 Correspondence analysis: the algorithm

This exercise illustrates the two-way weighted-averaging algorithm of CA (Table 5.2) with the small table of artificial data given below.

Species	Sites				
	1	2	3	4	5
A	1	0	0	1	0
B	0	0	1	0	1
C	0	2	0	1	0
D	3	0	0	1	1

The data appear rather chaotic now, but they will show a clear structure after having extracted the first CA ordination axis. The first axis is dealt with in Exercises 5.1.1-3, and the second axis in Exercises 5.1.4-6.

Exercise 5.1.1 Take as site scores the values 1, 2, ..., 5 as shown above the data table. Now, standardize the site scores by using the standardization procedure described in Table 5.2c.

Exercise 5.1.2 Use the site scores so standardized as initial site scores in the iteration process (Table 5.2a). Carry out at least five iteration cycles and in each cycle calculate the dispersion of the species scores. (Use an accuracy of three decimal places in the calculations for the site and species scores and of four decimal places for s.) Note that the scores keep changing from iteration to iteration, but that the rank order of the site scores and of the species scores remains the same from Iteration 4 onwards. Rearrange the species and sites of the table according to their rank order. Note also that the dispersion increases during the iterations.

Exercise 5.1.3 After 19 iterations, the site scores obtained are 0.101, -1.527, 1.998, -0.524, 1.113. Verify these scores for the first CA axis (within an accuracy of two decimal places) by carrying out one extra iteration cycle. What is the eigenvalue of this axis? Verify that Equation 5.1 holds true for the species scores and site scores finally obtained, but that Equation 5.2 does not hold true. Modify Equation 5.2 so that it does hold true.

Exercise 5.1.4 We now derive the second CA axis by using the same initial site scores as in Exercise 5.1.2. Orthogonalize these scores first with respect to the first axis by using the orthogonalization procedure described in Table 5.2b, and next standardize them (round the site scores of the first axis to two decimals and use four decimals for v and s and three for the new scores).

Exercise 5.1.5 Use the site scores so obtained as initial site scores to derive the second axis. The scores stabilize in four iterations (within an accuracy of two decimal places).

Exercise 5.1.6 Construct an ordination diagram of the first two CA axes. The diagram shows one of the major 'faults' in CA. What is this fault?

Exercise 5.2 Adding extra sites and species to a CA ordination

Exercise 5.2.1 We may want to add extra species to an existing CA ordination. In the Dune Meadow Data, *Hippophae rhamnoides* is such a species, occurring at Sites 9, 18 and 19 with abundances 1, 2 and 1, respectively. Calculate from the site scores in Table 5.1c the score for this species on the first CA axis in the way this is done with CA. Plot the abundance of the species against the site score. What does the species score mean in this plot? At which place does the species appear in Table 5.1c? Answer the same questions for *Poa annua*, which occurs at Sites 1, 2, 3, 4, 7, 9, 10, 11, 13 and 18 with abundances 3, 3, 6, 4, 2, 2, 3, 2, 3 and 4, respectively, and for *Ranunculus acris*, which occurs at Sites 5, 6, 7, 9, 14 and 15 with abundances 2, 3, 2, 2, 1 and 1, respectively.

Exercise 5.2.2 Similarly, we may want to add an extra site to an existing CA ordination. Calculate the score of the site where the species *Bellis perennis*, *Poa pratensis* and *Rumex acetosa* are present with abundances 5, 4 and 3, respectively (imaginary data). (Hint: recall how the site scores were obtained from the species scores in Exercise 5.1.3.) Species and sites so added to an ordination are called passive, to distinguish them from the active species and sites of Table 5.1. The scores on higher-order axes are obtained in the same way.

Exercise 5.2.3 Rescale the scores of Table 5.1c to Hill's scaling and verify that the resulting scores were used in Figure 5.4.

Exercise 5.3 Principal components analysis

Add the extra species and the extra site of Exercise 5.2 to the PCA ordination of Table 5.5c. Plot the abundance of the extra species against the site scores. What does the species score mean in this plot? At which places do the species appear in Table 5.5c?

Exercise 5.4 Length of gradient in DCA

Suppose DCA is applied to a table of abundances of species at sites and that the length of the first axis is 1.5 s.d. If, for each species, we made a plot of its abundance against the site scores of the first axis, would most plots suggest monotonic curves or unimodal curves? And what would the plots suggest if the length of the axis was 10 s.d.?

Exercise 5.5 Interpretation of joint plot and biplot

Exercise 5.5.1 Rank the sites in order of abundance of *Juncus bufonius* as inferred from Figure 5.7, as inferred from Figure 5.15 and as observed in Table 5.1a. Do the same for *Eleocharis palustris*.

Exercise 5.5.2 If Figure 5.15 is interpreted erroneously as a joint plot of DCA, one gets different inferred rank orders and, when Figure 5.7 is interpreted erroneously as a biplot, one also gets different rank orders. Is the difference in interpretation greatest for species that lie near the centre of an ordination diagram or for species that lie on the edge of an ordination diagram?

Exercise 5.6 Detrended canonical correspondence analysis

Cramer (1986) studied vegetational succession on the rising sea-shore of an island in the Stockholm Archipelago. In 1978 and 1984, the field layer was sampled on 135 plots of 1 m^2 along 4 transects. The transects ran from water level into mature forest. One of the questions was whether the vegetational succession keeps track with the land uplift (about 0.5 cm per year) or whether it lags behind. In both cases, the vegetation zones 'run down the shore', but in the latter case too slowly. Because succession in the forest plots was not expected to be due to land uplift, only the 63 plots up to the forest edge were used. These plots contained 68 species with a total of about 1000 occurrences on the two sampling occasions. An attempt was made to answer the question by using detrended canonical correspondence analysis (DCCA) with two explanatory variables, namely altitude above water level in 1984 (not corrected for land uplift; so each plot received the same value in 1978 as in 1984) and time (0 for 1978, 6 for 1984). The altitude ranged from -14 to 56 cm. The first two axes gave eigenvalues 0.56 and 0.10, lengths 4.4 and 0.9 s.d. and species–environment correlations 0.95 and 0.74, respectively. Table 5.14 shows that the first axis is strongly correlated with altitude and almost uncorrelated with time, whereas the second axis is strongly correlated with time and almost uncorrelated with altitude. However the canonical coefficients tell a more interesting story.

Exercise 5.6.1 With Table 5.14, show that the linear combination of altitude and time best separating the species in the sense of Section 5.5.2 is

$$x = 0.054\, z_1 + 0.041\, z_2 \qquad \text{Equation 5.47}$$

Table 5.14 Detrended canonical correspondence analysis of rising shore vegetation data: canonical coefficients ($100 \times c$) and intra-set correlations ($100 \times r$) for standardized environmental variables. In brackets, the approximate standard errors of the canonical coefficients. Also given are the mean and standard deviation (s.d.) of the variables.

Variable	Coefficients		Correlations		mean	s.d.
	Axis 1	Axis 2	Axis 1	Axis 2		
Altitude (cm)	100 (3)	4 (4)	99	19	22	18.5
Time (years)	12 (3)	-34 (3)	7	-99	3	2.9

where z_1 is the numeric value of altitude (cm) and z_2 is the numeric value of time (years) and where the intercept is, arbitrary, set to zero.

Hint: note that Table 5.14 shows standardized canonical coefficients, i.e. canonical coefficients corresponding to the standardized variables $z_1^* = (z_1 - 22)/18.5$ and $z_2^* = (z_2 - 3)/2.9$.
Similarly, show that the standard errors of estimate of $c_1 = 0.054$ is 0.0016 and of $c_2 = 0.041$ is 0.010.

Exercise 5.6.2 Each value of x in Equation 5.47 stands for a particular species composition (Figures 5.8 and 5.18) and changes in the value of x express species turnover along the altitude gradient in multiples of s.d. With Equation 5.47, calculate the species turnover between two plots that were 15 cm and 25 cm above water level in 1984, respectively. Does the answer depend on the particular altitudes of these plots or only on the difference in altitude? What is, according to Equation 5.47, the species turnover between these plots in 1978?

Exercise 5.6.3 With Equation 5.47, calculate the species turnover between 1978 and 1984 for a plot with an altitude of 15 cm in 1984? Does the answer depend on altitude?

Exercise 5.6.4 With Equation 5.47, calculate the altitude that gives the same species turnover as one year of succession.

Exercise 5.6.5 Is there evidence that the vegetational succession lags behind uplift?

Exercise 5.6.6 Roughly how long would it take to turn the species composition of the plot closest to the sea into that of the plot that is on the edge of the forest? Hint: use the length of the first axis. Is there evidence from the analysis that there might also be changes in species composition that are unrelated to land uplift? Hint: consider the length of the second axis.

5.11 Solutions to exercises

Exercise 5.1 Correspondence analysis: the algorithm

Exercise 5.1.1 The centroid of the site scores is $z = (4 \times 1 + 2 \times 2 + 1 \times 3 + 3 \times 4 + 2 \times 5)/12 = 2.750$ and their dispersion is $s^2 = [4 \times (1 - 2.750)^2 + ... + 2 \times (5 - 2.750)^2]/12 = 2.353$, thus $s = 1.5343$. The standardized initial score for the first site is thus $x_1 = (1 - 2.750)/1.5343 = -1.141$. The other scores are listed on the second line of Table 5.15.

Exercise 5.1.2 In the first iteration cycle at Step 2, we obtain for Species C, for example, the score $[2 \times (-0.489) + 1 \times 0.815)]/(2 + 1) = -0.054$, and for Site 5 at Step 3, the score $(0.815 - 0.228)/(1 + 1) = 0.294$. The dispersion of the species scores in the first iteration cycle is $\delta = (2 \times 0.163^2 + 2 \times 0.815^2 + 3 \times 0.054^2 + 5 \times 0.228^2)/12 = 0.138$. See further Table 5.15. In the iterations shown $z = 0.000$, apart from Iteration 3, where $z = -0.001$ (Step 5). The rearranged data table shows a Petrie matrix (Subsection 5.2.3).

Exercise 5.1.3 The standardized site scores obtained in the 19th and 20th iteration are equal within the accuracy of two decimal places; so the iteration process has converged (Table 5.15). The eigenvalue of the first axis is $\lambda_1 = 0.7799$, the value of s calculated last. Equation 5.2 does not hold true for the final site and species scores. But the site scores calculated in Step 3 are weighted averages of the species scores and are divided in the 20th iteration by $s = 0.7799$ to obtain the final site scores. On convergence, s equals the eigenvalue λ; thus the final site and species scores satisfy the relation $\lambda\ x_i = \Sigma_{k=1}^{m}\ y_{ki}\ u_k / \Sigma_{k=1}^{m} y_{ki}$. Applying Steps 3, (4) and 5 to the eigenvector (the scores x_i) thus transforms the eigenvector into a multiple of itself. The multiple is the 'eigenvalue' of the eigenvector. Note that δ equals λ within arithmetic accuracy.

Exercise 5.1.4 In Step 4.2, we obtain $v = [4 \times (-1.141) \times 0.10 + 2 \times (-0.489) \times (-0.53) + 1 \times 0.163 \times 2.00 + 3 \times 0.815 \times (-0.53) + 2 \times 1.466 \times 1.11]/12 = 0.2771$ and for Site 1 at Step 4.3, the score $-1.141\ -0.277\ 1 \times 0.10 = -1.169$. See further the first four lines of Table 5.16.

Exercise 5.1.5 See Table 5.16.

Exercise 5.1.6 The configuration of the site points looks like the letter V, with Site 1 at the bottom and Sites 2 and 3 at the two extremities. This is the arch effect of CA (Section 5.2.3).

Exercise 5.2 Adding extra sites and species to a CA ordination

Exercise 5.2.1 In CA, Equation 5.1 is used to obtain species scores from site scores. Thus the score for *Hippophae rhamnoides* is $[1 \times 0.09 + 2 \times (-0.31)$

Table 5.15 Two-way weighted averaging algorithm applied to the data of Exercise 5.1 to obtain the first ordination axis of CA. The initial site scores (Line 1) are first standardized (Line 2). The values in brackets are rank numbers of the scores of the line above. Column 1, iteration number; Column 2, step number in Table 5.2 ; Column 3, x is site score and u is species score; Column 4, dispersion of the species scores (δ) when preceded by u, or otherwise the square root of the dispersion of the site scores of the line above (s).

Column				Sites					Species			
1	2	3	4	1	2	3	4	5	A	B	C	D
0	1	x		1.000	2.000	3.000	4.000	5.000				
0	5	x	1.5343	-1.141	-0.489	0.163	0.815	1.466				
1	2	u	0.1375						-0.163	0.815	-0.054	-0.228
1	3	x		-0.212	-0.054	0.815	-0.148	0.294	(2)	(4)	(3)	(1)
1	5	x	0.3012	-0.704	-0.179	2.706	-0.491	0.976				
2	2	u	0.6885						-0.598	1.841	-0.283	-0.325
2	3	x		-0.393	-0.283	1.841	-0.402	0.758	(1)	(4)	(3)	(2)
2	5	x	0.6953	-0.567	-0.408	2.646	-0.580	1.089				
3	2	u	0.7171						-0.574	1.868	-0.465	-0.238
3	3	x		-0.322	-0.465	1.868	-0.426	0.815	(1)	(4)	(2)	(3)
3	5	x	0.7193	-0.448	-0.646	2.597	-0.592	1.133				
4	2	u	0.7342						-0.520	1.865	-0.628	-0.161
4	3	x		-0.251	-0.628	1.865	-0.436	0.852	(2)	(4)	(1)	(3)
4	5	x	0.7383	-0.340	-0.851	2.526	-0.591	1.154				
5	2	u	0.7498						-0.466	1.840	-0.764	-0.091
5	3	x		-0.185	-0.764	1.840	-0.440	0.875	(2)	(4)	(1)	(3)
5	5	x	0.7529	-0.246	-1.015	2.444	-0.584	1.162				
.	.	.	0.7606	(3)	(1)	(5)	(2)	(4)				
.	.	.										
.	.	.										
20	2	u	0.7800						-0.211	1.556	-1.193	0.178
20	3	x		0.081	-1.193	1.556	-0.409	0.867	(2)	(4)	(1)	(3)
20	5	x	0.7799	0.104	-1.530	1.995	-0.524	1.112				
				(3)	(1)	(5)	(2)	(4)				

$+ 1 \times (-0.68)]/(1 + 2 + 1) = -0.30$, for *Poa annua* -0.33 and for *Ranunculus acris* -0.19. All three species come in Table 5.1c between *Elymus repens* and *Leontodon autumnalis*. The plots asked for suggest unimodal response curves for *Hippophae rhamnoides* and *Poa annua*, but a bimodal curve for *Ranunculus acris*. The species score is the centroid (centre of gravity) of the site scores in which they occur. The score gives an indication of the optimum of the response curve for the former two species, but has no clear meaning for the latter species. In general, species with a score close to the centre of the ordination may either be unimodal, bimodal or unrelated to the axes (Subsection 5.2.5).

Exercise 5.2.2 The weighted average for the site is $[3 \times (-0.65) + 5 \times (-0.50) + 4 \times (-0.39)]/(3 + 5 + 4) = -0.50$, which must be divided as in Exercise 5.1.3 by λ ($= 0.536$) to obtain the site score -0.93. If we calculated the score for the

Table 5.16 Two-way weighted averaging algorithm applied to the data of Exercise 5.1 to obtain the second ordination axis of CA. The first line shows the site scores of the first ordination axis (f). The scores on the second line are used as the initial scores after orthogonalizing them with respect to the first axis (Line 3) and standardizing them (Line 4). Column 5 is v, defined in Table 5.2; the other columns are defined in Table 5.15.

Column					Sites					Species			
1	2	3	4	5	1	2	3	4	5	A	B	C	D
0	4.1	f			0.10	-1.53	2.00	-0.53	1.11				
0	4.1	x			-1.141	-0.489	0.163	0.815	1.466				
0	4.3	x		0.2771	-1.169	-0.065	-0.391	0.962	1.158				
0	5.3	x	0.9612		-1.216	-0.068	-0.407	1.001	1.205				
1	2	u	0.0837							-0.107	0.399	0.288	-0.288
1	3	x			-0.243	0.288	0.399	-0.036	0.056				
1	4	x		0.0001	-0.243	0.288	0.399	-0.036	0.056				
1	5	x	0.2182		-1.114	1.320	1.829	-0.165	0.257				
2	2	u	0.5956							-0.639	1.043	0.825	-0.650
2	3	x			-0.647	0.825	1.043	-0.155	0.197				
2	4	x		-0.0011	-0.647	0.823	1.045	-0.156	0.198				
2	5	x	0.5967		-1.084	1.379	1.751	-0.261	0.332				
3	2	u	0.5980							-0.673	1.042	0.832	-0.636
3	3	x			-0.645	0.832	1.042	-0.159	0.203				
3	4	x		-0.0014	-0.645	0.830	1.045	-0.160	0.205				
3	5	x	0.5982		-1.078	1.387	1.747	-0.267	0.343				
4	2	u	0.5984							-0.672	1.045	0.836	-0.632
4	3	x			-0.642	0.836	1.045	-0.156	0.206				
4	4	x		-0.0016	-0.642	0.834	1.048	-0.157	0.208				
4	5	x	0.5985		-1.073	1.393	1.751	-0.262	0.348				

second axis by the same method, the extra site would come somewhat below Site 5 in the ordination diagram (Figure 5.4).

Exercise 5.2.3 The site scores of Table 5.1c must be divided by $\sqrt{(1-\lambda)/\lambda} = \sqrt{(0.464/0.536)} = 0.93$ and the species scores by $\sqrt{\lambda(1-\lambda)} = \sqrt{(0.536 \times 0.464)} = 0.50$ (Subsection 5.2.2). For Site 20, for example, we obtain the score $1.95/0.93 = 2.10$ and for *Juncus articulatus* $1.28/0.50 = 2.56$. In Hill's scaling, the scores satisfy Equation 5.2 whereas Equation 5.1 must be modified analogously to the modification of Equation 5.2 in Exercise 5.1.3

Exercise 5.3 Principal components analysis

The mean abundance of *Hippophae rhamnoides* is 0.2. With Equation 5.8, we obtain the score $(0\text{-}0.2) \times (-0.31) + (0\ -0.2) \times (-0.30) + \ldots + (2\text{-}0.2) \times (-0.04) + (1\text{-}0.2) \times 0.00 + \ldots + (0\text{-}0.2) \times 0.45 = -0.03$. Similarly we obtain the scores -3.22 for *Poa annua* and -1.48 for *Ranunculus acris*. The plots suggest monotonic decreasing relations for the latter two species, and a unimodal relation

(if any) for the first species. If straight lines are fitted in these plots, the slope of regression turns out to equal the species score (Subsection 5.3.1). The species come at different places in Table 5.5c. For example, *Poa annua* comes just after *Bromus hordaceus*. The score for the extra site is calculated by dividing the weighted sum (Equation 5.9) by the eigenvalue: $3.90/471 = 0.008$.

Exercise 5.4 Length of gradient in DCA

In DCA, axes are scaled such that the standard deviation (tolerance) of the response curve of each species is close to one and is on average equal to one. Each response curve will therefore rise and decline over an interval of about 4 s.d. (Figure 3.6; Figure 5.3b). If the length of the first axis equals 1.5 s.d., the length of the axis covers only a small part of the response curve of each species. Most plots will therefore suggest monotonic curves, although the true response curves may be unimodal (Figure 3.3). If the length of the first axis is 10 s.d., the response curves of many species are contained within the length of the axis, so that many of the plots will suggest unimodal response curves.

5.5 Interpretation of joint plot and biplot

Exercise 5.5.1 Inferred rank orders of abundance are for *Juncus bufonius*

from Figure 5.7 (DCA) sites $12 > 8 > 13 > 9 > 4 \approx 18$
from Figure 5.15 (PCA) sites $13 \approx 3 > 4 > 9 \approx 12$
from Table 5.1a (data) sites $9 = 12 > 13 > 7$ - -

and for *Eleocharis palustris*

from Figure 5.7 (DCA) sites $16 > 14 \approx 15 > 20 > 8$
from Figure 5.15 (PCA) sites $16 > 20 > 15 > 14 > 19$
from Table 5.1a (data) Sites $16 > 15 > 8 = 14 = 20$.

Exercise 5.5.2 The difference in interpretation is greatest for species that lie at the centre of the ordination diagram. In a DCA diagram, the inferred abundance drops with distance from the species point in any direction, whereas in a PCA diagram the inferred abundance decreases or increases with distance from the species point, depending on the direction. This difference is rather unimportant for species that lie on the edge of the diagram, because the site points all lie on one side of the species point. One comes to the same conclusion by noting that a species point in a DCA diagram is its inferred optimum; if the optimum lies far outside the region of the sites the inferred abundance changes monotonically across the region of site points (*Eleocharis palustris* in Figure 5.7).

Exercise 5.6 Detrended canonical correspondence analysis

Exercise 5.6.1 From Table 5.14, we see that the best linear combination is

$x = 1.00\ z_1^* + 0.12\ z_2^*$. In terms of unstandardized variables, we obtain $x = 1.00 \times (z_1 - 22)/18.5 + 0.12 \times (z_2 - 3)/2.9 = (1.00/18.5)z_1 + (0.12/2.9)z_2 - 22/18.5 - 0.12 \times 3/2.9 = 0.054\ z_1 + 0.041\ z_2 - 1.31$. The standard error of c_1 is $0.03/18.5 = 0.00162$ and of c_2 is $0.03/2.9 = 0.010$.

Exercise 5.6.2 The value of x for the plot that was 15 cm above water level in 1984 is $x = 0.054 \times 15 + 0.041 \times 6 = 1.056$ s.d. For the plot 25 cm above water level, we obtain $x = 0.054 \times 25 + 0.041 \times 6 = 1.596$ s.d. Hence, the species turnover is $1.596 - 1.056 = 0.54$ s.d. According to Equation 5.47, turnover depends only on the difference in altitude between the plots: $0.054 \times (25 - 15) = 0.54$, and does not depend on the particular altitudes of the plots nor on the year of sampling. The species turnover between plots differing 10 cm in altitude is therefore 0.54 s.d. on both occasions of sampling.

Exercise 5.6.3 The value of x for a plot with an altitude of 15 cm in 1984 was 1.056 s.d. in 1984 (Exercise 5.6.2) and was $0.054 \times 15 + 0.041 \times 0 = 0.81$ s.d. in 1978. (Note that in the model altitude was not corrected for uplift; hence $z_1 = 15$ in 1984 and in 1978.) The species turnover is $1.056 - 0.81 = 0.246$ s.d., which equals 0.041×6 s.d. and which is independent of altitude. Hence, each plot changes about a quarter standard deviation in 6 years.

Exercise 5.6.4 The species turnover rate is 0.041 s.d. per year, whereas the species turnover due to altitude is 0.054 s.d. per centimetre. The change in altitude that results in 0.041 s.d. species turnover is therefore $0.041/0.054 = 0.76$ cm. An approximate 95% confidence interval can be obtained for this ratio from the standard error of c_1 and c_2 and their covariance by using Fieller's theorem (Finney 1964). Here the covariance is about zero. In this way, we so obtained the interval (0.4 cm, 1.1 cm).

Exercise 5.6.5 From Exercise 5.6.4, we would expect each particular species composition to occur next year 0.76 cm lower than its present position. Uplift (about 0.5 cm per year) is less; hence, there is no evidence that the vegetational succession lags behind the land uplift. The known uplift falls within the confidence interval given above. Further, because the value 0 cm lies outside the confidence interval, the effect of uplift on species composition is demonstrated. Uplift apparently drives the vegetational succession without lag.

Exercise 5.6.6 The length of the first axis is 4.4 s.d. From Exercise 5.6.3, we know that each plot changes about 0.25 s.d. in 6 years. The change from vegetation near the sea to vegetation at the edge of the forest therefore takes roughly $(4.4/0.25) \times 6$ years $\approx$ 100 years. The second axis is 0.9 s.d. and mainly represents the differences in species composition between the two sampling occasions that are unrelated to altitude and land uplift. More precisely, the canonical coefficient of time on the second axis is $-0.34/2.9 = -0.117$. It therefore accounts for 0.117×6 s.d. $= 0.70$ s.d. of the length of the second axis, whereas time accounted for 0.25 s.d. of the length of the first axis. There are apparently more changes going on than can be accounted for by uplift.

6 Cluster analysis

O.F.R. van Tongeren

6.1 Introduction

6.1.1 Aims and use

For ecological data, cluster analysis is a type of analysis that classifies sites, species or variables. Classification is intrinsic in observation: people observe objects or phenomena, compare them with other, earlier, observations and then assign them a name. Therefore one of the major methods used since the start of the study of ecology is the rearrangement of data tables of species by sites, followed by the definition of community types, each characterized by its characteristic species combination (Westhoff & van der Maarel 1978; Becking 1957). Scientists of different schools have different ideas about the characterization of community types and the borders between these types. In vegetation science, for instance, the Scandinavian school and the Zürich-Montpellier school differ markedly, the Scandinavians emphasizing the dominants and the Zürich-Montpellier school giving more weight to characteristic and differential species, which are supposed to have a narrower ecological amplitude and are therefore better indicators for the environment. Cluster analysis is an explicit way of identifying groups in raw data and helps us to find structure in the data. However even if there is a continuous structure in the data, cluster analysis may impose a group structure: a continuum is then arbitrarily partitioned into a discontinuous system of types or classes.

Aims of classification are:

- to give information on the concurrence of species (internal data structure)
- to establish community types for descriptive studies (syntaxonomy and mapping)
- to detect relations between communities and the environment by analysis of the groups formed by the cluster analysis with respect to the environmental variables (external analysis).

In Chapter 6, an introduction will be given to several types of cluster analysis. This chapter aims at a better understanding of the properties of several methods to facilitate the choice of a method, without pretending to show you how to find the one and only best structure in your data. It is impossible to choose a 'best' method because of the heuristic nature of the methods. If there is a markedly discontinuous structure, it will be detected by almost any method, a continuous structure will almost always be obscured by cluster analysis.

6.1.2 Types of cluster analysis

There are several types of cluster analysis, based on different ideas about the cluster concept. Reviews are found mainly in the taxonomic literature (Lance & Williams 1967; Dunn & Everitt 1982). Here a brief summary will be given of the main groups.

A major distinction can be made between divisive and agglomerative methods. Divisive methods start with all objects (in ecology mostly samples; in taxonomy operational taxonomic units, OTUs) as a group. First this group is divided into two smaller groups, which is repeated subsequently for all previously formed groups, until some kind of 'stopping rule' is satisfied. The idea of this way of clustering is that large differences should prevail over the less important smaller differences: the global structure of a group should determine the subgroups. Alternatively agglomerative methods start with individual objects, which are combined into groups by collection of objects or groups into larger groups. Here 'local' similarity prevails over the larger differences. Divisive methods will be described in Section 6.3, and agglomerative methods in Section 6.2. Most agglomerative methods require a similarity or dissimilarity matrix (site by site) to start from. Several indices of (dis)similarity will be introduced in Subsection 6.2.3.

A second way of distinguishing methods is to classify them by hierarchical and non-hierarchical methods. Hierarchical methods start from the idea that the groups can be arranged in a hierarchical system. In ecology, one could say that a certain difference is more important than another one and therefore should prevail: be expressed at a higher hierarchical level. Non-hierarchical methods do not impose such a hierarchical structure on the data. For data reduction, non-hierarchical methods are usually used.

Non-hierarchical classification handles

- redundancy: sites that are much like many other sites are grouped without considering the relations to other less similar sites
- noise: before subsequent hierarchical clustering, a 'composite sample' may be constructed
- outliers, which can be identified because they appear in small clusters or as single samples.

6.2 Agglomerative methods

6.2.1 Introduction

Agglomerative cluster analysis starts from single objects, which are agglomerated into larger clusters. In many sciences, agglomerative techniques are employed much more frequently than divisive techniques. The historical reason for this is the inefficient way early polythetic divisive techniques used computer resources, while the agglomerative ones were more efficient. Now, the opposite seems true. Nevertheless, there is a very large range of agglomerative techniques, each emphasizing other aspects of the data and therefore very useful.

All agglomerative methods share the idea that some kind of (dis)similarity

function between (groups of) objects (usually sites) is decisive for the fusions. Different methods, however, are based on different ideas on distance between clusters. Within most methods, there is also a choice between different 'indices of similarity or dissimilarity' (distance functions). Most of this section is devoted to similarity and dissimilarity indices.

6.2.2 Similarity and dissimilarity

Grouping of sites and species in many ecological studies is a matter of personal judgment on the part of the investigator: different investigators may have different opinions or different aims; they therefore obtain different results. There are, however, many different objective functions available with which to express similarity.

Ideally, similarity of two sites or species should express their ecological relation or resemblance; dissimilarity of two sites or species is the complement of their similarity. Since this idea of similarity includes an ecological relation, it is important which ecological relation is focused upon – so the objectives of a study may help to determine the applicability of certain indices. Most of the indices used in ecology do not have a firm theoretical basis. My attitude towards this problem is that practical experience, as well as some general characteristics of the indices, can help us choose the right one. Numerous indices of similarity or dissimilarity have been published, some of them are widely used, others are highly specific.

The aim of this section is to make the concepts of similarity and dissimilarity familiar and to examine some of the popular indices. Although most indices can be used to compute (dis)similarities between sites as well as between species, they are demonstrated here as if the site is the statistical 'sampling unit'. Computations of similarity can be made directly from the species-abundance values of sites or indirectly, after using some ordination technique from the site scores on the ordination axes. With indirect computation, dissimilarities refer to distances between sites in the ordination space.

Comparison of sites on the basis of presence-absence data

If detailed information on species abundance is irrelevant for our problem or if our data are qualitative (e.g. species lists), we use an index of similarity for qualitative characters. The basis of all similarity indices for qualitative characters is that two sites are more similar if they share more species and that they are more dissimilar if there are more species unique for one of both (two species are more similar if their distribution over the sites is more similar). One of the earliest indices is the index according to Jaccard (1912). This Jaccard index is the proportion of species out of the total species list of two sites, which is common to both sites:

$$SJ = c/(a + b + c) \qquad \text{Equation 6.1}$$

where
SJ is the similarity index of Jaccard
c is the number of species shared by the two sites
a and b are the numbers of species unique to each of the sites.

Often the equation is written in a different way:

$$SJ = c/(A + B - c) \qquad \text{Equation 6.2}$$

where c is the number of species shared and A and B are the total numbers of species for the samples: $A = a + c$ and $B = b + c$).

Sørensen (1948) proposed another similarity index, often referred to as coefficient of community *(CC).*

$$CC = 2c/(A + B) \text{ or } 2c/(a + b + 2c). \qquad \text{Equation 6.3}$$

Instead of dividing the number of species shared by the total number of species in both samples, the number of species shared is divided by the average number of species. Faith (1983) discusses the asymmetry of these indices with respect to presence or absence.

Comparison of samples on the basis of quantitative data

Quantitative data on species abundances always have many zeros (i.e. species are absent in many sites); the problems arising from this fact have been mentioned in Section 3.4. Therefore an index of similarity for quantitative characters should also consider the qualitative aspects of the data. The similarity indices in this subsection are different with respect to the weight that is given to presence or absence (the qualitative aspect) with regard to differences in abundance when the species is present. Some of them emphasize quantitative components more than others. Two of them are very much related to the Jaccard index and the coefficient of community, respectively: similarity ratio (Ball 1966) and percentage similarity (e.g. Gauch 1982). The other indices can easily be interpreted geometrically.

The similarity ratio is:

$$SR_{ij} = \Sigma_k\, y_{ki}\, y_{kj}/(\Sigma_k\, y_{ki}^2 + \Sigma_k\, y_{kj}^2 - \Sigma_k\, y_{ki}\, y_{kj}) \qquad \text{Equation 6.4}$$

where y_{ki} is the abundance of the k-th species at site i, so sites i and j are compared. For presence–absence data (0 indicating absence and 1 presence), this can be easily reduced to Equation 6.1, indicating that the Jaccard index is a special case of the similarity ratio. For Sørensen's index, Equation 6.3, the same can be said in respect to percentage similarity:

$$PS_{ij} = 200\, \Sigma_k \min(y_{ki}, y_{kj})/(\Sigma_k y_{ki} + \Sigma_k y_{kj}) \qquad \text{Equation 6.5}$$

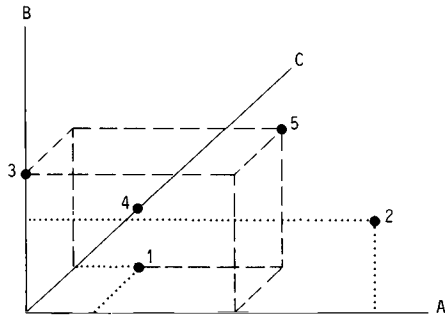


Figure 6.1 Five sites (1-5) in a three-dimensional space of which the axes are the thre species A, B and C. Site 1 is characterized by low abundance of species A and specie C, and absence of species B. In site 2, species A is dominant, species B is less importan and species C is absent. Sites 3 and 4 are monocultures of species B and C, respectively Site 5 has a mixture of all three species.

where min (y_{ki}, y_{kj}) is the minimum of y_{ki} and y_{kj}.

Some indices can be explained geometrically. To explain these indices, it i necessary to represent the sites by a set of points in a multi-dimensional spac (with as many dimensions as there are species). One can imagine such a spac with a maximum of three species (Figure 6.1) but conceptually there is no differenc if we use more species (see Subsection 5.3.3).

The position of a site is given by using the abundances of the species as coordinate (Figure 6.1), and therefore sites with similar species composition occupy nearb positions in species space. The Euclidean Distance, *ED*, between two sites is a obvious measure of dissimilarity:

$$ED = \sqrt{\Sigma_k (y_{ki} - y_{kj})^2} \qquad \text{Equation 6.}$$

Figure 6.1 shows that quantitative aspects play a major role in Euclidean Distance the distance between Sites 1 and 2, which share one species, is much larger tha the distance between Sites 1 and 3, not sharing a species.

More emphasis is given to qualitative aspects by not considering a site as point but as a vector (Figure 6.2). Understandably, the direction of this vecto tells us something about the relative abundances of species. The similarity o two sites can be expressed as some function of the angle between the vector of these sites. Quite common is the use of the cosine (or Ochiai coefficient):

$$cos = OS = \Sigma_k y_{ki} y_{kj} / \sqrt{\Sigma_k y_{ki}^2 \Sigma_k y_{kj}^2} \qquad \text{Equation 6.}$$

A dissimilarity index that is more sensitive to qualitative aspects than the Euclidea Distance is the chord distance:

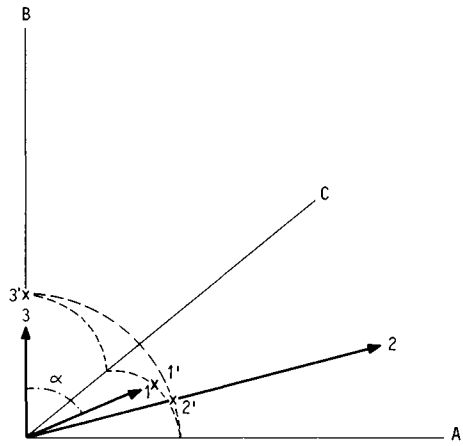


Figure 6.2 The same space as in Figure 6.1. Samples are now indicated by vectors. Crosses indicate where the sample vectors intersect the unit sphere (broken lines). Note that the distance between 1′ and 2′ is much lower than the distance between either of them and 3′. The angle between sample vectors 1 and 3 is indicated by α.

$$CD = \sqrt{\Sigma_k \left[y_{ki} / (\Sigma_k y_{ki}^2)^{1/2} - y_{kj} / (\Sigma_k y_{kj}^2)^{1/2} \right]^2} \qquad \text{Equation 6.8}$$

This chord distance is geometrically represented by the distance between the points where the sample vectors intersect the unit sphere (Figure 6.2).

Conversion of similarity to dissimilarity and vice versa

For some applications, one may have to convert a similarity index into a dissimilarity index. This conversion must be made if, for instance, no dissimilarity index with the desired properties is available, but the cluster algorithm needs an index of dissimilarity. For cluster algorithms merely using the rank order of dissimilarities, any conversion reversing the rank order is reasonable, but care must be taken for cluster algorithms that use the dissimilarities in a quantitative way (as forming an interval or ratio scale (Subsection 2.4.2). We mention two ways of making the conversion:

- subtracting each similarity value from a certain value: in this way the intervals between the values are preserved. Bray & Curtiss (1957), for instance, subtract similarity values from the expected similarity among replicate samples, the so-called internal association. In practice, the best estimate of internal association *(IA)* is the maximum similarity observed. Thus percentage similarity is converted to percentage distance, *PD*, using this subtraction:

$$PD = IA - PS \qquad \text{Equation 6.9}$$

- taking the reciprocal of each similarity value. In this way, the ratios between similarity values are preserved in the dissimilarity matrix.

6.2.3 Properties of the indices

Despite many studies (e.g. Williams et al. 1966; Hajdu 1982) addressing the problem of which index should be used, it is still difficult to give a direct answer. The choice of index must be guided by best professional judgment (or is it intuition?) of the investigator, by the type of data collected and by the ecological question that should be answered. Dunn & Everitt (1982) and Sneath & Sokal (1973) advise to choose the simplest coefficient applicable to the data, since this choice will generally ease the interpretation of final results.

However one can say a little bit more, though still not very much: the objectives of a study may help in deciding which index is to be applied. The length of the sampled gradient is important: the relative weight that is given to abundance (quantity) should be larger for short gradients, the relative weight given to presence or absence should be larger for long gradients (Lambert & Dale 1964; Greig-Smith 1971). Other criteria that should be considered are species richness (Is it very different at different sites?) and dominance or diversity of the sites (Are there substantial differences between sites?). The easiest way of getting some feeling for these aspects is to construct hypothetical matrices of species abundances and see how the various indices respond to changes in different aspects of the data. However this gives only an indication and one must be aware of complications whenever more characteristics of the data are different between samples.

To demonstrate the major responses to dominance/diversity, species richness and length of gradient a set of artificial species-by-site data, each referring to one major aspect of ecological samples, is given, together with graphs, to indicate the responses of the indices (Tables 6.1-6.4). To obtain comparable graphs (Figures

Table 6.1 Artificial species-by-sites table. Total abundance for each sample is 10, the number of species (α-diversity) is lower on the right side and the 'evenness' is constant (equal scores for all species in each sample).

Site	1	2	3	4	5	6	7	8	9	10
Species										
A	1.00	1.11	1.25	1.43	1.67	2.00	2.50	3.33	5.00	10.00
B	1.00	1.11	1.25	1.43	1.67	2.00	2.50	3.33	5.00	
C	1.00	1.11	1.25	1.43	1.67	2.00	2.50	3.33		
D	1.00	1.11	1.25	1.43	1.67	2.00	2.50			
E	1.00	1.11	1.25	1.43	1.67	2.00				
F	1.00	1.11	1.25	1.43	1.67					
G	1.00	1.11	1.25	1.43						
H	1.00	1.11	1.25							
I	1.00	1.11								
J	1.00									

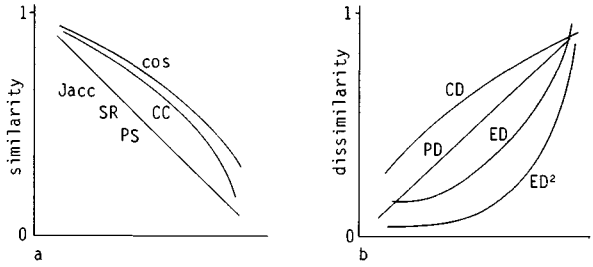


Figure 6.3 Standardized (dis)similarity (ordinate) between the first site and each of the other sites in Table 6.1 in corresponding order on the abscissa. Note that squared Euclidean Distance (ED^2) is strongly affected by higher abundances. a: similarity indices. b: dissimilarity indices.

Table 6.2 Artificial species-by-sites table. Evenness and species number are constant, the sample totals varying largely.

Site	1	2	3	4	5	6	7	8	9	10
Species										
A	1	2	3	4	5	6	7	8	9	10
B	1	2	3	4	5	6	7	8	9	10
C	1	2	3	4	5	6	7	8	9	10
D	1	2	3	4	5	6	7	8	9	10
E	1	2	3	4	5	6	7	8	9	10
F	1	2	3	4	5	6	7	8	9	10
G	1	2	3	4	5	6	7	8	9	10
H	1	2	3	4	5	6	7	8	9	10
I	1	2	3	4	5	6	7	8	9	10
J	1	2	3	4	5	6	7	8	9	10

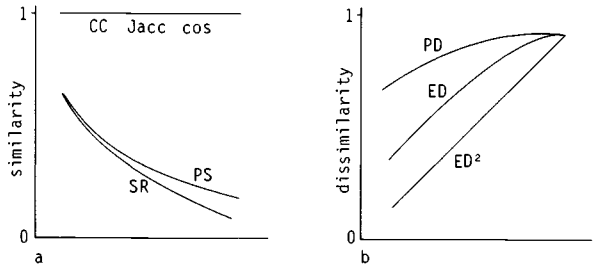


Figure 6.4 Standardized (dis)similarity (ordinate) between the first site and each of the other sites in Table 6.2 in corresponding order on the abscissa. Note that coefficient of community (CC), Jaccard index (Jacc) and cosine are at their maximum for all sites compared with the first site. a: similarity indices. b: dissimilarity indices.

Table 6.3 Artificial species-by-sites table. Number of species (2) and total abundance (10) are constant but the evenness varies.

Site	1	2	3	4	5	6	7	8	9
Species									
A	1	2	3	4	5	6	7	8	9
B	9	8	7	6	5	4	3	2	1

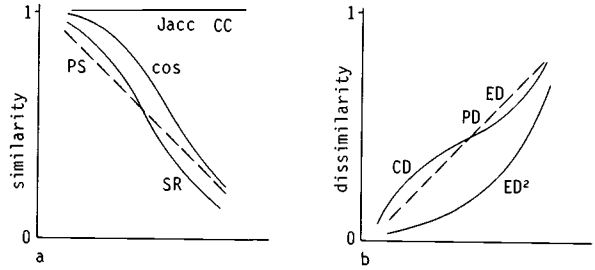


Figure 6.5 Standardized (dis)similarity (ordinate) between the first site and each of the other sites in Table 6.3 in corresponding order on the abscissa. Note the difference with Figure 6.4: the cosine is not at its maximum value for other sites as compared to the first site. a: similarity indices. b: dissimilarity indices.

Table 6.4 Artificial species-by-sites table. A regular gradient with equal species numbers in the samples, equal scores for all species: at each 'step' along the gradient one species is replaced by a new one.

Site	1	2	3	4	5	6	7	8	9
Species									
A	1								
B	1	1							
C	1	1	1						
D	1	1	1	1					
E	1	1	1	1	1				
F	1	1	1	1	1	1			
G		1	1	1	1	1	1		
H			1	1	1	1	1	1	
I				1	1	1	1	1	1
J					1	1	1	1	1
K						1	1	1	1
L							1	1	1
M								1	1
N									1

6.3-6.6), all indices are scaled from 0 to 1. Comparisons are always made between the first site of the artificial data and the other sites within that table. The captions to Tables 6.1-6.4 and Figures 6.3-6.6 give more information on the properties of the artificial data. Table 6.5 summarizes the major characteristics of the indices but it is only indicative.

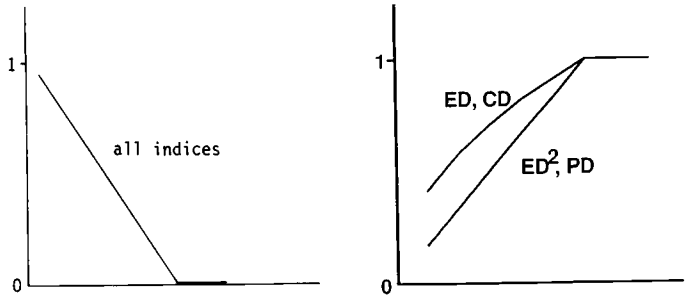


Figure 6.6 Standardized (dis)similarity (ordinate) between the first site and each of the other sites in Table 6.4 in corresponding order on the abscissa. Except for Euclidean Distance (ED) and Chord Distance (CD), all indices are linear until a certain maximum (or minimum) is reached, a: similarity indices. b: dissimilarity indices.

Table 6.5 Characteristics of the (dis)similarity indices. The asterisk (*) indicates qualitative characteristics. Sensitivity for certain properties of the data is indicated by: − not sensitive; + sensitive; ++ and +++ strongly sensitive.

	sensitivity to sample total	sensitivity to dominant species	sensitivity to species richness	similarity	dissimilarity	quantitative	qualitative	abbreviation
Similarity Ratio				*		*	*	SR
Percentage Similarity				*		*	*	PS
Cosine	+	+	+	*		*	*	Cos
Jaccard Index	−	−	++	*			*	SJ
Coefficient of Community	−	−	+	*			*	CC
Cord Distance	+	+	+		*	*	*	CD
Percentage Dissimilarity	+	+	++		*	*	*	PD
Euclidean Distance	++	++	++		*	*	*	ED
Squared Euclidean Distance	+++	+++	+++		*	*	*	ED^2

6.2.4 *Transformation, standardization and weighting*

Transformation, standardization and weighting of data are other ways of letting certain characteristics of the data express themselves more or less strongly. This paragraph is meant to give you some idea of how certain manipulations can be made with the data and what are the reasons for and the consequences of transformations and standardizations.

Transformation

Transformations are possible in many different ways. Most transformations used in ecology are essentially non-linear transformations: the result of such transformations is that certain parts of the scale of measurement for the variables are shrunk, while other parts are stretched.

Logarithmic transformation.

$y_{ij}^* = \log_e y_{ij}$ or (if zeros are present) $y_{ij}^* = \log_e (y_{ij} + 1)$ Equation 6.10

This transformation is often used for three essentially different purposes:

- to obtain the statistically attractive property of normal distribution for log-normally distributed variables (as in Subsection 2.4.4)
- to give less weight to dominant species, in other words to give more weight to qualitative aspects of the data
- in environmental variables, to reflect the linear response of many species to the logarithm of toxic agents or (in a limited range) to the logarithm of nutrient concentrations. Instead of '+1', take the minimum non-zero value.

Square-root transformation.

$y_{ij}^* = y_{ij}^{1/2}$ Equation 6.11

This transformation is used

- before analysis of Poisson-distributed variables (e.g. number of individuals of certain species caught in a trap over time)
- to give less weight to dominant species.

Exponential transformation.

$y_{ij}^* = a^{y_{ij}}$. Equation 6.12

If a is a real number greater than 1, the dominants are emphasized.

Transformation to an ordinal scale. The species abundances are combined into classes. The higher the class number, the higher the abundance. A higher class number always means a higher abundance, but an equal class number does not always mean an equal abundance: intervals between classes are almost meaningless. Dependent on the class limits, one can influence the results of a classification in all possible ways. An extreme is the transformation to presence–absence scale

(1/0). A transformation to ordinal scale always includes loss of information: if continuous data are available any other transformation is to be preferred. However it can be very useful to collect data on an ordinal scale (as is done in the Zürich–Montpellier school of vegetation science) for reduction of the work in the field.

Standardization

Several aspects of standardization have been treated in Subsection 2.4.4. Here we discuss some other types of standardization that are used in cluster analysis. Standardization can here be defined as the application of a certain standard to all variables (species) or objects (sites) before the computation of the (dis)similarities or before the application of cluster analysis. Possible ways of standardizing are as follows.

Standardization to site total. The abundances for each species in a site are summed and each abundance is divided by the total: in this way relative abundances for the species are computed, a correction is made for 'size' of the site (total number of individuals collected at the site or total biomass). Care should be taken if these sizes are very different, because rare species tend to appear in large sites: (dis)similarity measures that are sensitive to qualitative aspects of the data might still be inappropriate.

Standardization to species total. For each species the abundances are summed over all sites and then divided by the total. This standardization strongly over-weights the rare species and down-weights the common species. It is therefore recommended to use this standardization only if the species frequencies in the table do not differ too much. This type of standardization can be applied when different trophic levels are represented in the species list, because the higher trophic levels are less abundant (pyramids of biomass and numbers).

Standardization to site maximum. All species abundances are divided by the maximum abundance reached by any species in the site. This standardization is applied for the same reason as standardization to site total. It is less sensitive to species richness, but care should be taken if there are large differences in the 'evenness' of sites. If an index is used with a large weighting for abundance, sites with many equal scores will become extremely different from sites with a large range in their scores.

Standardization to species maximum. The reason for this standardization is that, in the opinion of many ecologists, less abundant species (in terms of biomass or numbers) should be equally weighted. As the standardization to species total, this type of standardization is recommended when different trophic levels are present in the species list. This standardization also makes data less dependent on the kind of data (biomass or numbers or cover) collected.

Standardization to unit site vector length. By dividing the species abundance in a site by the square root of their summed squared abundances, all end-points of the site vectors are situated on the unit sphere in species-space. Euclidean Distance then reduces to chord distance.

Weighting

There are several reasons for weighting species or sites. Depending on the reason for down-weighting several kinds of down-weighting can be applied.
Down-weighting of rare species. A lower weight dependent on the species frequency, is assigned to rare species to let them influence the final result to a lesser extent. This should be done if the occurrence of these species is merely by chance and if the (dis)similarity index or cluster technique is sensitive to rare species.
Down-weighting of species indicated by the ecologist. A lower weight is assigned to species (or sites) that are less reliable (determination of a species is difficult; a sample is taken by an inexperienced field-worker) or to species that are ecologically less relevant (planted trees; the crop species in a field). This kind of down-weighting is ad hoc and therefore arbitrary.

6.2.5 Agglomerative cluster algorithms

All agglomerative methods are based on fusion of single entities (sites) or clusters (groups of sites) into larger groups. The two groups that closest resemble each other are always fused, but the definition of (dis)similarity between groups differs between methods.

Often the results of hierarchical clustering are presented in the form of a dendrogram (tree-diagram, e.g. Figure 6.8). Such a dendrogram shows the relations between sites and groups of sites. The hierarchical structure is indicated by the branching pattern.

Single-linkage or nearest-neighbour clustering

The distance between two clusters is given by the minimum distance that can be measured between any two members of the clusters (Figure 6.7). A dendrogram of the classification of the Dune Meadow Data with single-linkage clustering,

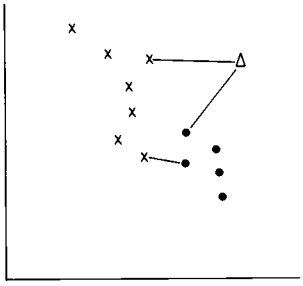


Figure 6.7 Distances (solid lines) between clusters in single linkage: samples within the same cluster are indicated with the same symbol.

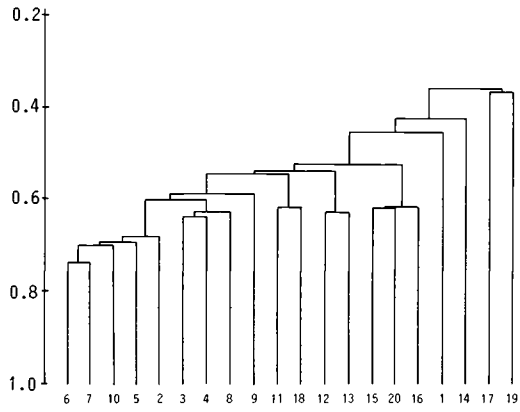


Figure 6.8 Dendrogram of single linkage, using the Dune Meadow Data and the similarity ratio.

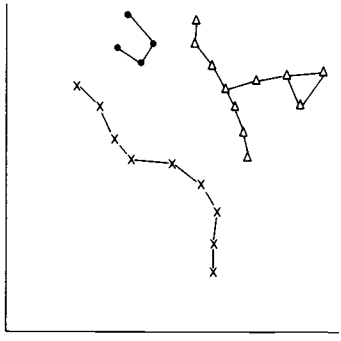


Figure 6.9 Hypothetical example of 'chaining', a problem occurring in single-linkage clustering.

using similarity ratio, is given in Figure 6.8. The dendrogram shows us that there are not very well defined clusters: our data are more or less continuous. Single-linkage clustering can be used very well to detect discontinuities in our data. For other research in community ecology, it is not appropriate because of its tendency to produce straggly clusters, distant sites being connected by chains of sites between them (Figure 6.9).

Complete-linkage or furthest-neighbour clustering

In contrast to the definition of distance in single-linkage clustering, the definition in complete-linkage clustering is as follows. The distance between two clusters is given by the maximum distance between any pair of members (one in each cluster) of both clusters (Figure 6.10). The dendrogram (Figure 6.11) suggests clear groups but, as can be seen in Figure 6.8, this may be an artefact. The group structure is imposed on the data by complete linkage: complete linkage tends to tight clusters, but between-cluster differences are over-estimated and therefore exaggerated in the dendrogram.

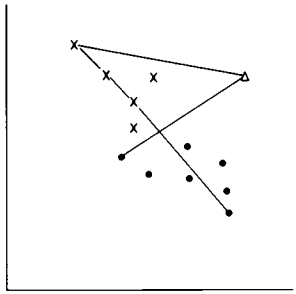


Figure 6.10 Distances (solid lines) between clusters in complete linkage: samples within the same cluster are indicated with the same symbol.

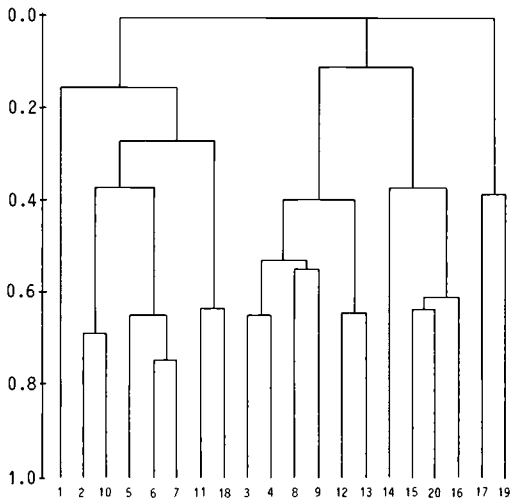


Figure 6.11 Complete-linkage dendrogram of the Dune Meadow Data using the similarity ratio.

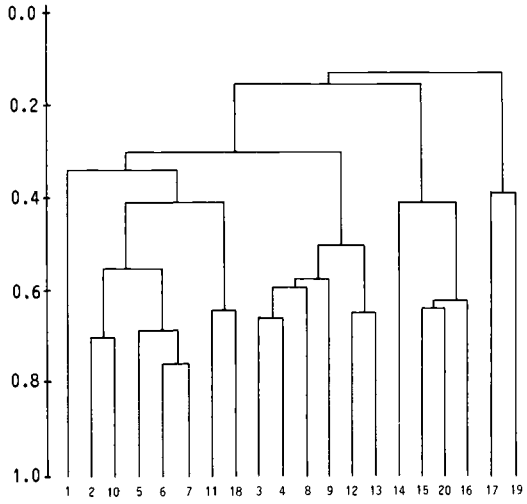


Figure 6.12 Average-linkage dendrogram of the Dune Meadow Data using the similarity ratio.

Average-linkage clustering

In average-linkage clustering, the between-group (dis)similarity is defined as the average (dis)similarity between all possible pairs of members (one of each group). This method is most widely used in ecology and in systematics (taxonomy). The algorithm maximizes the 'cophenetic correlation', the correlation between the original (dis)similarities and the (dis)similarities between samples, as can be derived from the dendrogram. For any sample pair, it is the lowest dissimilarity (or highest similarity) required to join them in the dendrogram (Sneath & Sokal 1973). As can be seen in the dendrogram of average linkage (Figure 6.12), this method is intermediate between complete and single linkage. The preceding explanation refers to UPGMA, the unweighted-pair groups method (Sokal & Michener 1958). There are variants of this technique in which a weighted average is computed (e.g. Lance & Williams 1967).

Centroid clustering

In centroid clustering, between-cluster distance is computed as the distance between the centroids of the clusters. These centroids are the points in species space defined by the average abundance value of each species over all sites in a cluster (Figure 6.13). Figure 6.14 shows subsequent steps in centroid clustering. For the Dune Meadow Data, the dendrogram (which is not presented) closely resembles the average-linkage dendrogram (Figure 6.12).

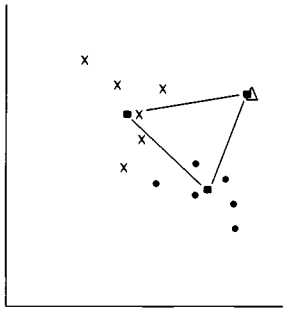


Figure 6.13 Between-cluster distances (solid lines) in centroid clustering: samples within the same cluster are indicated with the same symbol; cluster centroids are indicated by squares.

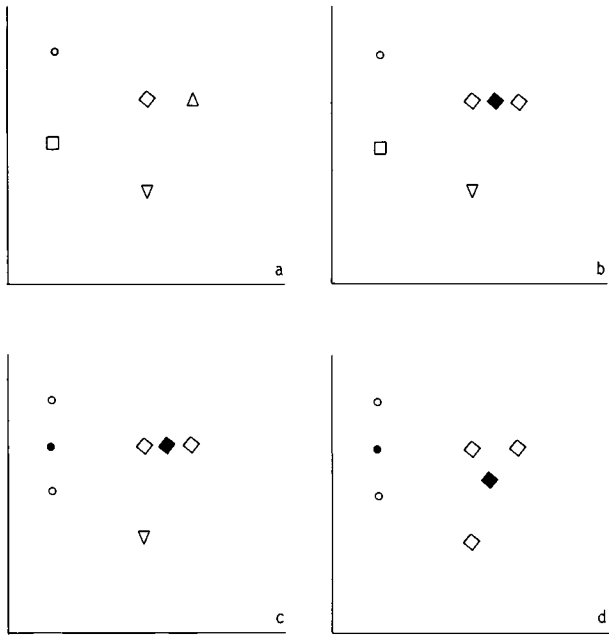


Figure 6.14 Subsequent steps in centroid clustering. Sites belonging to the same cluster are indicated with the same open symbol. Cluster centroids are indicated by the corresponding filled symbols. Upon fusion of two clusters, the symbols of sites change to indicate the new cluster to which they belong.

Ward's method or minimum variance clustering

Ward's method, also known as Orlóci's (1967) error sum of squares clustering, is in some respects similar to average-linkage clustering and centroid clustering. Between-cluster distance can either be computed as a squared Euclidean distance between all pairs of sites in a cluster weighted by cluster size (resembling average-linkage clustering) or as an increment in squared distances towards the cluster centroid when two clusters are fused (resembling centroid clustering). Penalty by squared distance and cluster size makes the clusters tighter than those in centroid clustering and average linkage, and more like those obtained in complete linkage. The algorithm proceeds as follows: with all samples in a separate cluster, the sum of squared distances is zero, since each sample coincides with the centroid of its cluster. In each step, the pair of clusters is fused, which minimizes the total within-group sum of squares (Subsection 3.2.1, residual sum of squares), which is equal to minimizing the increment (dE) in the total sum of squares:

$$dE = E_{p+q} - E_p - E_q$$

where
E is the total error sum of squares
E_{p+q} is the within-group sums of squares for the cluster in which p and q are fused together
E_p and E_q the sums of squares for the individual clusters p and q.

The within-group sum of squares for a cluster is:

$$E_p = 1/N \, \Sigma_{i \varepsilon p} \, \Sigma_k \, (y_{ki} - \bar{y}_k)^2$$

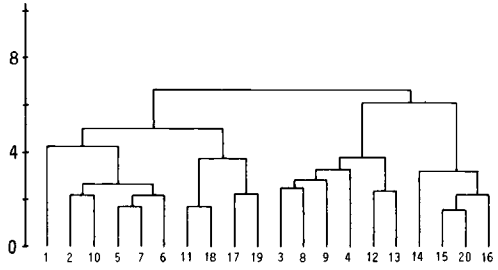
where the first summation is over all members of cluster p and the second summation is over all species.

The dendrograms of Ward's clustering, average linkage and complete linkage using squared Euclidean Distance are given in Figure 6.15.

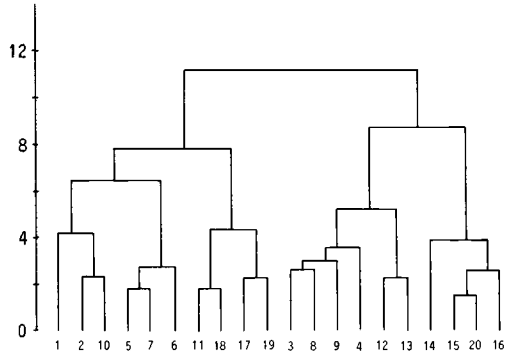
6.3 Divisive methods

6.3.1 Introduction

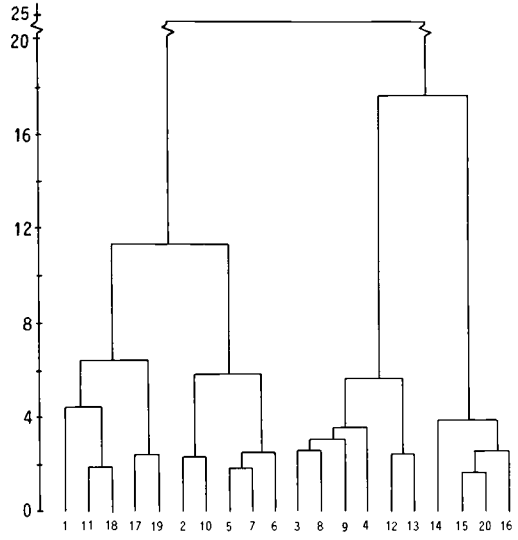
Divisive methods have long been neglected. The reason for this is that they were developed in the early years of numerical data analysis. At that time they failed either because of inefficiency (too many computational requirements) or because the classification obtained was inappropriate. Williams & Lambert (1960) developed the first efficient method for divisive clustering: association analysis. This method is monothetic: divisions are made on the basis of one attribute (e.g. character or species). Although it is not used very often now, some authors still use association analysis or variations of association analysis (e.g. Kirkpatrick et


8
4
0
1 2 10 5 7 6 11 18 17 19 3 8 9 4 12 13 14 15 20 16

a


12
8
4
0
1 2 10 5 7 6 11 18 17 19 3 8 9 4 12 13 14 15 20 16

b


25
20
16
12
8
4
0
1 11 18 17 19 2 10 5 7 6 3 8 9 4 12 13 14 15 20 16

c

al. 1985). Subsection 6.3.2 briefly describes association analysis. Efficient polythetic methods for divisive clustering appeared after Roux & Roux (1967) introduced the partitioning of ordination space for purposes of classification. Lambert et al. (Williams 1976b) wrote a program to perform a division based on partitioning of the first axis of principal component analysis. A divisive clustering is obtained by repeating this method of partitioning on the groups obtained in the previous step. More recently, Hill (Hill et al. 1975; Hill 1979b) developed a method based on the partitioning of the first axis of CA. Since this method has some remarkable features and in most cases leads to very interpretable solutions it will be treated in detail in Subsection 6.3.3.

6.3.2 Association analysis and related methods

Association analysis (Williams & Lambert 1959 1960 1961) starts selecting the species that is maximally associated to the other species: association between species is estimated as the qualitative correlation coefficient for presence–absence data, regardless of its sign. For each species, the sum of all associations is computed. The species having the highest summed association value is chosen to define the division. One group is the group of sites in which the species is absent, the other group is the group of sites in which the species is present. Because it is sensitive to the presence of rare species and to the absence of more common ones this method is not often used in its original form. Other functions defining association, chi-square and information statistics have been proposed. These functions produce better solutions. Groups obtained in monothetic methods are less homogeneous than groups resulting from polythetic methods, because in the latter case more than one character determines the division. Therefore if a polythetic method is available it should always be preferred over a monothetic one (Coetzee & Werger 1975; Hill et al. 1975).

6.3.3 Two Way INdicator SPecies ANalysis

This section deals with the method of Two Way INdicator SPecies ANalysis (TWINSPAN). The TWINSPAN program by Hill (1979b) not only classifies the sites, but also constructs an ordered two-way table from a sites-by-species matrix. The process of clustering sites and species and the construction of the two-way table are explained step by step to illustrate TWINSPAN's many features, some of which are available in other programs too. However the combination of these features in TWINSPAN has made it one of the most widely used programs in community ecology.

Figure 6.15 Comparison of average linkage, complete linkage and Ward's method using squared Euclidean Distance. a: average linkage. b: complete linkage. c: Ward's method. The dendrograms for average linkage and complete linkage are similar. By the use of squared Euclidean Distance, the larger distances have a higher weighting in average linkage. The result of Ward's method is different from both other methods, even at the four-cluster level.

Pseudo-species

One of the basic ideas in TWINSPAN stems from the original idea in phytosociology that each group of sites can be characterized by a group of differential species, species that appear to prevail in one side of a dichotomy. The interpretation of TWINSPAN results is, in this respect, similar to the interpretation of a table rearranged by hand. Since the idea of a differential species is essentially qualitative, but quantitative data must be handled effectively also, Hill et al. (1975) developed a qualitative equivalent of species abundance, the so-called pseudo-species (see Section 3.4). Each species abundance is replaced by the presence of one or more pseudo-species. The more abundant a species is, the more pseudo-species are defined. Each pseudo-species is defined by a minimum abundance of the corresponding species, the 'cut level'. This way of substituting a quantitative variable by several qualitative variables is called conjoint coding (Heiser 1981). An advantage of this conjoint coding is that if a species' abundance shows a unimodal response curve along a gradient, each pseudo-species also shows a unimodal response curve (see Section 3.4), and if the response curve for abundance is skewed, then the pseudo-species response curves differ in their optimum.

Making a dichotomy; iterative character weighting

A crude dichotomy is made by ordinating the samples. In TWINSPAN, this is done by the method of correspondence analysis (Hill 1973; Section 5.2) and division of the first ordination axis at its centre of gravity (the centroid). The groups formed are called the negative (left-hand) and positive (right-hand) side of the dichotomy. After this division the arrangement is improved by a process that is comparable to iterative character weighting (Hogeweg 1976) or to the application of a transfer algorithm (Gower 1974) that uses a simple discriminant function (Hill 1977). What follows is an account of this process of iterative character weighting in some more details; the reader may skip the rest of this passage at first reading.

A new dichotomy is constructed by using the frequencies of the species on the positive and negative sides of the first, crude dichotomy: differential species (species preferential for one of the sides of the dichotomy) are identified by computing a preference score. Positive scores are assigned to the species with preference for the positive side of the dichotomy, negative scores for those preferential for the negative side. An absolute preference score of 1 is assigned to each pseudo-species that is at least three times more frequent on one side of the dichotomy as on the other side. Rare pseudo-species and pseudo-species that are less markedly preferential are down-weighted. A first ordering of the sites is obtained by adding the species preference scores to each other as in PCA (Chapter 5, Equation 5.9). This weighted sum is standardized so that the maximum absolute value is 1. A second ordering is constructed by computing for each site the average preference scores (similar to the computation of weighted averages in correspondence analysis (Chapter 5, Equation 5.2)) without down-weighting

of the rare species. In comparison to the first ordering, this one polarizes less strongly when there are many common (non-preferential) species, which is to be expected at the lower levels of the hierarchy. At the higher levels of the hierarchy it polarizes more strongly than the first ordination because more rare species can be expected at the higher levels. Hill's preference scores have a maximum absolute value of 1, so the scores for the sites in this second ordering range from −1 to 1. The scores in both orderings are added to each other and this so-called refined ordination is divided at an appropriate point near its centre (see Hill 1979b). The refined ordination is repeated using the refined classification. With the exception of a few 'borderline' cases, this refined ordination determines the dichotomy. For borderline cases (sites that are close to the point where the refined ordination is divided), the final decision is made by a third ordination: the indicator ordination. The main aim of this indicator ordination is, however, not to assign these borderline cases to one of the sides of the dichotomy, but to reproduce the dichotomy suggested by the refined ordination by using simple discriminant functions based on a few of the most highly preferential species.

Hill (1979b) warns of confusion arising from the terms 'Indicator Species Analysis' in TWINSPAN's name, because indicator ordination is an appendage, not the real basis, of the method. He suggests the name 'dichotomized ordination analysis' as a generic term to describe a wide variety of similar methods (e.g. the program POLYDIV of Williams (1976b)). The indicator species (the set of most highly preferential species that reproduce as good a refined ordination as possible) can be used afterwards in the field to assign a not-previously-sampled stand to one of the types indicated by TWINSPAN.

The construction of a species-by-sites table

For the construction of a species-by-sites table two additional features are necessary. First, the dichotomies must be ordered and, second, the species must be classified. The order of the site groups is determined by comparison of the two site groups formed at any level with site groups at two higher hierarchical levels. Consider the hierarchy in Figure 6.16. Assume that the groups 4, 5, 6 and 7 have already been ordered. The ordering of subsequent groups is now decided upon. Without ordering we are free to swivel each of the dichotomies, and therefore

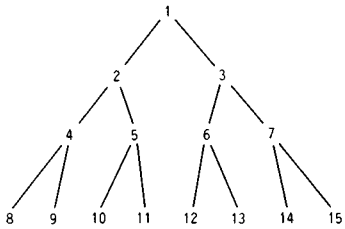


Figure 6.16 TWINSPAN dichotomy; cluster numbers are the numbers used by TWINSPAN.

Table 6.6 TWINSPAN table of the demonstration samples. Options are default TWINSPAN options, except the cut levels, which are 1, 2, 3, ... 9. Zeros and ones on the right-hand side and at the bottom of the table indicate the dichotomies.

```
                1111    1       111112
                17895670123489234560

 3 Air pra .2.3.................. 00000
12 Emp nig ...2.................. 00000
13 Hyp rad 22.5.................. 00000
28 Vic lat 2.1....1.............. 00000
 5 Ant odo .4.44324............. 00001
18 Pla lan 323.5553............. 00010
 1 Ach mil .2..222413.......... 000110
26 Tri pra ....252.............. 000110
 6 Bel per ..2.2..2.322........ 000111
 7 Bro hor ....2.24.4.3........ 000111
 9 Cir arv ...........2........ 000111
11 Ely rep ....4...4444.6...... 001
17 Lol per 7.2.2666758542...... 001
19 Poa pra 413.2344445444.2.... 001
23 Rum ace ....563......22..... 001
16 Leo aut 52563333.522322222.2 01
20 Poa tri ....645427654549..2. 01
27 Tri rep 3.222526.521233261.. 01
29 Bra rut 4.632622..22224..444 01
 4 Alo gen .........2725385..4. 10
24 Sag pro 2..3.......52242.... 10
25 Sal rep ..33.............5 10
 2 Agr sto ..........4843454475 110
10 Ele pal ............4...4584 11100
21 Pot pal ..............22.. 11100
22 Ran fla ............2..22224 11100
30 Cal cus ..............4.33 11100
14 Jun art ............44...334 11101
 8 Che alb ..............1.... 1111
15 Jun buf ......2......443.... 1111

                00000000000011111111
                00001111111100001111
                    00001111
```

this hierarchical structure only indicates that 8 should be next to 9, 10 next to 11, etc. The groups (e.g. 10 and 11) are ordered 11, 10 if group 11 is more similar to group 4 than group 10 and also less similar to group 3 than group 10. The ordering 10, 11 is better when the reverse holds. In this way, the ordering of the dichotomy is determined by relatively large groups, so that it depends on general relations more than on accidental observations.

After completing the site classification the species are classified by TWINSPAN in the light of the site classification. The species classification is based on fidelity, i.e. the degree to which species are confined to particular groups of sites. In other aspects the classification of the species closely resembles the site classification. A structured table is made from both classifications by ordering the species groups in such a way that an approximate 'positive diagonal' (from upper left to lower

right) is formed. A TWINSPAN table of the Dune Meadow example is given in Table 6.6.

6.4 Non-hierarchical clustering

A non-hierarchical clustering can be constructed by selecting sites to act as an initial point for a cluster and then assigning the other sites to the clusters. The methods vary in details. Gauch (1979) starts picking up a random site and clusters all sites within a specified radius from that site. COMPCLUS, as his program is called (composite clustering), repeats this process until all sites are accounted for. In a second phase, sites from small clusters are reassigned to larger clusters by specifying a larger radius. Janssen (1975) essentially proposes the same approach but picks up the first site from the data as initiating point for the first cluster. This method is applied in CLUSLA (Louppen & van der Maarel 1979). As soon as a site lies further away from the first site than specified by the radius, this site is the initiating point for the next cluster. Subsequent sites are compared to all previously formed clusters. In this way there is a strong dependence on the sequence in which the sites enter the classification. A second step in CLUSLA is introduced to reallocate the sites to the 'best' cluster. This is done by comparing all sites with all clusters: if they are more similar to another cluster then to their parent cluster at that moment, they are placed in the cluster to which they are most similar. In contrast to COMPCLUS, not only within-cluster homogeneity, but also between cluster distances are used by CLUSLA. A method combining the benefits of both methods is used in FLEXCLUS (van Tongeren 1986). From the total set of sites a certain number is selected at random or indicated by the user. All other sites are assigned to the nearest of the set. By relocation until stability is reached, a better clustering is achieved. Outliers are removed by reduction of the radius of the clusters afterwards. Variations of these methods are numerous; others have been presented by, for example, Benzécri (1973), Salton & Wong (1978) and Swain (1978).

Although hierarchical clustering has the advantage over non-hierarchical clustering that between-group relations are expressed in the classification, there is no guarantee that each level of the hierarchy is optimal. A combination of hierarchical and non-hierarchical methods can be made by allowing sites to be relocated, to improve clustering. Since clusters change by relocations, this can be repeated in an iterative process until no further changes occur.

If there is a clear group structure at any level of the hierarchy, no relocations will be made. An example of such a method is relocative centroid sorting. This method is demonstrated in Figure 6.17. Because of the possibility of relocations, a dendrogram cannot be constructed. By using relocative centroid sorting in a slightly different way - assigning each site to a cluster at random or by a sub-optimal, quick method - as shown in Figure 6.17, computing time can be saved because computation of a site-by-site (dis)similarity matrix can be replaced by computation of a site-by-cluster matrix. This is used in the table rearrangement program TABORD (van der Maarel et al. 1978), and also in CLUSLA (Louppen & van der Maarel 1979) and FLEXCLUS (van Tongeren 1986).

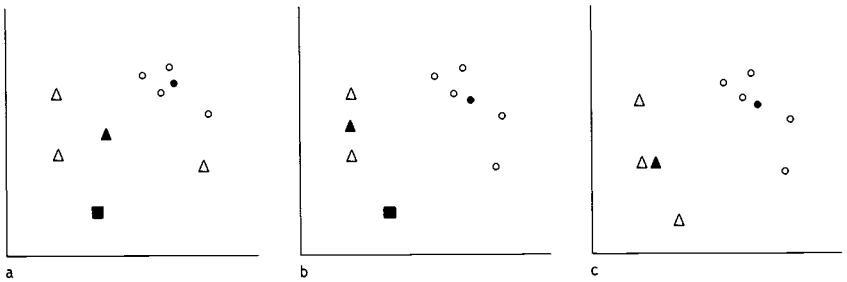


Figure 6.17 Three steps in relocative centroid sorting. a: arbitrary initial clustering. b: after relocation. c: after fusion of the most similar clusters. Samples in the same cluster are indicated by the same open symbol.The corresponding closed symbols indicate the cluster centroids.

6.5 Optimality of a clustering

It is difficult to decide which solution of the cluster analysis to choose. There are very different criteria used to do so: one can distinguish between external and internal criteria.

External criteria are not dependent on the method of clustering. Other data are used to test whether the clustering result is useful.

- In syntaxonomy (Westhoff & van der Maarel 1978), we look for sufficient differences in floristic composition to be able to interpret the results, for instance within the meaning of syntaxa, characteristic and differential species.
- In synecology, if we only used the information on the species composition for our clustering, we have the possibility to test for differences in other variables between the clusters (e.g. analysis of variance for continuous data or chi-square test for nominal variables, cf. Subsection 3.3.1).
- In survey mapping, we can have restrictions on the number of legend units, dependent on scale and technical facilities.

Internal criteria are dependent on the data used for obtaining the clustering and usually also the method of clustering. There are almost as many methods to decide which cluster method or which hierarchical level is best as there are methods for clustering. Some of these methods use statistical tests, but usually it would be better to use the word pseudo-statistics: the conditions for application of the tests are never satisfied because the same characters that are used to group the data are used to test for differences. Most other methods (a review can be found in Popma et al. 1983) use two criteria for the determination of the optimum solution:

- Homogeneity of the clusters (average (dis)similarity of the members of a cluster or some analogue).
- Separation of the clusters (average (dis)similarity of each cluster to its nearest neighbour, or some analogous criterion).

There are many possible definitions of homogeneity and separation of clusters, and each definition might indicate another clustering as the best one. The use of methods to determine the optimum clustering should therefore be restricted to evaluation of subsequent steps in one type of analysis (Hogeweg 1976; Popma et al. 1983).

6.6 Presentation of the results

Results of a classification can be presented in different ways. We have already mentioned:

- the species-by-sites table, giving as much information as possible on all separate sites and species abundances. In vegetation science, additional environmental information and information on the number of species in a site is usually provided in the head of the table.
- The dendrogram, a representation in which the hierarchical structure of the site groups is expressed.

When there are many sites, a species-by-sites table becomes quite large and it is not very easy to interpret the table. This is the reason for the construction of a so-called synoptical table. A synoptical table summarizes the results for each cluster. Classical synoptical tables in the Braun–Blanquet school of vegetation science present a presence class and minimum and maximum values for cover/abundance in each vegetation type for all species. Table 8.3 is an example of such a table. In Table 8.3 the presence classes I to V represent within cluster species frequencies (0-20, 20-40, 40-60, 60-80, 80-100%, respectively). Many other ways of presenting the summarized data in a synoptical table are possible. For example, one can form cross-tabulations of species groups by clusters, or instead of presence class and minimum and maximum scores, average values and standard deviations can be entered into the table.

A dendrogram and a species-by-sites table cannot be used for presentation in more than one dimension. Therefore it can be very useful to present the results of the classification in an ordination diagram of the same set of sites. In such a diagram, more complex relations with the environment can be clearly elucidated (cf. Figure 8.7).

6.7 Relation between community types and environment

The relation between community types as they are defined by cluster analysis on the basis of species data and their environment can be explored in several ways. A hierarchical classification can be evaluated at each level: at each dichotomy, a test can be done for differences between the two groups or all clusters at a certain level are evaluated simultaneously. In a non-hierarchical classification we are restricted to simultaneous comparison of all resulting clusters.The range of methods goes from exploratory analysis (Subsection 6.7.1) to statistical tests of significance (Subsection 6.7.2).

Table 6.7 FLEXCLUS table, similarity ratio, centroid sorting with relocations. Sample 1, which formed a separate cluster, is added to the second cluster by hand. Environmental data are summarized and added to the table.

Sites:				
	11 1 11		11	1112
	79617085	123489	23	4560
Leo aut	26353353	52232	45	4475
Bra rut	3942262	2222	4	444
Tri rep	2532622	52123	32	61
Agr sto		4843	45	4475
Ach mil	2 2 24 2	13		
Ant odo	443 24 4			
Pla lan	2 535335			
Poa pra	1 344432	44544	2	
Lol per	676622	75642		
Bel per	222	322		
Ely rep	4	4444 6		
Alo gen		27253	85	4
Poa tri	4 54 6	276545	49	2
Sag pro	3 2	522	42	
Jun buf	2	4	43	
Jun art		44		4 33
Cal cus				4 33
Ele pal		4		4584
Ran fla		2	2	2224
Air pra	23			
Bro hor	24 2	4 3		
Hyp rad	25 2			
Pot pal				22
Rum ace	6 3 5	2	2	
Sal rep	3 3			5
Tri pra	5 2 2			
Vic lat	2 11			
Che alb			1	
Cir arv		2		
Emp nig	2			
Environmental parameters:				
Dept A1				
mean	4.0	3.8	5.9	7.5
s.d.	1.1	0.6	0.1	3.6
% HF	38	33	0	0
% NM	38	0	0	75
Moisture				
cl 1	*****	*		
cl 2	**	**		
cl 3				
cl 4		*	*	
cl 5	*	*	*	****
Manure				
cl 1	*****	*		
cl 2	**	*		
cl 3	*	*	*	*
cl 4		***		

6.7.1 *The use of simple descriptive statistics*

For a continuous variable, mean and standard deviation can be computed for each cluster, if there is no reason to expect outliers or skewed distributions. For skewed distributions the alternative is to inspect the median and the mid-range, or range for each cluster. Table 6.7 gives means and standard deviations for the depth of the A1 soil horizon. There seems to be a weak relation between the clusters and the depth of the A1 horizon. A good alternative for continuous variables is the construction of stem and leaf diagrams for each cluster separately.

For ordinal and nominal variables, such as moisture class and fertilization class in the Dune Meadow Data, the construction of histograms give us quick insight into the relations. Table 6.7 clearly shows that moisture might be the most important environmental variable affecting the species composition. For nominal variables, frequencies or within cluster proportions of occurrence might also give insight in the relations. The proportion of plots managed by the Department of Forests, the Dutch governmental institution administering the country's nature reserves is, for instance, high in cluster 4, which are the nutrient-poor meadows.

In the preceding subsections, all variables have been evaluated separately and on one hierarchical level. A different, quick way to evaluate all levels of a hierarchical classification in the light of all appropriate environmental variables is the use of simple discriminant functions for each dichotomy as performed by DISCRIM (ter Braak 1982 1986). In the DISCRIM method, simple discriminant functions are constructed in which those environmental variables are selected that optimally predict the classification. Figure 6.18 shows the TWINSPAN classification of the sites (Table 6.6) in the form of a dendrogram and it shows at each branch the most discriminating variables selected by DISCRIM.

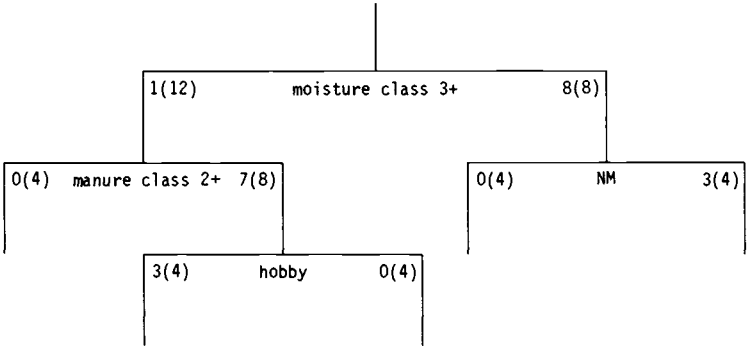


Figure 6.18 This is the same TWINSPAN site classification as in Table 6.6, but now presented as a dendrogram. At each branch the most discriminating variables, selected by DISCRIM, are shown. Numbers at the branches indicate the number of sites for which the conditions are true. Numbers in brackets indicate the number of sites in the clusters.

6.7.2 *The use of tests of significance*

Tests of significance are introduced in Chapter 3, Section 3.2. To test whether any environmental variable might be controlling the species composition (or might be controlled by the species composition, or simply related to the species composition) in the communities, we take as null hypothesis that the species composition is independent from the environmental variable. Rejecting the null hypothesis indicates that the environmental variable is related to the species composition of our community types in some way or another.

Analysis of variance

Analysis of variance is explained in Subsection 3.2.1. It can be used to detect relations between community types and continuous environmental variables. The systematic part consists of the expected values of the environmental variable, one for each community type and the error part is the variation in the values within each community type. Analysis of variance is based on the normal distribution. Therefore environmental variables must often be transformed, e.g. by using the logarithm of their values (see Subsection 2.4.4).

Chi-square test

Subsection 3.3.1 describes the chi-square test for $r \times k$ contingency tables. The chi-square test is used to test the null hypothesis: that a nominal environmental variable is not related to the community types. Care should be taken if the numbers of data are small or if the frequency of occurrence of the nominal variable is low (see Subsection 3.3.1).

The rank sum test for two independent samples

Analysis of variance and the t test (cf. Subsection 3.2.1) are not very resistant to outliers, because they are very much affected by gross errors in the observations. An alternative is to use a distribution-free method, like the rank sum test for two independent samples. As its name indicates, this test, developed by Wilcoxon, but also known as the Mann-Whitney test, can be used to test for differences between two groups. The test is described below.

All observations in both groups are put into a single array in increasing order (indicating also from which group they are) and rank numbers are given to the observations (the smallest observation has rank number 1). Tied observations (equal values) are given their average rank number. For equal sample sizes the smallest sum of rank numbers of both groups is directly referred to a table for the Mann-Whitney test. For unequal sample sizes the sum of the rank numbers is computed for the smallest sample (T_1). A second T (T_2) is computed:

$$T_2 = n_1 (n_1+n_2+1) - T_1$$

where n_1 and n_2 are the sizes of the smaller and the larger sample, respectively.

The test criterion T is the smaller of T_1 and T_2. For sample sizes outside the limits of the table, an approximate normal deviate Z is referred to the tables of the normal distribution to obtain the significance probability P:

$$Z = (|\mu - T| - 0.5)/\sigma$$

where $\mu = (n_1+n_2+1)/2$ and $\sigma = \sqrt{(n_2\,\mu/6)}$.

With this test two small groups (minimum sizes of 4 and 4, 3 and 5 or 2 and 8) can be compared without any assumption on the distribution of the environmental variable. For further details on this test and other tests of significance refer to a statistical handbook (e.g. Snedecor & Cochran 1980).

6.8 Bibliography

Scientific classification of communities can be traced back in history to von Humboldt (1807): he used associations of plants to define community types. Jaccard (1912) took the first step in the direction of multivariate analysis by the development of his index of similarity. Many years later he was followed by Sørensen (1948), who developed his 'method of establishing groups of equal amplitude in plant sociology' based on similarity of species content. The increasing access scientists have had to computers over the last thirty years has led to rapid developments in multivariate methods. An early work that is devoted to the use of multivariate methods in taxonomy is a book written by Sokal & Sneath (1963). In the late sixties, and 1970s there was a rapid increase in the use of cluster analysis (and ordination) by ecologists. Pielou (1969), Goodall (1970) and Williams (1976a) give a theoretical background to these methods.

Numerical classification in the phytosociological context is elucidated by Goodall (1973) and Mueller-Dombois & Ellenberg(1974). Everitt (1980) and Dunn & Everitt (1982) are more recent introductions to numerical taxonomy. Gauch (1982) gives an introduction to classification of communities and mentions many applications of cluster analysis.

6.9 Exercises

Exercise 6.1 Single-linkage clustering with Jaccard similarity

Exercise 6.1a Compute the Jaccard similarities for the sites in the artificial species-by-site table given below. Since the similarities are symmetrical and the diagonal elements all equal 1, you should only compute the elements below the diagonal of the site-by-site similarity matrix (cf. Exercise 6.1c for the species).

Site		1	2	3	4	5	6
Species	A		4	1		2	2
	B				1		
	C	1					
	D	2	1	1		1	
	E		1		4		5
	F		3	1	1		3
	G	1	1	3		1	

Exercise 6.1b Perform single-linkage clustering for the sites.

Exercise 6.1c The species similarities are:

	A	B	C	D	E	F
B	0					
C	0	0				
D	0.60	0	0.25			
E	0.40	0.33	0	0.17		
F	0.60	0.25	0	0.33	0.75	
G	0.60	0	0.25	1.0	0.17	0.33

Perform single-linkage clustering for the species.

Exercise 6.1d Rearrange the sites and the species to represent the hierarchical structure. Try also to obtain the best possible ordering.

Exercise 6.2 Complete-linkage clustering with percentage similarity

Exercise 6.2a Compute the missing elements (denoted by *) of the site-by-site similarity matrix (percentage similarity) for the table of Exercise 6.1a.

	1	2	3	4	5
2	29				
3	40	*			
4	0	25	17		
5	50	58	60	*	
6	0	60	*	63	29

Exercise 6.2b Perform complete-linkage clustering for all sites.

Exercise 6.3 Single-linkage clustering with Euclidean Distance

Exercise 6.3a Compute the missing elements (denoted by *) of the site-by-site Euclidean Distance matrix for the table of Exercise 6.1a.

	1	2	3	4	5
2	5.3				
3	*	4.2			
4	4.9	5.7	5.3		
5	2.5	3.8	2.5	4.9	
6	6.3	4.7	6.3	2.5	*

Exercise 6.3b Perform single linkage and try to find out why the result is so different from Exercise 6.1.

Exercise 6.4 Divisive clustering

This exercise is a demonstration of a simple classification procedure using a divisive strategy with iterative character weighting (this procedure is different from the procedures used in TWINSPAN and by Hogeweg (1976)). The species-by-sites table of Exercise 6.1 is used here too.

Step a Divide the sites in two groups, a positive and a negative one. You may choose a random division or monothetic division based on the presence of one species. In the solution we intially place Sites 1,3 and 6 in the negative group and Sites 5,2 and 4 in the positive group.

Step b Compute the sum of abundances for each species for both sides of the dichotomy (SPOS for the positive scores, SNEG for the negative scores).

Step c Compute a preference score for each species: pref = (SPOS−SNEG)/(SPOS+SNEG).

Step d Compute a weighted sum (or weighted average) of the species abundances for each site.

Step e Find maximum and minimum site score (MAX and MIN) and the midpoint of the site scores (MID = (MAX+MIN)/2).

Step f Assign all sites with a score less than the midpoint to the 'negative' group and all the other sites to the 'positive' group.

Step g Repeat steps b–f until the division is stable.

Step h Repeat steps a–g for each subgroup.

Exercice 6.5 Cluster interpretation with nominal environmental data

Do the chi-square test for moisture classes 1 and 2 combined, and 3, 4, and 5 combined, for:

Exercise 6.5a The first division of TWINSPAN (Table 6.6)

Exercise 6.5b The first three clusters of Table 6.7 combined and the last cluster (highest hierarchical level). What is your null-hypothesis? Has it to be rejected or not? Is it correct to use the chi-square test in this case?

Exercise 6.6. Cluster interpretation with ordinal environmental data

Perform Wilcoxon's test for 'Depth of A1' for the first division of TWINSPAN (Table 6.6).

6.10 Solutions to exercises

Exercise 6.1 Single-linkage clustering with Jaccard similarity

Exercise 6.1a The number of species per site (A) and the values of c and ($A+B-c$) in the notation of Subsection 6.2.2 are given below:

Site number	1	2	3	4	5	6
A (or B)	3	5	4	3	3	3
c 2:	2					
3:	2	4				
4:	0	2	1			
5:	2	3	3	0		
6:	0	3	2	2	1	
$(A+B-c)$ 2:	6					
3:	5	5				
4:	6	6	6			
5:	4	5	4	6		
6:	6	5	6	4	5	

Jaccard similarity is obtained by dividing corresponding elements of the tables: c and $(A+B-c)$

	1	2	3	4	5
Jaccard 2	0.33				
3	0.40	0.80			
4	0.00	0.33	0.17		
5	0.50	0.06	0.75	0.00	
6	0.00	0.60	0.40	0.50	0.20

Exercise 6.1b In order to obtain the single-linkage clustering we only have to find the highest remaining similarity as demonstrated in the following table:

Fusion	Highest similarity	Between sites	Clusters fused
1	0.80	2,3	2,3
2	0.75	3,5	(2,3),5
3	0.60	2,6	(2,3,5),6
4	0.50	1,5	1,(2,3,5,6)
5	0.50	4,6	(1,2,3,5,6),4

Exercise 6.1c

Fusion	Highest similarity	Between species	Clusters fused
1	1.0	D,G	D,G
2	0.75	E,F	E,F
3	0.60	A,D	A,(D,G)
4	0.60	A,F	(A,D,G),(E,F)
5	0.33	B,E	(A,D,E,F,G),B
6	0.25	C,D	(A,B,D,E,F,G),C

Exercise 6.1d The hierarchy can be represented in a dendrogram in which each dichotomy can be swivelled. In order to obtain the best possible ordering each site is placed next to its nearest neighbour.

2 next to 3:	2	3				
5 next to 3:	2	3	5			
6 next to 2:	6	2	3	5		
1 next to 5:	6	2	3	5	1	
4 next to 6:	4	6	2	3	5	1
or the reverse	1	5	3	2	6	4

If we use the same procedure for the species, there is a problem for species C.

D next to G: D G or G D
E next to F: E F or F E
A next to D: A D G or G D A
A next to F: E F A D G or G D A F E
B next to E: B E F A D G or G D A F E B

C between A and D is not a good solution because A resembles D much more than C does.

C next to G is better, but still there are two possibilities, C between D and G or C at the end. Because C resembles A less than D and G do, C is placed at the end of the ordering:

B E F A D G C or C G D A F E B

The rearranged table:

	1	5	3	2	6	4
C	1					
G	1	1	3	1		
D	2	1	1	1		
A		2	1	4	2	
F			1	3	3	1
E				1	5	4
B						1

Exercise 6.2 Complete linkage clustering with percentage similarity

Exercise 6.2a Computation of the similarities:

$PS_{2,3}$: Sum of scores site 2: 10
Sum of scores site 3: 6
Minimum scores: A = 1, B = 0, C = 0, D = 1, E = 0, F = 1, G = 1
Sum of minimum scores: $c = 1+0+0+1+0+1+1 = 4$
$PS_{2,3} = 200\times4/(10+6) = 50$
$PS_{3,6}$: $c = 1+0+0+0+0+1 = 2$
$PS_{3,6} = 200\times2/(6+10) = 25$
$PS_{4,5}$: no common species: $c = 0$ $PS_{4,5} = 0$

The complete similarity matrix is now:

	1	2	3	4	5
2	29				
3	40	50			
4	0	25	17		
5	50	58	60	0	
6	0	60	25	63	29

Exercise 6.2b The first step is the same as with single linkage: the two most similar samples are fused. Then we construct a new similarity matrix:

fusion 1:4 and 6, similarity 63

	1	2	3	(4,6)
2	29			
3	40	50		
(4,6)	0	25')	17")	
5	50	58	60	0"')

Note: ')min (25,60) = 25
")min (17,25) = 17
"')min (0,29) = 0

Fusion 2: 5 and 3, similarity 60

new similarity matrix:

	1	2	(3,5)
2	29		
(3,5)	40')	50")	
(4,6)	0	25	

Note: ') min (40,50) = 40
") min (50,58) = 50

Fusion 3: 2 and (3,5), similarity 50

new similarity matrix:

	1	(2,3,5)
(2,3,5)	29	
(4,6)	0	0

Fusion 4: (2,3,5) and 1, similarity 29

last fusion: (1,2,3,5) and (4,6), similarity 0

Exercise 6.3 Single-linkage clustering with Euclidean Distance

Exercise 6.3a

$$ED_{1,3} = [(0-1)^2+(0-0)^2+(1-0)^2+(2-1)^2+ (0-0)^2 +(0-1)^2+(1-3)^2]^{1/2}$$
$$= (1+0+1+1+0+1+4)^{1/2} = 8^{1/2} = 2.8$$

$$ED_{5,6} = [(2-0)^2+(0-0)^2+(0-5)^2+(0-3)^2+ (1-0)^2]^{1/2} = (4+1+25+9+1)^{1/2}$$
$$= 40^{1/2} = 6.3$$

Exercise 6.3b Instead of looking for the highest similarity values, we look for the lowest dissimilarity values.

Fusion	Dissimilarity	Between sites	Clusters fused
1	2.5	1,5	1,5
2	2.5	3,5	(1,5),3
3	2.5	4,6	4,6
4	4.2	2,3	(1,3,5),2
5	4.7	2,6	(1,2,3,5),(4,6)

Now Sites 4 and 6 group together because of the dominance of Species E. Sites 2 and 3 are more different and fuse therefore later, because of the different abundances for Species A, F and G. Sites 1 and 5 are fused first, because of their low species abundances, and so on.

Exercise 6.4 Divisive clustering

Steps a–g The species-by-sites table (Table 6.8) is rearranged according to the solution from Exercise 6.1.

Table 6.8 Species-by-sites table rearranged according to the solution for Exercise 6.1, with SPOS, SNEG and PREF computed in Steps b and c of the iteration algorithm of Exercise 6.4.

						step 1 b		step 1 c	step 2 b		step 2 c	step 3 b		step 3 c
						SPOS	SNEG	PREF	SPOS	SNEG	PREF	SPOS	SNEG	PREF
1						0	1	-1	0	1	-1	0	1	-1
1	1	3	1			2	4	-0.33	2	4	-0.33	1	5	-0.67
2	1	1	1			2	3	-0.20	2	3	-0.20	1	4	-0.75
	2	1	4	2		6	3	0.33	8	1	0.78	6	3	0.33
		1	3	3	1	4	4	0	7	1	0.75	7	1	0.75
			1	5	4	5	5	0	10	0	1	10	0	1
					1	1	0	1	1	0	1	1	0	1

Step 1a Initial choice: Sites 1, 3 and 6 in the negative group; Sites 2, 4 and 5 in the positive group.
Step 1b–1c See Table 6.8.
Step 1d For Sites 1–6 the weighted sums of the preference scores are −1.73, 0.79, −0.86, 1, 0.13 and 0.66, respectively.
Step 1e MID = (−1.73 + 1)/2 = −0.36.
Step 1f Sites 1 and 3 in the negative group; the other sites in the positive group.
Step 2b–2c See Table 6.8.
Step 2d For Sites 1–6 the weighted sums are −1.73, 5.84, 0.34, 5.75, 1.03 and 8.81, respectively.
Step 2e MID = (−1.73 + 8.81)/2 = 3.56.
Step 2f Sites 1, 3 and 5 in the negative group; the other sites in the positive group.
Step 3b–3c See Table 6.8.
Step 3d For Sites 1–6 the weighted sums are −3.17, 2.15, −1.68, 5.75, −0.76 and 7.91, respectively.
Step 3e MID = (−3.17 + 7.91)/2 = 2.37.
Step 3f Same as Step 2f, so the classification is stable now.

Step h This is solved in essentially the same way for further subdivisions.

Exercise 6.5 Cluster interpretation with nominal environmental data
We start by making two-way cross-tabulations of the observed frequencies (o):

	TWINSPAN	(Table 6.6)		FLEXCLUS	(Table 6.7)	
Cluster number:	0	1	total	1+2+3	4	total
Moisture class: 1+2	11	0	11	11	0	11
3+4+5	1	8	9	5	4	9
Total:	12	8	20	16	4	20

The expected cell frequencies can be obtained by dividing the product of the corresponding row and column totals by the overall total, e.g. (11×12)/20 = 6.6 (first cell of the 'TWINSPAN' table) The two-way cross-tabulation of the expected cell frequencies (*e*) becomes:

Cluster number:	0	1	total	1+2+3	4	total
Moisture class: 1+2	6.6	4.4	11	8.8	2.2	11
3+4+5	5.4	3.6	9	7.2	1.8	9
Total:	12	8	20	16	4	20

For a two-by-two table all deviations from the expected values are equal (check this for yourself): $o-e = 4.4$ and $(o-e)^2 = 19.36$ (TWINSPAN), $o-e = 2.2$ and $(o-e)^2 = 4.84$ (FLEXCLUS, highest level).

$\chi^2 = 19.36(1/6.6+1/4.4+1/5.4+1/3.6) = 16.30$ (TWINSPAN)

$\chi^2 = 4.84(1/8.8+1/2.2+1/7.2+1/1.8) = 6.11$ (FLEXCLUS, highest level).

Our null hypothesis is that the classification is not related to moisture class. We have 1 degree of freedom since the number of rows and the number of columns both are equal to 2 : ν = (number of rows - 1) × (number of columns - 1). Referring to a table of the chi-square distribution we see that the null hypothesis should be rejected ($P < 0.005$ for the TWINSPAN classification and $0.01 < P < 0.025$ for the FLEXCLUS classification), which means that the types are related to moisture level for both classifications.

We should not use the chi-square test because the expected cell frequencies are too low.

Exercise 6.6 Cluster interpretation with ordinal environmental data

The following table shows the sites ordered by increasing depth of the A1 horizon; the sites belonging to the right-hand side of the TWINSPAN dichotomy are indicated with an asterisk. Rank numbers are assigned, in case of ties the average rank number is used. Site 18 is not included in the list since its value for this variable is missing.

Site	7	1	10	2	11	20*	9*	19	17	4	8*	6	3	16*	12*	13*	5	14*	15*
A1	2.8	2.8	3.3	3.5	3.5	3.5	3.7	3.7	4.0	4.2	4.2	4.3	4.3	5.7	5.8	6.0	6.3	9.3	12.
Rank	1.5	1.5	3	5	5	5	7.5	7.5	9	10.	10.	12.	12.	14	15	16	17	18	19

Rank numbers 10. and 12. indicate 10.5 and 12.5 respectively. Value of T_1 (for the smaller, right-hand-side group) is the sum of the rank numbers: 105. Value of $T_2 = 8(8+11+1)-105 = 55$. T, which is referred to a table of Wilcoxon's test, is the smaller of these two: 55 Looking up this value in the table, we conclude that we cannot reject our null hypothesis. There is no evidence that the types are related to the depth of the A1 horizon.