Michael Greenacre
Raul Primicerio
Multivariate Analysis of Ecological Data

Fundación BBVA

Multivariate Analysis of Ecological Data

Multivariate Analysis of Ecological Data

Michael Greenacre
Raul Primicerio

Fundación BBVA

First published December 2013

Plaza de San Nicolás, 4. 48005 Bilbao
www.fbbva.es
publicaciones@fbbva.es

Supporting website www.multivariatestatistics.org provide the data sets used in the book, **R** scripts for the computations, glossary of terms, and additional material.

A digital copy of this book can be downloaded free of charge at www.fbbva.es.

Edition and production: Rubes Editorial

ISBN: 978-84-92937-50-9
Legal deposit no.: BI-1622-2013

Printed in Spain

Printed by Comgrafic on 100% recycled paper, manufactured to the most stringent European environmental standards.

To Zerrin and Maria

Contents

INTERPRETATION, INFERENCE AND MODELLING

CASE STUDIES

APPENDICES

Preface

The world around us – and especially the biological world – is inherently multidimensional. Biological diversity is the product of the interaction between many species, be they marine, plant or animal life, and of the many limiting factors that characterize the environment in which the species live. The environment itself is a complex mix of natural and man-induced parameters: for example, meteorological parameters such as temperature or rainfall, physical parameters such as soil composition or sea depth, and chemical parameters such as level of carbon dioxide or heavy metal pollution.

The properties and patterns we focus on in ecology and environmental biology consist of these many covarying components. Evolutionary ecology has shown that phenotypic traits, with their functional implications, tend to covary due to correlational selection and trade-offs. Community ecology has uncovered gradients in community composition. Much of this biological variation is organized along axes of environmental heterogeneity, consisting of several correlated physical and chemical characteristics. Spectra of functional traits, ecological and environmental gradients, all imply correlated properties that will respond collectively to natural and human perturbations. Small wonder that scientific inference in these fields must rely on statistical tools that help discern structure in datasets with many variables (i.e., multivariate data). These methods are comprehensively referred to as multivariate analysis, or multivariate statistics, the topic of this book. Multivariate analysis uses relationships between variables to *order* the objects of study according to their collective properties, that is to highlight spectra and gradients, and to *classify* the objects of study, that is to group species or ecosystems in distinct classes each containing entities with similar properties.

Although multivariate analysis is widely applied in ecology and environmental biology, also thanks to statistical software that makes the variety of methods more accessible, its concepts, potentials and limitations are not always transparent to practitioners. A scattered methodological literature, heterogeneous terminology, and paucity of introductory texts sufficiently comprehensive to provide a methodological overview and synthesis, are partly responsible for this. Another reason is the fact that biologists receive a formal quantitative training often limited to univariate, parametric statistics (regression and analysis of variance), with some

exposure to statistical modelling. In order to provide a training opportunity that could compensate for this, we collaborated on an introductory, intensive workshop in multivariate analysis of ecological data, generously supported and hosted several times by the BBVA Foundation in Madrid, Spain. The material for the workshop, consisting of lectures and practical sessions (R being our choice of software for the daily practicals) developed out of a graduate and postgraduate course at the University of Tromsø, Norway, now in its tenth year. Further intensive courses for professional ecologists and biologists were given at research institutions and universities in Iceland, Norway, United Kingdom, Italy and South Africa.

The aim of the material, developed for the various teaching and training purposes, refined, expanded and organized in this book, was always to provide the practitioner with the necessary tools to (i) choose the appropriate method for a given problem and data, (ii) implement the method correctly with the help of a computer, (iii) interpret the results with regard to the question asked, and (iv) clearly communicate the results and interpretation with the help of graphical illustrations. The last point about the importance of publishing quantitative results has been an emphasis of ours. As the ecologist Robert MacArthur has put it, "you have a choice, you can either keep up with the literature or you can contribute to it". For your scientific contribution to be effective, quantitative results and their interpretation must be presented in an understandable and accessible way.

The book, aimed at graduate and post-graduate students and professional biologists, is organized in a series of topics, labelled as parts consisting of multiple chapters, reflecting the sequence of lectures of our courses. The background for understanding multivariate methods and their applications is presented in the first introductory part, summarizing the character of ecological data and reviewing multivariate methods. The second part defines the basic concepts of distance and correlation measures for multivariate data, measuring inter-sample and inter-variable relationships. Initial approaches to analysing multivariate data are given in the third part, in the form of clustering and multidimensional scaling, both of which visualize these relationships in a fairly simple way. The fourth part introduces the core concept of the biplot, which explains how a complete data set can be explored using well-known ideas of linear regression and geometry, leading up to the method of principal component analysis. The fifth part is devoted to correspondence analysis and the related method of log-ratio analysis, and ending with canonical correspondence analysis, one of the key methodologies in ecology, which attempts to relate multivariate biological responses to multivariate environmental predictors. The sixth part is dedicated to aids to interpretation of results, statistical inference, and modelling, including an introduction to permutation testing and bootstrapping for the difficult problem of hypothesis testing in the multivariate context. Throughout the book the methods are illustrated using

small to medium-sized data sets. The seventh and last part of the main text of the book consists of two case studies that apply the above multivariate techniques to larger data sets, where the reader can see the challenge for analysis, interpretation and communication when dealing with large studies and complex designs. Finally, three appendices are provided on theoretical, bibliographical and computational aspects. All the analyses presented in the book can be replicated using the R scripts and data sets that are available on the website www.multivariatestatistics.org.

All the above topics contain material accessible to graduate students and practitioners with a basic statistical training. To make the material accessible we relied on the more visual (geometric) and intuitive aspects of the subjects. But chapters, and sections within chapters, do vary with regard to technicality and difficulty. A suggestion for the reader unfamiliar with multivariate methods is to first focus on the more general, accessible sections of the book, respecting the suggested sequence of parts and chapters, and wait before dwelling into the deeper, more technical layers of explanation upon a second reading. With some basic exposure to multivariate methods, the text can also be used as a handbook, with individual core chapters covering the rationales, strengths and weaknesses of various standard methods of interest, and providing illustrations (based on case studies) of appropriate visualization and pertinent interpretation of results.

Our most sincere gratitude goes to the colleagues and institutions that have hosted and helped organize our workshops and courses. First and foremost, we thank the BBVA Foundation and its director, Prof. Rafael Pardo, for continual support and assistance from their dedicated and friendly staff, who are now helping us further to publish the book that summarizes it all. A special thanks goes to the University of Tromsø, which has helped us maintain our course over a prolonged period. Many other institutions have provided help during the planning and running of our intensive courses. The Marine Research Institutes of Iceland and Norway, the University of Lancaster, UK, the Universities of Stellenbosch and Potchefstroom in South Africa, the Universities of Parma and Ancona, Italy, and the Italian Ecological Society.

Many colleagues have helped, directly or indirectly, with the preparation of this book: Giampaolo Rossetti, for his hospitality and support in Parma; Michaela Aschan, Maria Fossheim, Magnus Wiedmann, Grégoire Certain, Benjamin Planque, Andrej Dolgov, Edda Johannesen and Lis Lindal Jørgensen, our co-workers on the Barents Sea Ecosystem Resilience project *(BarEcoRe)* of the Norwegian Research Council; Paul Renaud, Sabine Cochrane, Michael Carroll, Reinhold Fieler and Salve Dahle, colleagues from Akvaplan-niva in Tromsø; Janne Søreide, Eva Leu, Anette Wold and Stig Falk-Petersen, for interesting discussions on fatty acid compositional data; and our families, for their patience and support.

Beginning with our first course at the University of Tromsø in 2004 and our first workshop at the BBVA Foundation in 2005, which developed into courses and workshops in six countries, we have had the privilege of a continuous exposure to insightful comments and friendly exchange with over 500 attendants sharing our passion for science and the environment. To all of them go our sincere gratitude and hope for a long and rewarding career. To us remains the intense satisfaction of interacting with motivated and enthusiastic people. We hope for more, thanks to this book.

Michael Greenacre
Raul Primicerio
Barcelona and Tromsø
November 2013

ECOLOGICAL DATA AND MULTIVARIATE METHODS

CHAPTER 1

Multivariate Data in Environmental Science

In this introductory chapter we take a simple *univariate* or *bivariate* view of multivariate data, using a small educational example taken from marine biology. This means we will not venture beyond studying one or two variables at a time, using graphical displays as much as possible. Often we will show many of these representations simultaneously, which facilitates comparison and interpretation. The descriptive graphical methods that we use here – histograms, bar-charts and box-and-whisker plots – are well-known in any basic statistical course, and are invaluable starting points to what we could call a "marginal" understanding of our data before embarking on multivariate analysis. We encourage researchers to make as many graphical displays as possible of their data, to become acquainted with each variable, and to be aware of problems at an early stage, such as incorrect or very unusual values, or unusual distributions.

Contents

Data set "bioenv": Introductory data set from marine biology

As a simple introductory example to motivate and illustrate the concepts and methods explained in this book, consider the data in Exhibit 1.1. These are biological and environmental observations made at 30 sampling points, or *sites*, on the sea-bed. Typically, a number of grabs (for example, five) are made close by at each site and then a fixed volume of each grab is sent to a biological laboratory for analysis. The more grabs one takes at a site, the more species are eventually identified. The biological data consist of species

Exhibit 1.1: *Typical set of multivariate biological and environmental data: the species data are counts, whereas the environmental data are continuous measurements, with each variable on a different scale; the last variable is a categorical variable classifying the sediment of the sample as mainly* C *(= clay/silt),* S *(= sand) or* G *(= gravel/stone)*

Site No.	Species counts					Environmental variables			
	a	*b*	*c*	*d*	*e*	*Depth (x)*	*Pollution (y)*	*Temperature (z)*	*Sediment (s)*
s1	0	2	9	14	2	72	4.8	3.5	S
s2	26	4	13	11	0	75	2.8	2.5	C
s3	0	10	9	8	0	59	5.4	2.7	C
s4	0	0	15	3	0	64	8.2	2.9	S
s5	13	5	3	10	7	61	3.9	3.1	C
s6	31	21	13	16	5	94	2.6	3.5	G
s7	9	6	0	11	2	53	4.6	2.9	S
s8	2	0	0	0	1	61	5.1	3.3	C
s9	17	7	10	14	6	68	3.9	3.4	C
s10	0	5	26	9	0	69	10.0	3.0	S
s11	0	8	8	6	7	57	6.5	3.3	C
s12	14	11	13	15	0	84	3.8	3.1	S
s13	0	0	19	0	6	53	9.4	3.0	S
s14	13	0	0	9	0	83	4.7	2.5	C
s15	4	0	10	12	0	100	6.7	2.8	C
s16	42	20	0	3	6	84	2.8	3.0	G
s17	4	0	0	0	0	96	6.4	3.1	C
s18	21	15	33	20	0	74	4.4	2.8	G
s19	2	5	12	16	3	79	3.1	3.6	S
s20	0	10	14	9	0	73	5.6	3.0	S
s21	8	0	0	4	6	59	4.3	3.4	C
s22	35	10	0	9	17	54	1.9	2.8	S
s23	6	7	1	17	10	95	2.4	2.9	G
s24	18	12	20	7	0	64	4.3	3.0	C
s25	32	26	0	23	0	97	2.0	3.0	G
s26	32	21	0	10	2	78	2.5	3.4	S
s27	24	17	0	25	6	85	2.1	3.0	G
s28	16	3	12	20	2	92	3.4	3.3	G
s29	11	0	7	8	0	51	6.0	3.0	S
s30	24	37	5	18	1	99	1.9	2.9	G

abundances obtained by summing the counts of the species identified in the grabs for each site.

Usually there are dozens or hundreds of species found in an ecological study. Exhibit 1.1 is intentionally a small data set with only five species, labelled *a* to e. The number of sites, 30 in this case, is more realistic, because there are usually few

sampling locations in marine environmental sampling. As well as the biological data, several environmental variables are typically available that characterize the sites. As examples of these we give four variables, three measurements and one classification – Exhibit 1.1 shows the values of depth x (in metres), a pollution index y, the temperature z in °C and the sediment type (three categories). The *pollution index* is based on data for heavy metal concentrations such as barium, cadmium and lead, as measured in the sea-bed samples – the higher the index, the higher is the overall level of pollution. The last column gives the classification of the sediment in the sample as clay/silt (C), sand (S) or gravel/stone (G). In this chapter we look at well-known univariate and bivariate summaries of these data, before we move on to a multivariate treatment.

Continuous variables

The three variables pollution, depth and temperature are called *continuous variables* because they can – theoretically, at least – have any value on a continuous scale. To take a look at the range of values as well as the shape of the distribution of a continuous variable, we typically plot a *histogram* for each variable – see the first three histograms in Exhibit 1.2. A histogram divides the range of a continuous variable into intervals, counts how many observations are in each interval, and then plots these frequencies. Pollution and temperature are seen to have single-peaked distributions, while the distribution of depth seems more uniform across its range.

Categorical variables

The sediment variable is a *categorical* (or *discrete*) variable because it can take only a few "values", which in this case are sediment types – see the *bar-chart* form of its distribution in Exhibit 1.2. A bar-chart simply counts how many observations correspond to each given category – this is why the bar-chart shows spaces between the categories, whereas in the histogram the frequency bars touch one another. The bar-chart shows a fairly even distribution, with gravel/rock (G) being the least frequent sediment category.

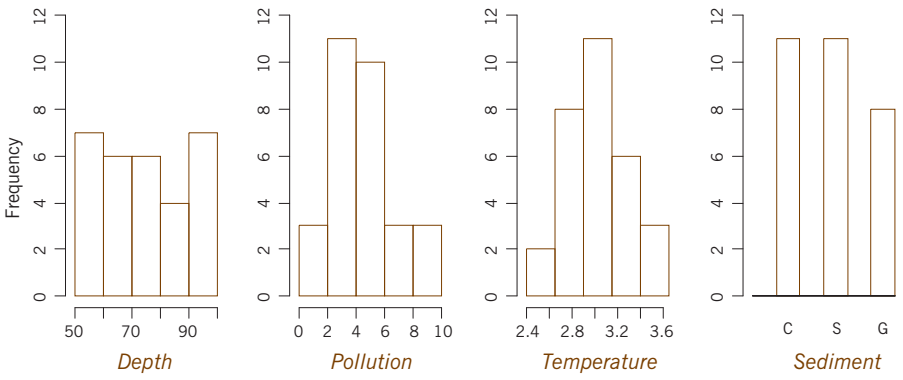


Exhibit 1.2: *Histograms of three environmental variables and bar-chart of the categorical variable*

Categorical variables are either *ordinal* or *nominal* depending on whether the categories can be ordered or not. In our case, the categories could be considered ordered in terms of granularity of the sediment, from finest (clay/silt) to coarsest (gravel/rock). An example of a nominal variable, where categories have no inherent ordering, might be "sampling vessel" (if more than one ship was used to do the sampling) or "region of sampling" (if sampling was done in more than one region). Often, continuous variables are categorized into intervals (i.e., discretized), giving an ordinal variable; for example, "depth" could be categorized into several categories of depth, from shallow to deep.

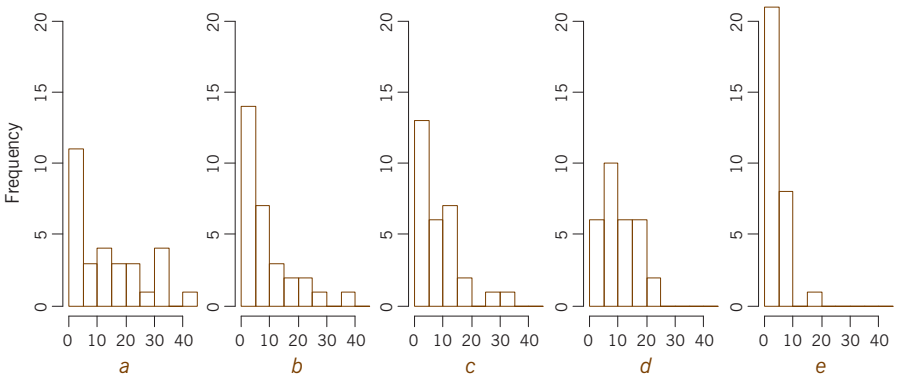
Count variables

The biological data are measured on a different scale from the others – these are *counts*, and have an integer scale: 0, 1, 2, and so on. Counts have a special place in statistics, lying somewhere between a continuous variable and a categorical variable. For the moment, however, we shall treat these data as if they were continuous variables; later we shall discuss various special ways to analyse them. Exhibit 1.3 shows histograms of the five species, with highly skew distributions owing to the many zeros typically found in such species abundance data.

Relationships amongst the environmental variables

The usual way to investigate the relationships between two continuous variables is to make a scatterplot of their relationship. The scatterplots of the three pairs of variables are shown in Exhibit 1.4, as well as the numerical value of their correlation coefficients. The only statistically significant correlation is between depth and pollution, a negative correlation of −0.396 ($p = 0.0305$, using the two-tailed t-test[1]).

Exhibit 1.3: *Histograms of the five species, showing the usual high frequencies of low values that are mostly zeros, especially in species* e



[1] Note that the t-test is not the correct test to use on such nonnormal data. An alternative is the distribution-free permutation test, which gives an estimated p-value of 0.0315, very close to the 0.0305 of the t-test. The permutation test for a correlation is described in Chapter 6, with a full treatment of permutation testing in Chapter 17.

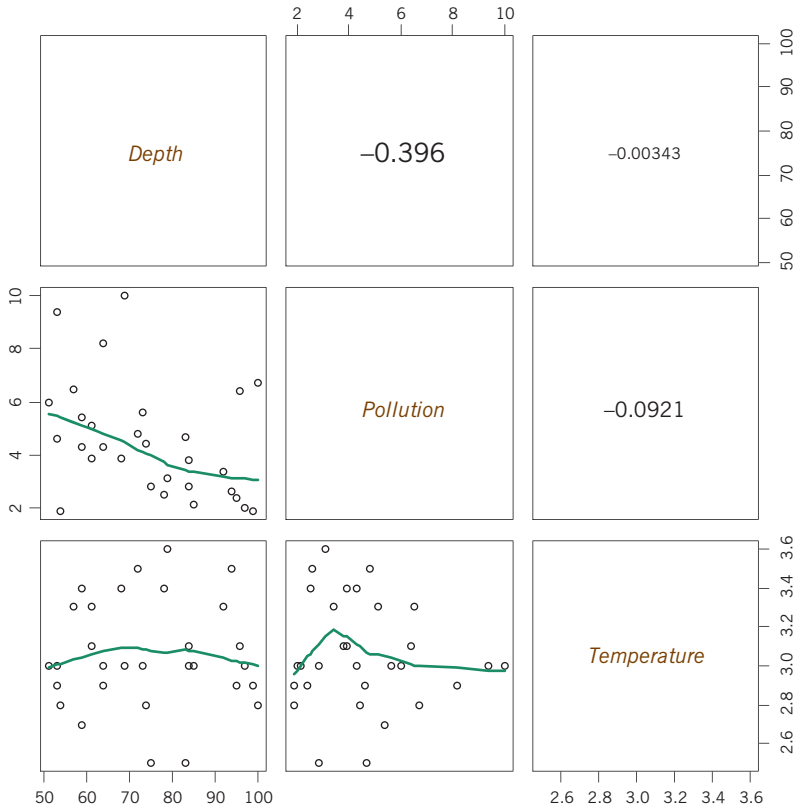


Exhibit 1.4: *Pairwise scatterplots of the three continuous variables in the lower triangle, showing smooth relationships (in brown, a type of moving average) of the vertical variable with respect to the horizontal one; for example, at the intersection of depth an pollution, pollution defines the vertical ("y") axis and depth the horizontal ("x") one. The upper triangle gives the correlation coefficients, with size of numbers proportional to their absolute values*

To show the relationship between the continuous environmental variables and the categorical one (sediment), the *box-and-whisker* plots in Exhibit 1.5 compare the distributions of each continuous variable within each category. The boxes are drawn between the lower and upper quartiles of the distribution, hence the central 50% of the data values lie in the box. The median value is indicated by a line inside the box, while the whiskers extend to the minimum and maximum values in each case. These displays show differences between the gravel samples (G) and the other samples for the depth and pollution variables, but no differences amongst the sediment types with respect to temperature. A correlation can be calculated if the categorical variable has only two categories, i.e., if it is *dichotomous*. For example, if clay and sand are coded as 0 and gravel as 1, then the correlations[2] between the three variables depth, pollution and temperature, and

[2] This correlation between a continuous variable and a dichotomous categorical variable is called the *point biserial* correlation. Based on permutation testing (see Chapters 6 and 17), the *p*-values associated with these correlations are estimated as 0.0011, 0.0044 and 0.945 respectively.

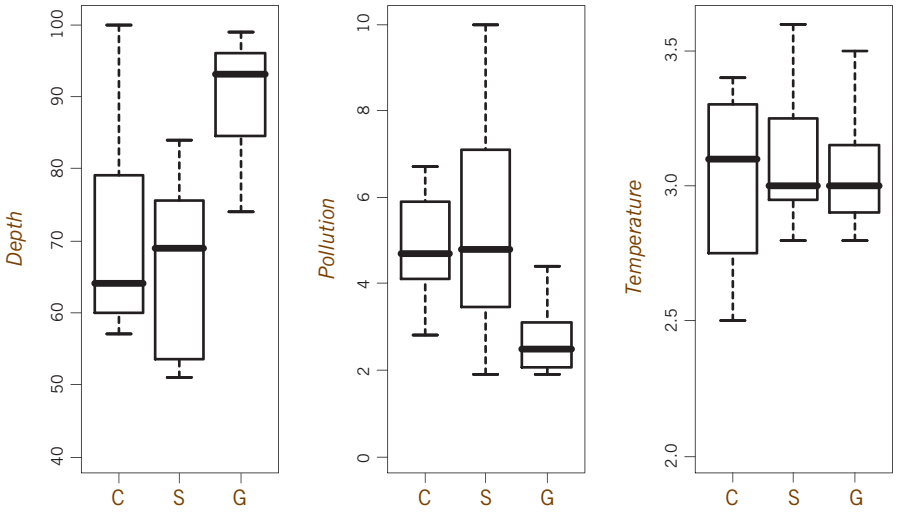


Exhibit 1.5: *Box-and-whisker plots showing the distribution of each continuous environmental variable within each of the three categories of sediment (C = clay/silt, S = sand, G = gravel/stone). In each case the central horizontal line is the median of the distribution, the boxes extend to the first and third quartiles, and the dashed lines extend to the minimum and maximum values*

this dichotomous sediment variable, are 0.611, –0.520 and –0.015 respectively, confirming our visual interpretation of Exhibit 1.5.

Relationships amongst the species abundances

Similar to Exhibit 1.4 the pairwise relationships between the species abundances can be shown in a matrix of scatterplots (Exhibit 1.6), giving the correlation coefficient for each pair. Species *a*, *b* and *d* have positive inter-correlations, whereas *c* tends to be correlated negatively with the others. Species *e* does not have a consistent pattern of relationship with the others.

Relationships between the species and the continuous environmental variables

Again, using scatterplots, we can make a first inspection of these relationships by looking at each species-environmental variable pair in a scatterplot. The simplest way of modelling the relationship, although perhaps not the most appropriate way (see Chapter 18), is by a linear regression, shown in each mini-plot of Exhibit 1.7. The coefficient of determination R^2 (variance explained by the regression) is given in each case, which for simple linear regression is just the square of the correlation coefficient. The critical point for a 5% significance level, with $n = 30$ observations, is $R^2 = 0.121$ ($|R| = 0.348$); but because there are 15 regressions we should reduce the significance level accordingly. A conservative way of doing this is to divide the significance level by the number of tests, in which case the R^2 for significance is 0.236 ($|R| = 0.486$).[3] This would lead to the conclusion

[3] This is known as the *Bonferroni correction*. If many tests are performed, then the chance of finding a significant result by chance increases; that is, there is higher probability of committing a "type I" error. If the significance level is α and there are M tests, the Bonferroni correction is to divide α by M, then use the α/M significance level for the tests. This is a conservative strategy because the tests are usually not independent, so the correction overcompensates for the problem. But in any case, it is good to be conservative, at least in statistics!

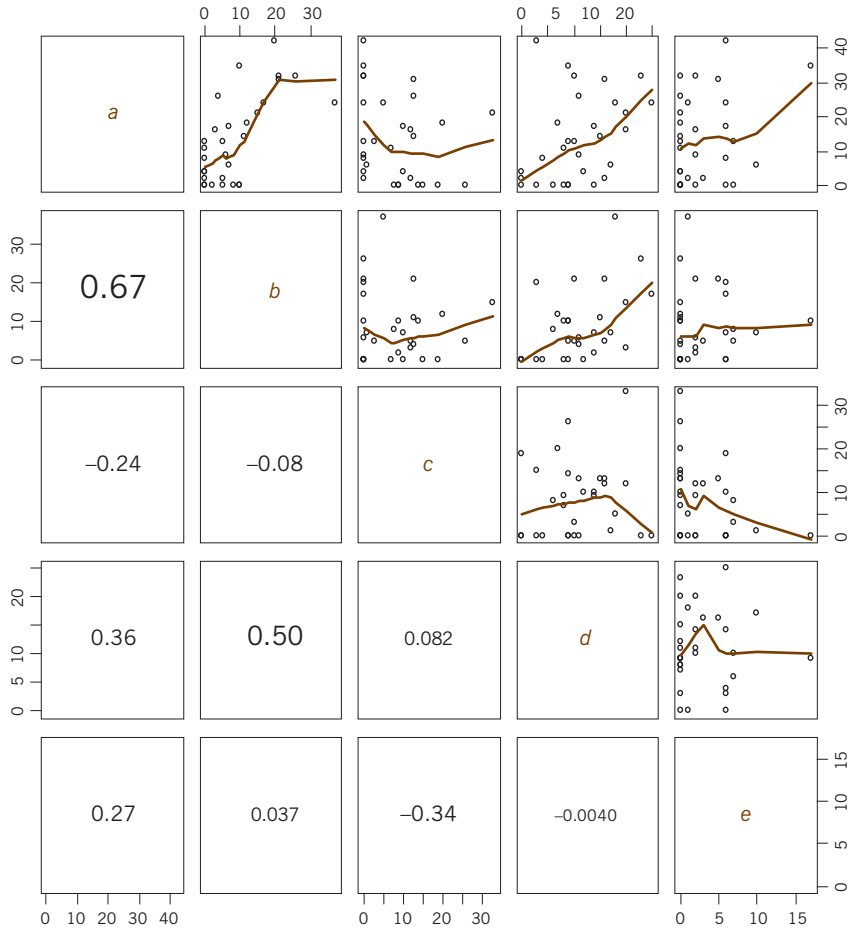


Exhibit 1.6: *Pairwise scatterplots of the five species abundances, showing in each case the smooth relationship of the vertical variable with respect to the horizontal one; the lower triangle gives the correlation coefficients, with size of numbers proportional to their absolute values*

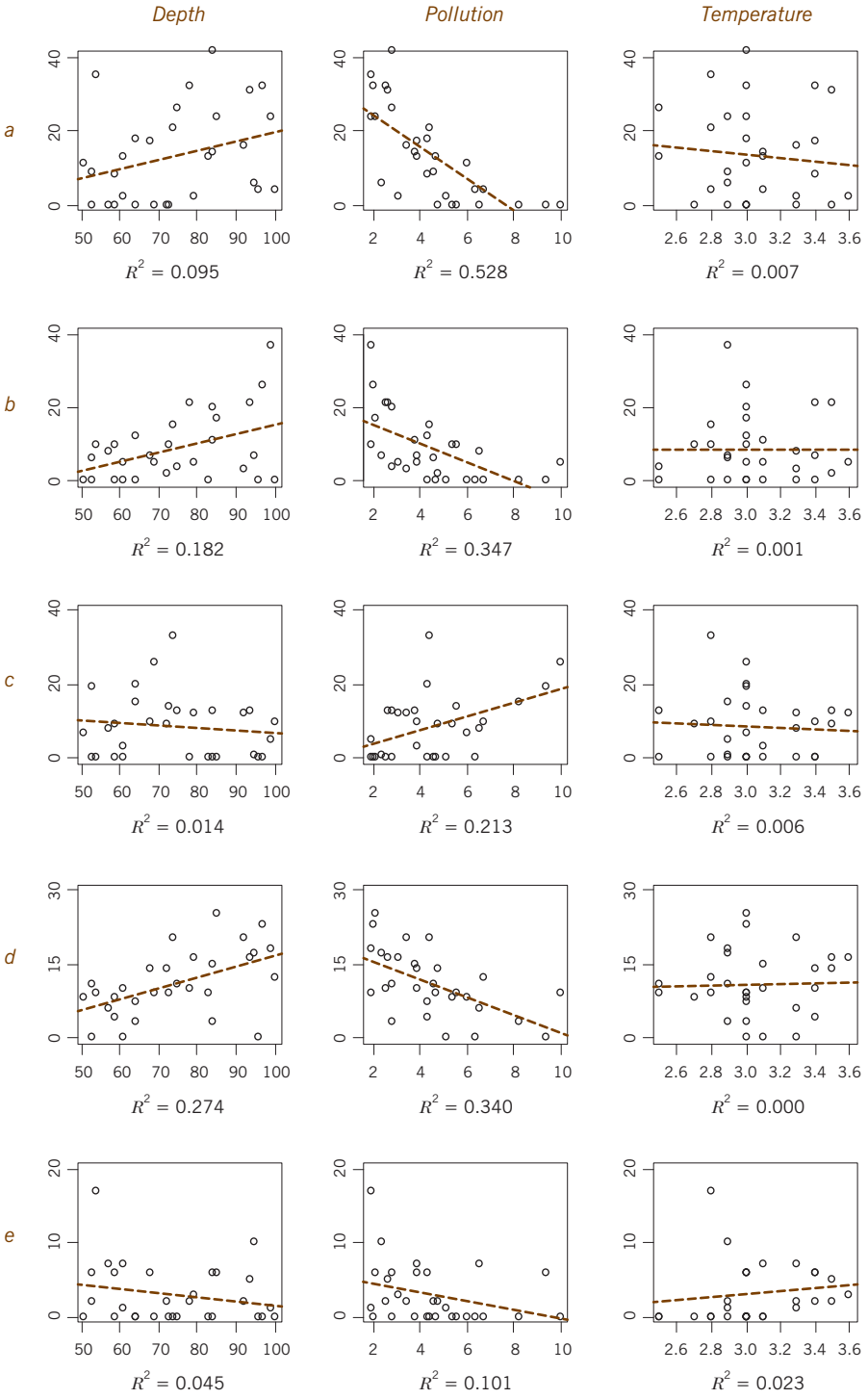
that *a*, *b* and *d* are significantly correlated with pollution, and that *d* is also significantly correlated with depth.

Relationships between the species and the categorical environmental variables

To show how the species abundances vary across the sediment groups, we use the same boxplot style of analysis as Exhibit 1.5, shown in Exhibit 1.8. The statistic that can be used to assess the relationship is the *F*-statistic from the corresponding analysis of variance (ANOVA). For this size data set and testing across three groups, the 5% significance point of the *F* distribution[4] is $F = 3.34$. Thus, species groups *a*, *b* and *d* show apparent significant differences between the sediment types, with the abundances generally showing an increase in gravel/stone.

[4] Note that using the *F*-distribution is not the appropriate way to test differences between count data, but we use it anyway here as an illustration of the *F* test.

Exhibit 1.7: *Pairwise scatterplots of the five groups of species with the three continuous environmental variables, showing the simple least-squares regression lines and coefficients of determination (R^2)*



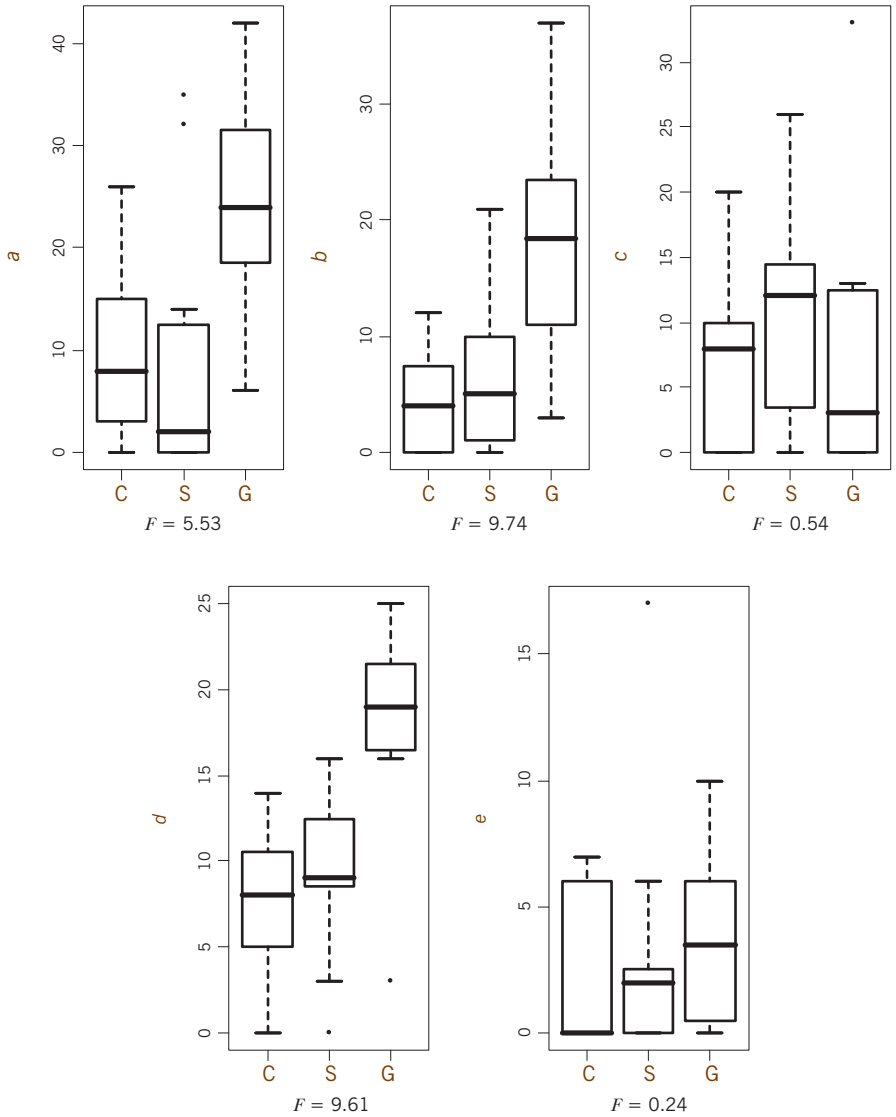


Exhibit 1.8: *Box-and-whisker plots showing the distribution of each count variable across the three sediment types* (C = *clay/silt,* S = *sand,* G = *gravel/stone) and the F-statistics of the respective ANOVAs*

SUMMARY: Multivariate data in environmental science

1. Large numbers of variables are typically collected in environmental research: it is not unusual to have more than 100 species, and more than 10 environmental variables.

2. The scale of the variables is either continuous, categorical or in the form of counts.

3. For the moment we treat counts and continuous data in the same way, whereas categorical data are distinct in that they usually have very few values.

4. The categorical data values do not have any numerical meaning, but they might have an inherent order, in which case they are called *ordinal*. If not, they are called *nominal*.

5. The univariate distributions of count and continuous variables are summarized in histograms, whereas those of categorical variables are summarized in bar-charts.

6. The bivariate distributions of continuous and count variables are summarized in typical "*x*-*y*" scatterplots. Numerically, the relationship between a pair of variables can be summarized by the correlation coefficient.

7. The relationship between a continuous and a categorical variable can be summarized using box-and-whisker plots side by side, one for each category of the categorical variable. The usual correlation coefficient can be calculated between a continuous variable and a dichotomous categorical variable (i.e., with only two categories).

CHAPTER 2

The Four Corners of Multivariate Analysis

Multivariate analysis is a wide and diverse field in modern statistics. In this chapter we shall give an overview of all the multivariate methods encountered in ecology. Most textbooks merely list the methods, whereas our approach is to structure the whole area in terms of the principal objective of the methods, divided into two main types – functional methods and structural methods. Each of these types is subdivided into two types again, depending on whether the variable or variables of main interest are continuous or categorical. This gives four basic classes of methods, which we call the "four corners" of multivariate analysis, and all multivariate methods can be classified into one of these corners. Some methodologies, especially more recently developed ones that are formulated more generally, are of a hybrid nature in that they lie in two or more corners of this general scheme.

Contents

The basic data structure: a rectangular data matrix

In multivariate statistics the basic structure of the data is in the form of a cases-by-variables rectangular table. This is also the usual way that data are physically stored in a computer file, be it a text file or a spreadsheet, with cases as rows and variables as columns. In some particular contexts there are very many more variables than cases and for practical purposes the variables are defined as rows of the matrix: in genetics, for example, there can be thousands of genes observed on a few samples, and in community ecology species (in their hundreds) can be listed as rows and the samples (less than a hundred) as columns of the data

table. By convention, however, we will always assume that the rows are the cases or sampling units of the study (for example, sampling locations, individual animals or plants, laboratory samples), while the columns are the variables (for example: species, chemical compounds, environmental parameters, morphometric measurements).

Variables in a data matrix can be on the same measurement scale or not (in the next chapter we treat measurement scales in more detail). For example, the matrix might consist entirely of species abundances, in which case we say that we have "same-scale" data: all data in this case are counts. A matrix of morphometric measurements, all in millimetres, is also same-scale. Often we have a data matrix with variables on two or more different measurement scales – the data are then called "mixed-scale". For example, on a sample of fish we might have the composition of each one's stomach contents, expressed as a set of percentages, along with morphometric measurements in millimetres and categorical classifications such as sex (male or female) and habitat (littoral or pelagic).

Functional and structural methods

We distinguish two main classes of data matrix, shown schematically in Exhibit 2.1. On the left is a data matrix where one of the observed variables is separated from the rest because it has a special role in the study – it is often called a *response variable*. By convention we denote the response data by the column vector **y**, while data on the other variables – called *predictor,* or *explanatory, variables* – are gathered in a matrix **X**. We could have several response variables, gathered in a matrix **Y**. On the right is a different situation, a data matrix **Y** of several response variables to be studied together, with no set of explanatory variables. For this case we have indicated by a dashed box the existence of an unobserved variable

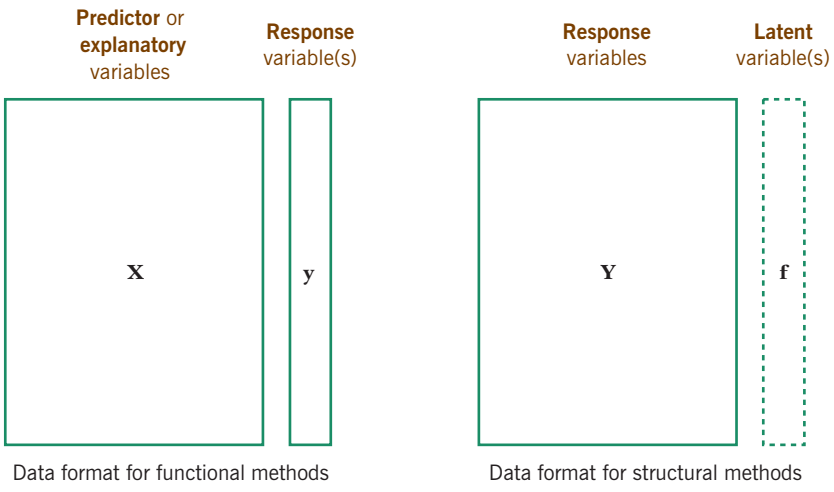


Exhibit 2.1: *Schematic diagram of the two main types of situations in multivariate analysis: on the left, a data matrix where a variable* **y** *is singled out as being a response variable and can be partially explained in terms of the variables in* **X**. *On the right, a data matrix* **Y** *with a set of response variables but no observed predictors, where* **Y** *is regarded as being explained by an unobserved, latent variable* **f**

f, called a *latent variable*, which we assume has been responsible, at least partially, for generating the data **Y** that we have observed. The vector **f** could also consist of several variables and thus be a matrix **F** of latent variables.

We call the multivariate methods which treat the left hand situation *functional methods*, because they aim to come up with a model which relates the response variable **y** as a function of the explanatory variables **X**. As we shall see in the course of this book, the nature of this model need not be a mathematical formula, often referred to as a *parametric* model, but could also be a more general *nonparametric* concept such as a tree or a set of smooth functions (these terms will be explained more fully later). The methods which treat the right hand situation of Exhibit 2.1 are called *structural methods*, because they look for a structure underlying the data matrix **Y**. This latent structure can be of different forms, for example gradients or typologies.

One major distinction within each of the classes of functional and structural methods will be whether the response variable (for functional methods) or the latent variable (for structural methods) is of a continuous or a categorical nature. This leads us to a subdivision within each class, and thus to what we call the "four corners" of multivariate analysis.

The four corners of multivariate analysis

Exhibit 2.2 shows a vertical division between functional and structural methods and a horizontal division between continuous and discrete variables of interest, where "of interest" refers to the response variable(s) **y** (or **Y**) for functional methods and the latent variable(s) **f** (or **F**) for the structural methods (see Exhibit 2.1). The four quadrants of this scheme contain classes of methods, which we shall treat one at a time, starting from regression at top left and then moving clockwise.

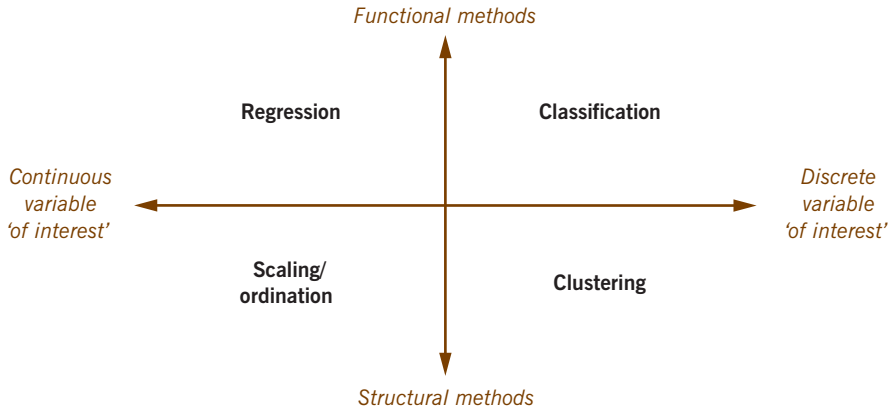


Exhibit 2.2:
The four corners of multivariate analysis. Vertically, functional and structural methods are distinguished. Horizontally, continuous and discrete variables of interest are contrasted: the response variable(s) in the case of functional methods, and the latent variable(s) in the case of structural methods

Regression: Functional methods explaining a given continuous variable

At top left we have what is probably the widest and most prolific area of statistics, generically called *regression*. In fact, some practitioners operate almost entirely in this area and hardly move beyond it. This class of methods attempts to use multivariate data to explain one or more continuous response variables. In our context, an example of a response variable might be the abundance of a particular plant species, in which case the explanatory variables could be environmental characteristics such as soil texture, pH level, altitude and whether the sampled area is in direct sunshine or not. Notice that the explanatory variables can be of any type, continuous or categorical, whereas it is the continuous nature of the response variable – in this example, species abundance – that implies that the appropriate methodology is of a regression nature.

In this class of regression methods are included multiple linear regression, analysis of variance, the general linear model and regression trees. Regression methods can have either or both of the following purposes: to explain relationships and/or to predict the response from new observations of the explanatory variables. For example, on the one hand, a botanist can use regression to study and quantify the relationship between a plant species and several variables that are believed to influence the abundance of the plant. But on the other hand, the objective could also be to ask "what if?" type questions: what if the rainfall decreased by so much % and the pH level rose to such-and-such a value, what would the average abundance of the species be, and how accurate is this prediction?

Classification: Functional methods explaining a given categorical variable

Moving to the top right corner of Exhibit 2.2, we have the class of methods analogous to regression but with the crucial difference that the response variable is not continuous but categorical. That is, we are attempting to model and predict a variable that takes on a small number of discrete values, not necessarily in any order. This area of *classification* methodology is omnipresent in the biological and life sciences: given a set of observations measured on a person just admitted to hospital having experienced a heart attack – age, body mass index, pulse, blood pressure and glucose level, having diabetes or not, etc. – can we predict whether the patient will survive in the next 24 hours? Having found a fossil skull at an archeological site, and made several morphometric measurements, can we say whether it is a human skull or not, and with what certainty?

The above questions are all phrased in terms of predicting categories, but our investigation would also include trying to understand the relationship between a categorical variable, with categories such as "healthy" and "sick", and a host of variables that are measured on each individual. Especially, we would like to know which of these variables is the most important for discriminating between the categories. Classification methods can also be used just to quantify differences between groups, as in Exhibits 1.5 and 1.8 of Chapter 1. There we observed some

differences between the sediment types for one variable at a time; the multivariate challenge will be to see if we can quantify *combinations* of variables that explain group differences.

Clustering: Structural methods uncovering a latent categorical variable

We now move down to the structural methods, where the unobserved latent variable **f** is sought that "explains" the many observed variables **Y**. This is a much more subtle area than the functional one, almost abstract in nature: how can a variable be "unobserved"? Well, let us suppose that we have collected a whole bunch of data on clams found in the Arctic. Are they all of the same species? Suppose they are not and there are really two species involved, but we cannot observe for a particular clam whether it is of species A or B, we just do not know. So species is an unobserved categorical variable. Because it is categorical, we are in the bottom right area of the scheme of Exhibit 2.2. The idea of clustering is to look for similarities between the individual clams, not on single variables but across all measured variables. Can we come up with a grouping (i.e., *clustering*) of the clams into two clusters, each of which consists internally of quite similar individuals, but which are quite different if we compare individuals from different clusters? This is the objective of cluster analysis, to create a categorical structure on the data which assigns each individual to a cluster category. Supposing that the cluster analysis does come up with two clusters of clams, it is then up to the marine biologist to consider the differences between the two clusters to assess if these are large enough to warrant evidence of two different species.

Cluster analysis is often called *unsupervised learning* because the agenda is open to whether groups really do exist and how many there are; hence we are learning without guidance, as it were. Classification methods, on the other hand, are sometimes called *supervised learning* because we know exactly what the groups are and we are trying to learn how to predict them.

Scaling/ordination: Structural methods uncovering a latent continuous variable

The final class of methods, at bottom left in Exhibit 2.2, comprise the various techniques of *scaling*, more often referred to as *ordination* by ecologists. Ordination is just like clustering except the structures that we are looking for in the data are not of a categorical nature, but continuous. Examples of ordination abound in environmental science, so this will be one of the golden threads throughout this book. The origins of scaling, however, come from psychology where measurement is an issue more than in any other scientific discipline. It is relatively simple for a marine biologist to measure a general level of "pollution" – although the various chemical analyses may be expensive, reliable figures can be obtained of heavy metal concentrations and organic materials in any given sample. A psychologist interested in emotions such as anxiety or satisfaction, has a much more difficult job arriving at a reliable quantification. Dozens of measurements could be made to assess the level of anxiety, for example, most of them "soft" in the

sense that they could be answers to a battery of questions on how the respondent feels. Scaling attempts to discern whether there is some underlying dimension (i.e., a scale) in these data which is ordering the respondents from one extreme to another. If this dimension can be picked up, it is then up to the psychologist, to decide whether it validly orders people along a continuous construct that one might call "anxiety".

In the large data sets collected by environmental biologists, the search for continuous constructs can be the identification of various environmental *gradients* in the data, for example pollution or temperature gradients, and of geographical gradients (e.g., north–south). Almost always, several gradients (i.e., several continuous latent variables **f**) can be identified in the data, and these provide new ways of interpreting the data, not in terms of their original variables, which are many, but in terms of these fewer latent dimensions. Because of the importance of ordination and reduction of dimensionality in environmental research, a large part of this book will be devoted to this area.

Hybrid methods

It is in the nature of scientific endeavour that generalizations are made that move the field ahead while including everything that has been discovered before. Statistics is no exception and there are many examples of methods developed as generalizations of previous work, or a gathering together of interconnected methodologies. "General linear modelling" and "generalized linear modelling" (similar in name but different in scope) are two such examples.

In classical linear regression of a continuous response variable there are several variants: multiple regression (where all the explanatory variables are continuous), analysis of variance (ANOVA, where all the explanatory variables are categorical), and analysis of covariance (ANCOVA, where the explanatory variables are continuous and categorical). Each of these has its own quirks and diagnostics and terminology. Design and analysis of experiments usually involve ANOVA or ANCOVA, where the cases are assigned to various treatments in order to be able to estimate their effects. All of these methods are subsumed under the umbrella of the *general linear model*, which falls into the regression corner of Exhibit 2.2.

A more fundamental gathering together of methodologies has taken place in the form of *generalized* linear modelling. Many techniques of regression and classification, which we grouped together as functional methods, have an inherent similarity in that the explanatory variables are combined linearly in order to make models or predictions of the response variable. The aspect that distinguishes them is how that linear function is used to connect with the response variable, and what probability distribution is assumed for the conditional distributions of the response. The *generalized linear model* involves firstly the choice of a function that

acts as the link between the mean of the response variable and the predictors, and secondly the choice of a distribution of the response around this mean – these choices lead to many well-known modelling methods as special cases. Generalized linear modelling straddles both functional corners of our four corner multivariate analysis scheme. Multiple linear regression is the simplest generalized linear model, while logistic regression (when responses are categorical) and Poisson regression (when responses are counts) are other examples – more details will be given in Chapter 18.

Well-known methods in environmental science are *canonical correspondence analysis* and *redundancy analysis* (see Chapter 15). These are basically ordination methods but force the ordination scales to be functions of observed explanatory variables, which recalls the idea of regression. Hence canonical correspondence analysis can be said to straddle the upper and lower corners on the left of our scheme. The method of *partial least squares* has a similar objective, but allows the inclusion of a very large number of explanatory variables.

Finally, the generalization of all generalizations is potentially *structural equation modelling*. We say "potentially" because it is presently covering at least the two left hand continuous corners of our scheme and, as the field develops, moving to cover the right hand ones as well. This area of methodology and its accompanying terminology are very specific to psychological and sociological research at the moment, but could easily find wider use in the environmental sciences as more options are added to handle count and categorical response data.

SUMMARY: The four corners of multivariate analysis

1. Methods of multivariate analysis treat rectangular data matrices where the rows are usually cases, individuals, sampling or experimental units, and the columns are variables.

2. A basic classification of methods can be achieved by first distinguishing the overall objective of a study as either (i) explaining an observed "response" variable in terms of the others, or (ii) ascertaining the inherent structure in the form of a "latent" variable that underlies the set of observed variables. This separates functional from structural methods, respectively.

3. Functional and structural methods can be subdivided into those where (in the case of functional methods) the response variable is continuous or categorical, or (in the case of structural methods) where the identified latent structure is of a continuous or categorical nature.

4. Thus four main classes of methods exist: functional methods explaining a continuous variable (regression and related methods), functional methods explaining a categorical variable (classification), structural methods with latent

structure that is continuous (scaling/ordination) and structural methods with latent categorical structure (clustering).

5. Several general methodologies, such as general and generalized linear models, canonical correspondence analysis, partial least squares and structural equation modelling, can cover more than one of these classes.

CHAPTER 3

Measurement Scales, Transformation and Standardization

To conclude this introductory part on multivariate data analysis, we present a discussion about scales of measurement and the various possibilities for transforming variables. Questions such as the following can plague environmental biologists: "Should I log-transform my data?", "How do I analyse a data set where there is a mixture of continuous and categorical variables?", "My data are not normally distributed, does this matter? And if it does, help!", "Do I need to standardize my data?" and "My data are percentages that add up to 100: does this make a difference to the analysis?" The answers to some of these questions will only become fully apparent later, but at least in this chapter we will catalogue some of the issues involved and list some of the standard ways of transforming data. Readers can optionally skip this chapter for the moment if they are keen to proceed, and dip into it later as we refer back to these issues when they come up in real applications.

Contents

Data theory

Data are the manifestation of statistical variables, and these variables can be classified into various types according to their scales of measurement. Our own personal hierarchical classification of measurement scales is depicted in Exhibit 3.1. The main division, as we have mentioned already, is between categorical and continuous scales. This is a pragmatic distinction, because in reality all observed data are categorical. As Michael types these words his age is 59.5475888 years

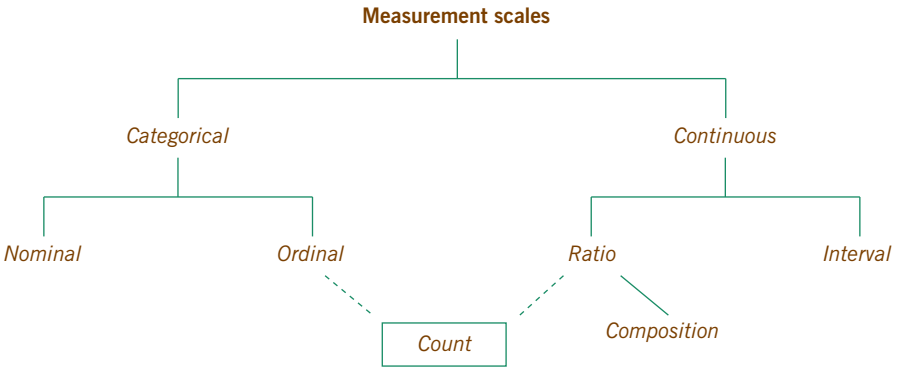


Exhibit 3.1: *A classification of data in terms of their measurement scales. A variable can be categorical (nominal or ordinal) or continuous (ratio or interval). Count data have a special place: they are usually thought of as ratio variables but the discreteness of their values links them to ordinal categorical data. Compositional data are a special case of ratio data that are compositional in a collective sense because of their "unit-sum" constraint*

(to seven decimal places and, theoretically, we could give it to you with even more accuracy), but it has now already advanced to 59.5475902 years in the interim, and is increasing by the second! Of course, in any statistical study, for example an epidemiological study, his age would be recorded simply as 59, having been *discretized*. But we still consider the value 59 to be the manifestation of the continuous variable "age".

Categorical data can be measured on a *nominal* or *ordinal* scale. Nominal categories have no ordering: for example, region of sampling (1. Tarehola, 2. Skognes, 3. Njosken, or 4. Storura), and habitat (1. pelagic, or 2. littoral), hence the numbers recorded in the database have no numerical meaning apart from assigning the samples into groups. Ordinal categories do have an ordering: for example, nature of substrate (1. clay, 2. silt, 3. sand, 4. gravel, or 5. stone – these are ordered by grain size), and month of sampling (1. June, 2. July, 3. August, 4. September), hence the ordering of the numbers (but not their actual values) can be taken into account in the subsequent statistical analysis. Circular ordering of categories (e.g., directions N, NE, E, SE, S, SW, W, NW) is a very special case, as are angular data in the continuous case, where 360° is identical to 0°.

Continuous data can be measured on a *ratio* or *interval* scale. A continuous scale is classified as ratio when two numbers on the scale are compared *multiplicatively*, and an interval scale is when they are compared *additively*. For example, age – in fact, any variable measuring time – is an interval variable. We would not say that Michael's age increased by 0.000002% (the multiplicative increase) in the time it took him to write that sentence above, but we would simply say that it increased by 44 seconds (the additive increase). Ratio variables are almost always nonnegative and have a fixed zero value: for example, biomass, concentration, length, euros and tonnage. Temperature, even though it does have an absolute zero, is an interval variable, unless you like to say that today is 2.6% hotter than yesterday

(with respect to absolute zero) – we prefer to say that the temperature today has risen by 7 °C compared to yesterday's 20 °C.

Count data have a special place in the scheme of Exhibit 3.1, as they can be considered both ordinal and ratio. When 23 individuals of *Galothowenia oculata* are counted in a marine benthic sample, is that a continuous variable? We could not have counted a fraction of an individual, so this sounds like an ordinal categorical observation, but with many possible categories. On the other hand, in a survey of family sizes in Europe, we find only a few values – 0, 1, 2, 3 and 4 children and a sprinkling of families with 5 or more. This sounds more ordinal categorical and less continuous than the *Galothowenia oculata* count. The truth is that they can be validly considered one or the other, depending on how many possible values are observable. If there are many possible values, as in the case of species abundance, then we tend to think of it as a ratio variable. Another aspect is whether we model the expected, or average, count, which is theoretically continuous: for example, at a given sampling location we might predict an average *Galothowenia oculata* abundance of 10.57, even though individual counts are, of course, integers.

Finally, we have singled out compositional data as a special case – these are proportions that add up to 1, a property called *closure*, or the *unit-sum constraint*. The compositional label applies to a set of variables, not to a single one, since it is the property of the set that gives it that nature. Compositional data are usually created from a set of counts or a set of ratio variables when their total is not as relevant as the composition formed by the parts. For example, when we count different species sampled at a particular site, it is likely that the total number is not so relevant, but rather the proportion that each species contributes to the overall count. But if the sampling sites were exactly the same size, as in quadrat sampling in botany, then the overall counts would also be valid measures of overall abundance per unit area sampled. By contrast, a geochemist looking at a mineral sample is not concerned about the weight or volume of the particular sample but in the breakdown of that sample into its components. The situation is identical for fatty acid studies in biology where the data are inherently proportions or percentages, with the overall size of the material sampled having no relevance at all.

The myth of the normal distribution

One of the thorniest issues for applied researchers is that of the normal distribution – most would think that their data should be normal or close to normal in order to arrive at valid conclusions subsequently. This belief is mostly misguided, however, and is a myth created in idealized statistics courses that assume that everything is normally distributed and teach very little about nonparametric statistics, categorical data analysis and modern hypothesis testing using computer-based algorithms such as permutation testing and bootstrapping (see Chapter 17). In any case, it is important to distinguish between *exploratory* and *confirmatory*

data analysis. In data exploration, which is actually the theme of most of the present book, we are considering methods to summarize and interpret large data sets, to give us an understanding of the information that we have painstakingly collected and to diagnose relationships between the observed variables. The normal distribution is a minor issue here, but outliers and standardization and transformations are major ones, which we deal with soon. In the second case of confirmatory analysis, which we will touch on now and again in passing, data are assumed to be representative of a wider population and we want to make conclusions, called *inferences*, about that population. An example of an inference might be that a particular underlying gradient detected in the sample exists in the population with a high degree of probability, based on statistical hypothesis testing. Here we need to know the probabilistic characteristics of the population, and the assumption of normality is the easiest (and most studied) choice. There are, however, other solutions which do not depend on this assumption at all. But, having said this, the idea of data being approximately normally distributed, or at least symmetrically distributed, does have some advantages in exploratory analysis too.

Most of the methods we use are what we call *least-squares* methods that were developed in the context of well-behaved normally distributed data. By "least squares" we mean that solutions are found by minimizing an error criterion defined as a sum of squared differences between our estimated (or "fitted") solution and the observed data. Even in our simple data set of Chapter 1 (Exhibit 1.1) we have seen that the variables are generally not symmetrically distributed around their means. The count variables in Exhibit 1.3, for example, show very skew distributions, with mostly low values and a few much higher ones. Data analysis with these variables, using standard least-squares procedures to fit the models, will be sensitive to the higher values, where the larger error in fitting the high values is even larger when squared. There are several solutions to this problem: one is to use a different theory – for example, maximum likelihood rather than least squares – or make some transformation of the data to make the distributions more symmetric and closer to "well-behaved" normal. Another possibility, used often in the case of count data, is to introduce weights into the analysis, where rare or unusual values are downweighted and contribute less to the results (for example, see Chapters 13 and 14 on correspondence analysis and log-ratio analysis).

Logarithmic transformation of ratio data

Since most ratio variables are skew with long tails to the right, a very good all-purpose transformation is the logarithmic one. This not only pulls in the long tails but also converts multiplicative relationships to additive ones, since $\log(ab) = \log(a) + \log(b)$ – this is advantageous not only for interpretation but also because most of the methods we use involve addition and subtraction. The logarithmic function is shown in Exhibit 3.2 (the lowest curve) as well as other

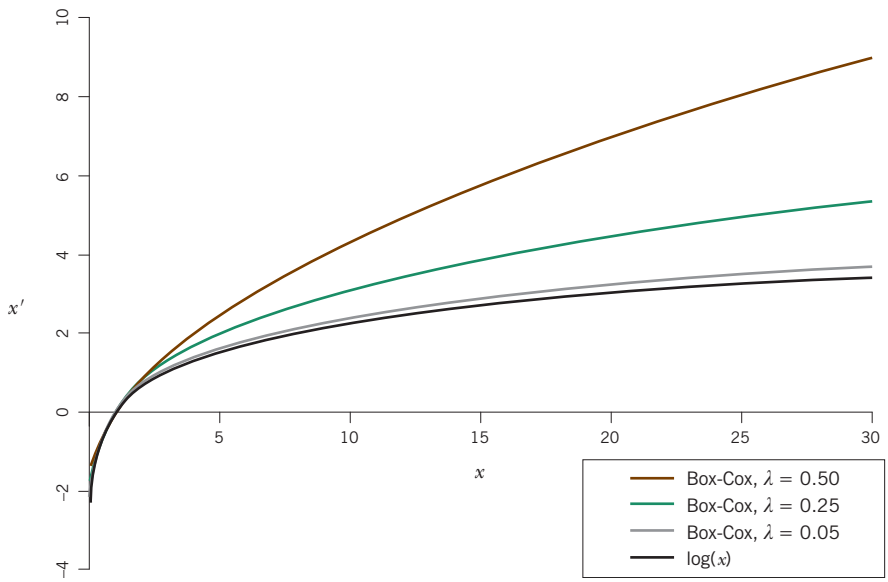


Exhibit 3.2:
The natural logarithmic transformation $x' = log(x)$ and a few Box-Cox power transformations, for powers $\lambda = ½$ (square root), ¼ (double square root, or fourth root) and 0.05

functions that will be described in the next section. Notice how large values of the original variable *a* are pulled down by the log-transformation.

To illustrate the effect the log-transformation has on the interpretation of a variable, consider first the simple linear additive relationship expressed in this equation between the average abundance of a certain marine species and the concentration of the heavy metal barium:

$$abundance = \mathrm{C} - 0.023\ Ba \tag{3.1}$$

where C is some constant. The interpretation is that abundance decreases on average by 0.023 per unit increase of barium (measured in ppm), or 2.3 per 100 units increase in barium. Now consider another equation where abundance has been log-transformed using the natural logarithm (sometimes denoted by "ln"):

$$\log(abundance) = \mathrm{C}' - 0.0017\ Ba \tag{3.2}$$

where C′ is another constant. A unit increase in *Ba* now decreases the logarithm of *abundance* on average by 0.0017. If we exponentiate both sides of equation (3.2), which is the inverse transformation of the natural logarithm, we obtain:

$$abundance = e^{(\mathrm{C}' - 0.0017\ Ba)} = e^{\mathrm{C}'} \cdot e^{(-0.0017\ Ba)} \tag{3.3}$$

That is, a unit increase in barium changes $\exp(-0.0017\,Ba)$ to $\exp(-0.0017\,[Ba+1]) = \exp(-0.0017\,Ba)\cdot\exp(-0.0017)$. So the effect is that abundance is multiplied by $\exp(-0.0017) = 0.9983$, in other words a 0.17% decrease. For a 100 unit increase in barium, abundance is multiplied by $\exp(-0.0017\times 100) = \exp(-0.17) = 0.8437$, a 15.63% decrease. Notice that this is not a $100\times 0.17\% = 17\%$ decrease since the multiplicative effect is compounded (just like interest calculations in finance where the "capital" is being diminished). The above example shows how the logarithmic transformation converts an additive effect into a multiplicative one.

Power transformations and Box-Cox

In Exhibit 3.2 three other curves are shown corresponding to power transformations of the variable x, called *Box-Cox transformations* after two of the most influential statisticians of the 20th century, the American George Box and the Englishman Sir David Cox. These are a slight modification of a simple power transformation x^λ and take the following form:

$$x' = \frac{1}{\lambda}(x^\lambda - 1) \tag{3.4}$$

The advantage of this form is that it tends to the log-transformation as the power parameter tends to 0, as shown in Exhibit 3.2 – as λ decreases the curve approaches the logarithmic curve. The division by λ conveniently keeps the scale of the original variable from collapsing: for example, if you take the 20th roots (that is, $x^{0.05}$) of a set of data, you will quickly see that all the values are close to 1, so the division by 0.05, which multiplies the values by 20, restores them to an almost logarithmic scale.

Box-Cox transformations serve as a flexible way of symmetrizing data and have found extensive application in regression analysis. The inverse transformation is:

$$x = (1+\lambda x')^{\frac{1}{\lambda}} \tag{3.5}$$

where x' is the transformed value in (3.4). We shall refer to these transformations in Chapter 14 in our treatment of compositional data.

Dummy variables

In functional methods of regression and classification, there is no problem at all to have some continuous and some categorical predictors. The categorical variables are coded as *dummy variables*, which are variables that take on the values 0 or 1. For example, suppose one of the predictors is sampling region, with four regions. This variable is coded as four dummy variables which have values [1 0 0 0] for region A, [0 1 0 0] for region B, [0 0 1 0] for region C and

[0 0 0 1] for region D. For a technical reason only 3 out of these 4 dummies can be used – the statistical program usually does this all automatically, omitting (for example) the last dummy for region D. Then the results for the three included dummies are interpreted as the differences of those three regions compared to the omitted region. If the categorical variable has only two categories, for example pelagic or littoral habitat, then only one dummy variable is included, omitting the one for littoral, for example, in which case the model estimates the effect of the difference between pelagic and littoral habitats.

For structural methods, however, the situation is more complicated, because we are trying to explore structure amongst all the variables and here the coding does matter. We could resort to dummy variable coding of all the categorical variables but this is not satisfactory because of the inherently different variances in the dummy variables compared to the continuous ones. For example, a danger might exist that the dummy variables have much less variance than the continuous variables, so when we look for structure we only see patterns in the continuous variables while those in the categorical variables are more or less "invisible" to our investigation. We need to balance the contributions of the variables in some way that gives them all a fair chance of competing for our attention. This is a problem of *standardization*, which we treat in detail in a later section.

Fuzzy coding

An alternative approach to the problem of mixed-scale data is to recode the continuous variables also as dummy variables, so that we put them on the same scale as the categorical dummies. This can be achieved by dividing up the continuous scale into intervals, for example three intervals which can be labelled "low", "medium" and "high". Clearly, this loses a lot of information in the continuous variables, so there is a way to avoid data loss called *fuzzy coding*. If we again choose the three-category option, then a continuous variable can be fuzzy coded as shown in Exhibit 3.3.

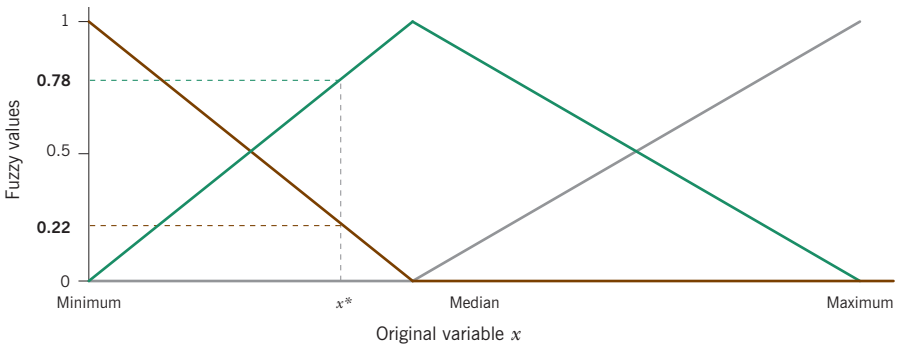


Exhibit 3.3:
Fuzzy coding of a continuous variable x into three categories, using triangular membership functions. The minimum, median and maximum are used as hinge points. An example is given of a value x just below the median being fuzzy coded as [0.22 0.78 0]*

In this example we have used the simplest procedure, *triangular membership functions,* for fuzzy coding. For three categories we need three *hinge points*, which we have chosen to be the minimum, median and maximum of the continuous variable *x*. Triangles are drawn as shown and these provide the fuzzy values for the three categories – notice that the third category, drawn in gray, has value zero below the median. The general algorithm for computing the three fuzzy values $[z_1\ z_2\ z_3]$ is as follows:

$$z_1(x) = \begin{cases} \dfrac{m_2 - x}{m_2 - m_1}, & \text{for } x \le m_2 \\ 0 \ \text{ otherwise} \end{cases}$$

$$z_2(x) = \begin{cases} \dfrac{x - m_1}{m_2 - m_1}, & \text{for } x \le m_2 \\ \dfrac{m_3 - x}{m_3 - m_2}, & \text{for } x > m_2 \end{cases} \qquad (3.6)$$

$$z_3(x) = \begin{cases} \dfrac{x - m_2}{m_3 - m_2}, & \text{for } x > m_2 \\ 0 \ \text{ otherwise} \end{cases}$$

where m_1, m_2 and m_3 denote the three hinge points. For example, in Exhibit 3.3, the hinges were $m_1 = 3.69$, $m_2 = 8.64$ and $m_3 = 19.65$. The value x^* was 7.55 and was fuzzy coded as $z_1(7.55) = (8.64 - 7.55) / (8.64 - 3.69) = 0.22$; $z_2(7.55) = (7.55 - 3.69) / (8.64 - 3.69) = 0.78$, and $z_3(7.55) = 0$.

The advantage of this coding is that it is invertible – we can recover the original value from the fuzzy values as a linear combination of the hinge values (in fact, a weighted average since the fuzzy values add up to 1):

$$x = z_1\ m_1 + z_2\ m_2 + z_3\ m_3 \qquad (3.7)$$

for example, $0.22 \times 3.69 + 0.78 \times 8.64 + 0 \times 19.65 = 7.55$. This reverse process of going from the fuzzy values back to the original data is called *defuzzification.* The fact that the fuzzy coding is reversible means that we have conserved all the information in the coded values, while gaining the advantage of converting the continuous variable to a form similar to the categorical dummies. However, there is still a problem of balancing the variances, which we now discuss.

Standardization

Standardization is an important issue in structural methods of multivariate analysis. Variables on different scales have natural variances which depend mostly on

their scales. For example, suppose we measure the length of a dorsal fin of a sample of fish in centimeters – the variance of our measurements across the sample might be 0.503 cm^2, and the standard deviation 0.709 cm (the square root of the variance). Then we decide to express the lengths in millimeters, because most of the other measurements are in millimeters; so the variance is now 50.3 mm^2, a hundred times the previous value, while the standard deviation is 7.09 mm, ten times more. The fact that some variables can have high variances just because of the chosen scale of measurement causes problems when we look for structure amongst the variables. The variables with high variance will dominate our search because they appear to contain more information, while those with low variance are swamped because of their small differences between values.

The answer is clearly to balance out the variances so that each variable can play an equal role in our analysis – this is exactly what standardization tries to achieve. The simplest form of standardization is to make all variances in the data set exactly the same. For a bunch of continuous variables, for example, we would divide the values of each variable by its corresponding sample standard deviation so that each variable has variance (and also standard deviation) equal to 1. Often this is accompanied by *centering* the variable as well, that is, subtracting its mean, in which case we often refer to the standardized variable as a *Z-score*. This terminology originates in the standardization of a normally distributed variable *X*, which after subtracting its mean and dividing by its standard deviation is customarily denoted by the letter *Z* and called a *standard normal variable,* with mean 0 and variance 1.

Standardization can also be thought of as a form of weighting. That is, by dividing variables with large variances by their large standard deviations, we are actually multiplying them by small numbers and reducing their weight. The variables with small variances, on the other hand, are divided by smaller standard deviations and thus have their weight increased relative to the others.

Other forms of standardization are:

- by the range: each variable is linearly transformed to lie between 0 and 1, where 0 is its minimum and 1 its maximum value;

- by chosen percentiles: because the range is sensitive to outliers, we can "peg" the 0 and 1 values of the linearly transformed variable to, say, the 5^{th} and 95^{th} percentile of the sample distribution;

- by the mean: the values of a variable are divided by their mean, so that they have standard deviations equal to what is called their *coefficient of variation*.

There are various forms of standardization which rely on the assumed theoretical characteristics of the variable. For example, count data are often assumed to come from a *Poisson distribution*. This distribution has the property that the variance is theoretically equal to the mean. Thus, dividing by the square root of the mean would be like dividing by the standard deviation (this is, in fact, the standardization inherent in correspondence analysis – see Chapter 13). Another theoretical result is that, while a Poisson variable has variance that increases as the average count increases, its square root has a variance tending to a constant value of ¼. Hence, an alternative form of standardization that is regularly used to "stabilize" the variance of count data is simply to square root transform them.

Finally, coming back to the handling of continuous and categorical variables jointly, where the continuous variables have been coded into fuzzy dummy variables and the categorical variables into "crisp" (zero-one) dummies, we could standardize by calculating the collective variance of each set of dummies corresponding to one variable and then weighting the set accordingly. That is, we do not standardize individual dummy variables, which would be incorrect, but each group as a whole.

SUMMARY: Measurement scales, transformation and standardization

1. Variables can be either categorical or continuous, although all measurements are categorical in the sense of being discretized. Continuous variables are those that have very many categories, for example a count variable, or are discretized versions of a variable which could, at least theoretically, be measured on a continuous scale, for example a length or a concentration.
2. Categorical variables can be either ordinal or nominal, depending on whether the categories have an inherent ordering or not.
3. Continuous variables can be either ratio or interval, depending on whether we compare two observations on that variable multiplicatively (as a ratio) or additively (as a difference).
4. The logarithmic transformation is a very useful transformation for most positive ratio measurements, because multiplicative comparisons are converted to additive ones and because high values are pulled in, making the distribution of the variable more symmetric.
5. Box-Cox transformations are a very flexible class of power transformations which include the log-transformation as a limiting case.
6. Categorical variables are usually coded as dummy variables in order to be able to judge the effect or relationship of individual categories.
7. Continuous variables can also be dummy coded but this loses a lot of information. A better option is to fuzzy code them into a small number of categories,

which allows continuous variables to be analysed together with categorical ones more easily, especially in the case of structural multivariate methods.

8. In structural methods standardization is a major issue for consideration. Variances of the variables being analysed need to be balanced in some way that gives each variable a fair chance of being involved in the determination of the latent structure. Results should not depend on the scale of measurement. Standardization is not an issue for functional methods because the effect of a variable on a response is measured independently of the scale.

MEASURING DISTANCE AND CORRELATION

CHAPTER 4

Measures of Distance between Samples: Euclidean

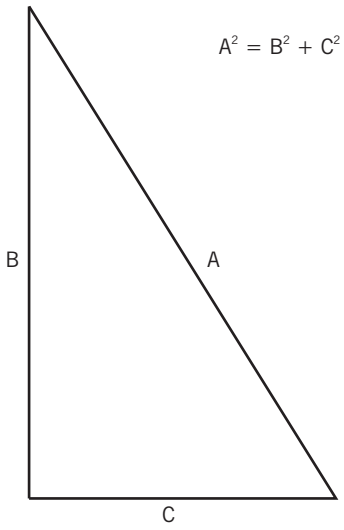
We will be talking a lot about distances in this book. The concept of distance between two samples or between two variables is fundamental in multivariate analysis – almost everything we do has a relation with this measure. If we talk about a single variable we take this concept for granted. If one sample has a pH of 6.1 and another a pH of 7.5, the absolute difference between them is 1.4. But on the pH line, the values 6.1 and 7.5 are at a distance apart of 1.4 units, and this is how we want to start thinking about data: points on a line, points in a plane, … even points in a 10-dimensional space! So, given two samples with not one measurement on them but several, how do we measure the difference between them? There are many possible answers to this question, and we devote three chapters to this topic. In the present chapter we consider what are called *Euclidean* distances, which coincide with our basic physical idea of distance, but generalized to multidimensional space.

Contents

Pythagoras' theorem

Pythagoras' theorem is at the heart of most of the multivariate analysis presented in this book, and particularly the graphical approach to data analysis that we are strongly promoting. When you see the word "square" mentioned in a statistical text (for example, chi-square or least squares), you can be almost sure that the corresponding theory has some relation to this theorem. We first

Exhibit 4.1:
Pythagoras' theorem in the familiar right-angled triangle, and the monument to this triangle in the port of Pythagorion, Samos island, Greece, with Pythagoras himself forming one of the sides (Photo: Michael Greenacre)



show the theorem in its simplest and most familiar two-dimensional form, before showing how easy it is to generalize it to multidimensional space. In a right-angled triangle, the square on the hypotenuse (the side denoted by A in Exhibit 4.1) is equal to the sum of the squares on the other two sides (B and C); that is, $A^2 = B^2 + C^2$.

Euclidean distance

The immediate consequence of this is that the squared length of a vector $\mathbf{x} = [\,x_1 \;\; x_2\,]$ is the sum of the squares of its coordinates (see triangle OPA in Exhibit 4.2, or triangle OPB – $|OP|^2$ denotes the squared length of **x**, that is the distance between point O, with both co-ordinates zero, and P); and the squared distance between two vectors $\mathbf{x} = [\,x_1 \;\; x_2\,]$ and $\mathbf{y} = [\,y_1 \;\; y_2\,]$ is the sum of squared differences in their coordinates (see triangle PQD in Exhibit 4.2; $|PQ|^2$ denotes the squared distance between points P and Q). To denote the distance between vectors **x** and **y** we can use the notation $d_{\mathbf{x},\mathbf{y}}$ so that this last result can be written as:

$$d^2_{x,y} = (x_1 - y_1)^2 + (x_2 - y_2)^2 \tag{4.1}$$

that is, the distance itself is the square root

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \tag{4.2}$$

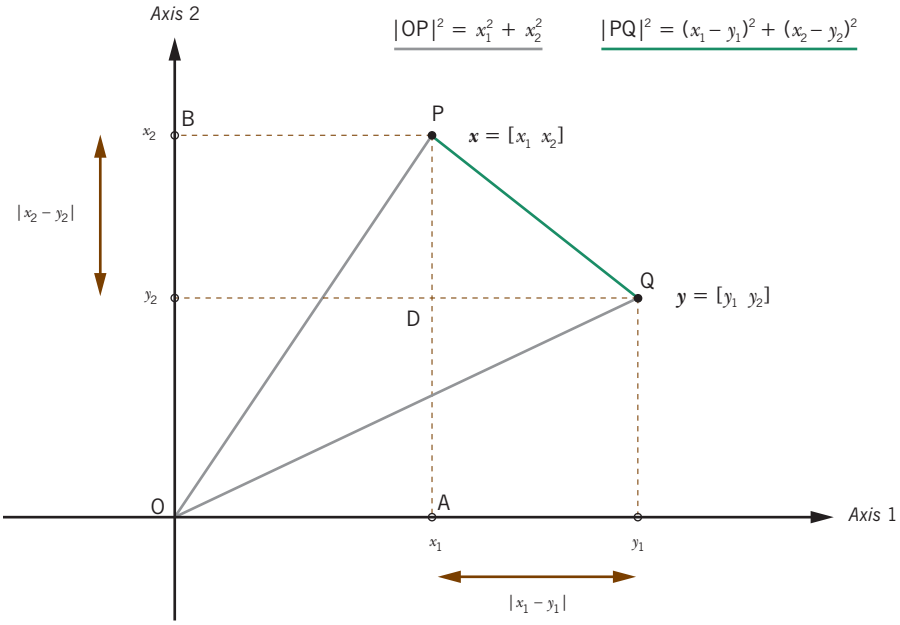


Exhibit 4.2: *Pythagoras' theorem applied to distances in two-dimensional space*

What we called the *squared length* of **x**, the distance between points P and O in Exhibit 4.2, is the distance between the vector $\mathbf{x} = [x_1 \ x_2]$ and the zero vector $\mathbf{0} = [0 \ 0]$:

$$d_{x,0} = \sqrt{x_1^2 + x_2^2} \tag{4.3}$$

which we could just denote by $d_{\mathbf{x}}$. The zero vector is called the *origin* of the space.

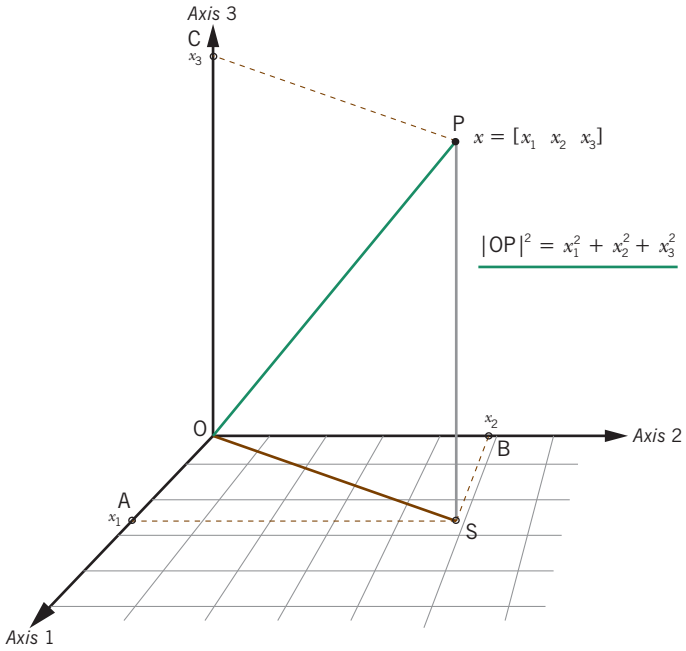
We move immediately to a three-dimensional point $\mathbf{x} = [x_1 \ x_2 \ x_3]$, shown in Exhibit 4.3. This figure has to be imagined in a room where the origin O is at the corner – to reinforce this idea "floor tiles" have been drawn on the plane of axes 1 and 2, which is the "floor" of the room. The three coordinates are at points A, B and C along the axes, and the angles AOB, AOC and COB are all 90° as well as the angle OSP at S, where the point P (depicting vector **x**) is projected onto the "floor". Using Pythagoras' theorem twice we have:

$|OP|^2 = |OS|^2 + |PS|^2$ (because of right-angle at S)
$|OS|^2 = |OA|^2 + |AS|^2$ (because of right-angle at A)

and so

$$|OP|^2 = |OA|^2 + |AS|^2 + |PS|^2$$

Exhibit 4.3:
Pythagoras' theorem extended into three dimensional space



that is, the squared length of **x** is the sum of its three squared coordinates, hence the length is

$$d_x = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

It is also clear that placing a point Q in Exhibit 4.3 to depict another vector **y** and going through the motions to calculate the distance between **x** and **y** will lead to

$$d_{x,y} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2} \tag{4.4}$$

Furthermore, we can carry on like this into four or more dimensions, in general J dimensions, where J is the number of variables. Although we cannot draw the geometry any more, we can express the distance between two J-dimensional vectors **x** and **y** as:

$$d_{x,y} = \sqrt{\sum_{j=1}^{J} (x_j - y_j)^2} \tag{4.5}$$

This well-known distance measure, which generalizes our notion of physical distance in two- or three-dimensional space to multidimensional space, is called the *Euclidean distance*.

Standardized Euclidean distance

Let us consider measuring the distances between our 30 samples in Exhibit 1.1, using the three continuous variables depth, pollution and temperature. What would happen if we applied formula (4.5) to measure distance between the last two samples, s29 and s30, for example? Here is the calculation:

$$d_{s29,s30} = \sqrt{(51-99)^2 + (6.0-1.9)^2 + (3.0-2.9)^2}$$

$$= \sqrt{2304 + 16.81 + 0.01} = \sqrt{2320.82} = 48.17$$

The contribution of the first variable depth to this calculation is huge – one could say that the distance is practically just the absolute difference in the depth values (equal to |51-99| = 48) with only tiny additional contributions from pollution and temperature. This is the problem of standardization discussed in Chapter 3 – the three variables have completely different units of measurement and the larger depth values have larger inter-sample differences, so they will dominate in the calculation of Euclidean distances.

Some form of transformation of the data is necessary to balance out the contributions, and the conventional way to do this is to make all variables have the same variance of 1. At the same time we centre the variables at their means – this centring is not necessary for calculating distance, but it makes the variables all have mean zero and thus easier to compare. This transformation, commonly called *standardization*, is thus as follows:

standardized value = (original value – mean) / standard deviation (4.6)

The means and standard deviations (sd) of the three variables are:

	Depth	*Pollution*	*Temperature*
mean	74.433	4.517	3.057
sd	15.615	2.141	0.281

leading to the table of standardized values given in Exhibit 4.4. These values are now on comparable standardized scales, in units of standard deviation with respect to the mean. For example, the standardized pollution value 0.693 for row s29 would signify 0.693 standard deviations above the mean, while –1.222 for row s30 would signify 1.222 standard deviations below the mean. The distance calculation thus aggregates squared differences in standard deviation units of each variable. As an example, the distance between the last two sites of the table in Exhibit 4.4 is:

$$d_{s29,s30} = \sqrt{[-1.501 - 1.573]^2 + [0.693 - (-1.222)]^2 + [-0.201 - (-.557)]^2}$$

$$= \sqrt{9.449 + 3.667 + 0.127} = \sqrt{13.243} = 3.639$$

For this particular pair of sites the difference in temperatures is still small but pollution now has a higher contribution than before. Depth still plays the largest role in this particular example, even after standardization, but this contribution is

Exhibit 4.4: *Standardized values of the three continuous variables of Exhibit 1.1*

Site No.	Environmental variables		
	Depth	*Pollution*	*Temperature*
s1	–0.156	0.132	1.576
s2	0.036	–0.802	–1.979
s3	–0.988	0.413	–1.268
s4	–0.668	1.720	–0.557
s5	–0.860	–0.288	0.154
s6	1.253	–0.895	1.576
s7	–1.373	0.039	–0.557
s8	–0.860	0.272	0.865
s9	–0.412	–0.288	1.221
s10	–0.348	2.561	–0.201
s11	–1.116	0.926	0.865
s12	0.613	–0.335	0.154
s13	–1.373	2.281	–0.201
s14	0.549	0.086	–1.979
s15	1.637	1.020	–0.913
s16	0.613	–0.802	–0.201
s17	1.381	0.880	0.154
s18	–0.028	–0.054	–0.913
s19	0.292	–0.662	1.932
s20	–0.092	0.506	–0.201
s21	–0.988	–0.101	1.221
s22	–1.309	–1.222	–0.913
s23	1.317	–0.989	–0.557
s24	–0.668	–0.101	–0.201
s25	1.445	–1.175	–0.201
s26	0.228	–0.942	1.221
s27	0.677	–1.129	–0.201
s28	1.125	–0.522	0.865
s29	–1.501	0.693	–0.201
s30	1.573	–1.222	–0.557

	s1	s2	s3	s4	s5	s6	···	s24	s25	s26	s27	s28	s29
s2	3.681												
s3	2.977	1.741											
s4	2.708	2.980	1.523										
s5	1.642	2.371	1.591	2.139									
s6	1.744	3.759	3.850	3.884	2.619								
s7	2.458	2.171	0.890	1.823	0.935	3.510							
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱						
s25	2.727	2.299	3.095	3.602	2.496	1.810	···	2.371					
s26	1.195	3.209	3.084	3.324	1.658	1.086	···	1.880	1.886				
s27	2.333	1.918	2.507	3.170	1.788	1.884	···	1.692	0.770	1.503			
s28	1.604	3.059	3.145	3.204	2.122	0.813	···	2.128	1.291	1.052	1.307		
s29	2.299	2.785	1.216	1.369	1.224	3.642	···	1.150	3.488	2.772	2.839	3.083	
s30	3.062	2.136	3.121	3.699	2.702	2.182	···	2.531	0.381	2.247	0.969	1.648	3.639

Exhibit 4.5: *Standardized Euclidean distances between the 30 samples, based on the three continuous environmental variables, showing part of the triangular distance matrix*

justified now, since depth does show the biggest standardized difference between the samples. We call this the *standardized Euclidean distance*, meaning that it is the Euclidean distance calculated on standardized data. It will be assumed that standardization refers to the form defined by (4.6), unless specified otherwise.

We can repeat this calculation for all pairs of samples. Since the distance between sample A and sample B will be the same as between sample B and sample A, we can report these distances in a triangular matrix – Exhibit 4.5 shows part of this distance matrix, which contains a total of $\frac{1}{2} \times 30 \times 29 = 435$ distances.

Readers might ask how all this has helped them – why convert a data table with 90 numbers into one that has 435, almost five times more? Were the histograms and scatterplots in Exhibits 1.2 and 1.4 not enough to understand these three variables? This is a good question, but we shall have to leave the answer to Part 3 of the book, from Chapter 7 onwards, when we describe actual analyses of these distance matrices. At this early stage in the book, we can only ask readers to accept that the computation of interpoint distances is an intermediate step in a process that will lead to an eventual simplification in interpreting the data structure – having a measure of distance (i.e., difference) between samples based on several variables is the key to this process.

Weighted Euclidean distance

The standardized Euclidean distance between two J-dimensional vectors can be written as:

$$d_{x,y} = \sqrt{\sum_{j=1}^{J}\left(\frac{x_j}{s_j} - \frac{y_j}{s_j}\right)^2} \tag{4.7}$$

where s_j is the sample standard deviation of the j-th variable. Notice that we need not subtract the j-th mean from x_j and y_j because the means will just cancel out in the differencing. Now (4.7) can be rewritten in the following equivalent way:

$$d_{x,y} = \sqrt{\sum_{j=1}^{J}\frac{1}{s_j^2}(x_j - y_j)^2} = \sqrt{\sum_{j=1}^{J} w_j (x_j - y_j)^2} \tag{4.8}$$

where $w_j = 1/s_j^2$ is the inverse of the j-th variance. We can think of w_j as a *weight* attached to the j-th variable: in other words, we compute the usual squared differences between the variables on their original scales, as we did in the (unstandardized) Euclidean distance, but then multiply these squared differences by their corresponding weights. Notice in this case how the weight of a variable with high variance is low, while the weight of a variable with low variance is high, which is another way of thinking about the compensatory effect produced by standardization. The weights of the three variables in our example are (to 4 significant figures) 0.004101, 0.2181 and 12.64 respectively, showing how much the depth variable is downweighted and the temperature variable upweighted: depth has over 3000 times the variance of temperature, so each squared difference in (4.8) is downweighted relatively by that much. We call (4.8) *weighted Euclidean distance*.

Distances for count data

So far we have looked at the distances between samples based on continuous data, now we consider distances on count data, for example the abundance data for the five species labelled *a*, *b*, *c*, *d* and *e* in Exhibit 1.1. First, notice that these five variables apparently do not have the problem of different measurement units that we had for the continuous environmental variables – all variables are counts. There are, however, different average frequencies of counts, and as we mentioned in Chapter 3, variances of count variables can be positively related to their means. The means and variances of these five variables are as follows:

	a	*b*	*c*	*d*	*e*
mean	13.47	8.73	8.40	10.90	2.97
variance	157.67	83.44	73.62	44.44	15.69

Variable *a* with the highest mean also has the highest variance, while *e* with the lowest mean has the lowest variance. Only *d* is out of line with the others, having smaller variance than *b* and *c* but a higher mean. Because this variance–mean relationship is a natural phenomenon for count variables, not one that is just par-

ticular to any given example, some form of compensation of the variances needs to be performed, as before. It is not common for count data to be standardized as Z-scores (i.e., with mean 0, variance 1), as was the case for continuous variables in (4.6). The most common ways of balancing the contributions of count variables to the distance measure are:

- *a power transformation:* usually square root $n^{1/2}$, where n is the count value, but also double square root (i.e., fourth root $n^{1/4}$) when the variance increases faster than the mean (this situation is called *overdispersion* in the literature);

- *a "shifted log" transformation:* because of the many zeros in ecological count data, a positive number, usually 1, has to be added to the data before log-transforming; that is, $\log(1 + n)$;

- *chi-square distance:* this is a weighted Euclidean distance of the form (4.8), which we shall discuss now.

Chi-square distance

The chi-square distance is special because it is at the heart of correspondence analysis, used extensively in ecological research. The first premise of this distance function is that it is calculated on relative counts,[1] and not on the original ones, and the second is that it standardizes by the mean and not by the variance.

In our example, the count data are first converted into relative counts by dividing the rows by their row totals so that each row contains relative proportions across the species, which add up to 1. These sets of proportions are called *profiles*, site profiles in this example – see Exhibit 4.6.

The extra row at the end of Exhibit 4.6 gives the set of proportions called the *average profile*. These are the proportions calculated on the set of column totals, which are equal to 404, 262, 252, 327 and 89 respectively, with grand total 1334. Hence, 404/1334 = 0.303, 262/1334 = 0.196, etc. Chi-square distances are then calculated between the profiles, in a weighted Euclidean fashion, using the inverse of the average proportions as weights. Suppose c_j denotes the j-th element of the average profile, that is the abundance proportion of the j-th species in the whole data set. Then the *chi-square*[2] *distance*, denoted by χ, between two sites with profiles $\mathbf{x} = [x_1 \;\; x_2 \;\; \cdots \;\; x_J]$ and $\mathbf{y} = [y_1 \;\; y_2 \;\; \cdots \;\; y_J]$ is defined as:

[1] A definition of *chi-square distance* on raw counts is referred to in the bibliographical appendix.
[2] From the definition of this distance function it would have been better to call it the *chi distance function*, because it is not squared, as in the *chi-square statistic!* But the "chi-square" epithet persists in the literature, so when we talk of its square we say the "squared chi-square distance".

$$\chi_{x,y} = \sqrt{\sum_{j=1}^{J} \frac{1}{c_j} (x_j - y_j)^2} \tag{4.8}$$

Exhibit 4.6: *Profiles of the sites, obtained by dividing the rows of counts in Exhibit 1.1 by their respective row totals. The last row is the average profile, computed in the same way, as proportions of the column totals of the original table of counts*

Site No.	Species proportions				
	a	*b*	*c*	*d*	*e*
s1	0.000	0.074	0.333	0.519	0.074
s2	0.481	0.074	0.241	0.204	0.000
s3	0.000	0.370	0.333	0.296	0.000
s4	0.000	0.000	0.833	0.167	0.000
s5	0.342	0.132	0.079	0.263	0.184
s6	0.360	0.244	0.151	0.186	0.058
s7	0.321	0.214	0.000	0.393	0.071
s8	0.667	0.000	0.000	0.000	0.333
s9	0.315	0.130	0.185	0.259	0.111
s10	0.000	0.125	0.650	0.225	0.000
s11	0.000	0.276	0.276	0.207	0.241
s12	0.264	0.208	0.245	0.283	0.000
s13	0.000	0.000	0.760	0.000	0.240
s14	0.591	0.000	0.000	0.409	0.000
s15	0.154	0.000	0.385	0.462	0.000
s16	0.592	0.282	0.000	0.042	0.085
s17	1.000	0.000	0.000	0.000	0.000
s18	0.236	0.169	0.371	0.225	0.000
s19	0.053	0.132	0.316	0.421	0.079
s20	0.000	0.303	0.424	0.273	0.000
s21	0.444	0.000	0.000	0.222	0.333
s22	0.493	0.141	0.000	0.127	0.239
s23	0.146	0.171	0.024	0.415	0.244
s24	0.316	0.211	0.351	0.123	0.000
s25	0.395	0.321	0.000	0.284	0.000
s26	0.492	0.323	0.000	0.154	0.031
s27	0.333	0.236	0.000	0.347	0.083
s28	0.302	0.057	0.226	0.377	0.038
s29	0.423	0.000	0.269	0.308	0.000
s30	0.282	0.435	0.059	0.212	0.012
ave.	0.303	0.196	0.189	0.245	0.067

Exhibit 4.7 shows part of the 30 × 30 triangular matrix of chi-square distances. Once again, this is a large matrix with more numbers (435) than the original table of counts (150), and we shall see the benefit of calculating these distances

	s1	s2	s3	s4	s5	s6	···	s24	s25	s26	s27	s28	s29
s2	1.139												
s3	0.855	1.137											
s4	1.392	1.630	1.446										
s5	1.093	0.862	1.238	2.008									
s6	1.099	0.539	0.887	1.802	0.597								
s7	1.046	0.845	1.081	2.130	0.573	0.555							
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱						
s25	1.312	0.817	1.057	2.185	0.858	0.495	···	0.917					
s26	1.508	0.805	1.224	2.241	0.834	0.475	···	0.915	0.338				
s27	1.100	0.837	1.078	2.136	0.520	0.489	···	0.983	0.412	0.562			
s28	0.681	0.504	0.954	1.572	0.724	0.613	···	0.699	0.844	0.978	0.688		
s29	0.951	0.296	1.145	1.535	0.905	0.708	···	0.662	0.956	1.021	0.897	0.340	
s30	1.330	0.986	0.846	2.101	0.970	0.535	···	0.864	0.388	0.497	0.617	1.001	1.142

Exhibit 4.7: *Chi-square distances between the 30 samples, based on the biological count data, showing part of the triangular distance matrix*

from Part 3 onwards. For the moment, think of Exhibit 4.5 as a way of measuring similarities and differences between the 30 samples based on the (continuous) environmental data, while Exhibit 4.7 is the similar idea but based on the count data. Notice that the scale of distances in Exhibit 4.5 is not comparable to that of Exhibit 4.7, but the ordering of the values does have some meaning: for example, in Exhibit 4.5 the smallest standardized Euclidean distance (amongst those that we report there) is 0.381, between sites s30 and s25. In Exhibit 4.7 these two sites have one of the smallest chi-square distances as well. This means that these two sites are relatively similar in their environmental variables and also in their biological compositions. This might be an interesting finding, but we will need to study all the pairwise distances, and not just this isolated one, in order to see if there is any connection between the biological abundances and the environmental variables (this will come later).

Distances for categorical data

In our introductory example we have only one categorical variable (sediment), so the question of computing distance is fairly trivial: if two samples have the same sediment then their distance is 0, and if different then it is 1. But what if there were several categorical variables, say K of them? There are several possibilities, one of the simplest being to count how many matches and mismatches there are between samples, with optional averaging over variables. For example, suppose that there are five categorical variables, *C1* to *C5*, each with three categories, which we denote by a/b/c and that there are two samples with the following characteristics:

	C1	C2	C3	C4	C5
Sample 1	a	c	c	b	a
Sample 2	b	c	b	a	a

Then the number of matches is 2 and the number of mismatches is 3, hence the distance between the two samples is 3 divided by 5 (the number of variables), that is 0.6. This is called the *simple matching coefficient*. Sometimes this coefficient is expressed in terms of similarity, not dissimilarity, in which case the similarity would be equal to 0.4, the relative number of matches – so one should check which way it is being defined. Here we stick to distances, in other words dissimilarities or mismatches. Note that this coefficient is directly proportional to the squared Euclidean distance calculated between these data in dummy variable form, where each category defines a zero-one variable:

	C1a	C1b	C1c	C2a	C2b	C2c	C3a	C3b	C3c	C4a	C4b	C4c	C5a	C5b	C5c
Sample 1	1	0	0	0	0	1	0	0	1	0	1	0	1	0	0
Sample 2	0	1	0	0	0	1	0	1	0	1	0	0	1	0	0

The squared Euclidean distance sums the squared differences between these two vectors: if there is an agreement (there are two matches in this example) there is zero sum of squared differences, but if there is a discrepancy there are two differences, +1 and −1, which give a sum of squares of 2. So the sum of squared differences here is 6, and if this is expressed relative to the maximum discrepancy that can be achieved, namely 10 when there are no matches in the 5 variables, then this gives exactly the same value 0.6 as before.

There are several variations on the theme of the matching coefficient, and one of them is the chi-square distance for multivariate categorical data, which introduces a weighting of each category inverse to its mean value, as for profile data based on counts. Suppose that there are J categories in total (in the above example $J = 15$) and that the total occurrences of each category are denoted by $n_1,\ldots,n_J$, with total $n = \Sigma_j\, n_j$ (since the totals for each variable equal the sample size, n will be the sample size times the number of variables). Then define c_j as follows: $c_j = n_j/n$ and use $1/c_j$ as weights in a weighted Euclidean distance between the samples coded in dummy variable form. The idea here is, as before, that mismatches on a rare category should have a higher weight in the distance calculation than that of a frequent category. Just like the chi-square distance function is at the heart of correspondence analysis of abundance data, so this form of the chi-square for multivariate categorical data is at the heart of *multiple correspondence analysis*. We do not treat multiple correspondence analysis specifically in this book,

as it is more common in the social sciences where almost all the data are categorical, for example in survey research.

SUMMARY: Measures of distance between samples: Euclidean

1. Pythagoras' theorem extends to sets of observations (called *vectors*) in multidimensional space, for example sets of observations corresponding to a series of samples: the squared length of a vector is the sum of squares of its coordinates.
2. As a consequence, squared distances between two vectors (e.g., between two samples) in multidimensional space are the sum of squared differences in their coordinates. This multidimensional distance is called the *Euclidean distance*, and is the natural generalization of our three-dimensional notion of physical distance to more dimensions.
3. When variables are on different measurement scales, standardization is necessary to balance the contributions of the variables in the computation of distance. The Euclidean distance computed on standardized variables is called the *standardized Euclidean distance*.
4. Standardization in the calculation of distances is equivalently thought of as *weighting* the variables – this leads to the notion of Euclidean distances with any choice of weights, called *weighted Euclidean distance*.
5. A particular weighted Euclidean distance applicable to count data is the *chi-square distance*, which is calculated between the relative counts for each sample, called *profiles*, and weights each variable by the inverse of the variable's overall mean count.

CHAPTER 5

Measures of Distance between Samples: Non-Euclidean

Euclidean distances are special because they conform to our physical concept of distance. But there are many other distance measures which can be defined between multivariate samples. These non-Euclidean distances are of different types: some still satisfy the basic axioms of what mathematicians call a distance metric, while others are not even true metrics but still make good sense as a measure of difference between samples in the context of certain data. In this chapter we shall consider several non-Euclidean distance measures that are popular in the environmental sciences: the Bray-Curtis dissimilarity, the L_1 distance (also called the *city-block* or *Manhattan distance)* and the Jaccard index for presence-absence data. We also consider how to measure dissimilarity between samples for which we have mixed-scale data.

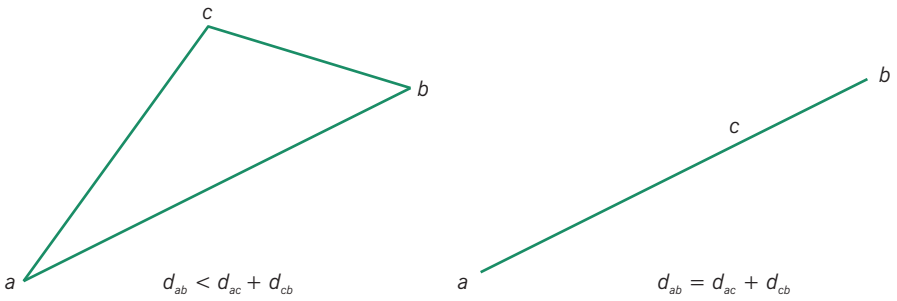
Contents

The axioms of distance

In mathematics, a true measure of distance, also called a *metric*, obeys three properties. These metric axioms are as follows (Exhibit 5.1), where d_{ab} denotes the distance between objects a and b:

$$\begin{aligned} &1.\quad d_{ab} = d_{ba} \\ &2.\quad d_{ab} \geq 0 \quad \text{and} = 0 \text{ if and only if } a = b \\ &3.\quad d_{ab} \leq d_{ac} + d_{ca} \end{aligned} \tag{5.1}$$

Exhibit 5.1: *Illustration of the triangle inequality for distances in Euclidean space*



The first two axioms seem self-evident: the first says that the distance from *a* to *b* is the same as from *b* to *a*, in other words the measure is symmetric; the second says that distances are always positive except when the objects are identical, in which case the distance is necessarily 0. The third axiom, called the *triangle inequality*, may also seem intuitively obvious but is the more difficult one to satisfy. If we draw a triangle *abc* in our Euclidean world, for example in Exhibit 5.1, then it is obvious that the distance from *a* to *b* must be shorter than the sum of the distances via another point *c*, that is from *a* to *c* and from *c* to *b*. The triangle inequality can only be an equality if *c* lies exactly on the line connecting *a* and *b* (see the right hand sketch in Exhibit 5.1).

But there are many apparently acceptable measures of distance that do not satisfy this property: with those it would be theoretically possible to get a "route" from *a* to *b* via a third point *c* which is shorter than from *a* to *b* "directly". Because such measures that do not satisfy the triangle inequality are not true distances (in the mathematical sense) they are usually called *dissimilarities*.

Bray-Curtis dissimilarity

When it comes to species abundance data collected at different sampling locations, the *Bray-Curtis* (or *Sørensen*) *dissimilarity* is one of the most well-known ways of quantifying the difference between samples. This measure appears to be a very reasonable way of achieving this goal but it does not satisfy the triangle inequality axiom, and hence is not a true distance (we shall discuss the implications of this in later chapters when we analyse Bray-Curtis dissimilarities). To illustrate its definition, we consider again the count data for the last two samples of Exhibit 1.1, which we recall here:

	a	*b*	*c*	*d*	*e*	*Sum*
s29	11	0	7	8	0	*26*
s30	24	37	5	18	1	*85*

One of the assumptions of the Bray-Curtis measure is that the sampled areas or volumes are of the same size. This is because dissimilarity will be computed on raw counts, not on relative counts, so the fact that there is higher overall abundance at site s30 is part of the difference between these two samples – that is, "size" and "shape" of the count vectors will be taken into account in the measure.[1]

The computation involves summing the absolute differences between the counts and dividing this by the sum of the abundances in the two samples, denoted here by *b*:

$$b_{s29,s30} = \frac{|11-24| + |0-37| + |7-5| + |8-18| + |0-1|}{26+85} = \frac{63}{111} = 0.568$$

The general formula for calculating the *Bray-Curtis dissimilarity* between samples i and i' is as follows, supposing that the counts are denoted by n_{ij} and that their sample (row) totals are n_{i+}:

$$b_{ii'} = \frac{\sum_{j=1}^{J} |n_{ij} - n_{i'j}|}{n_{i+} + n_{i'+}} \tag{5.2}$$

This measure takes on values between 0 (for identical samples: $n_{ij} = n_{i'j}$ for all j) and 1 (samples completely disjoint; that is, when there is a nonzero abundance of a species in one sample, then it is zero in the other: $n_{ij} > 0$ implies $n_{i'j} = 0$) – hence it is often multiplied by 100 and interpreted as a percentage. Exhibit 5.2 shows part of the Bray-Curtis dissimilarities between the 30 samples (the caption points out a violation of the triangle inequality).

If the Bray-Curtis dissimilarity is subtracted from 100, a measure of *similarity* is obtained, called the *Bray-Curtis index*. For example, the similarity between sites s25 and s4 is 100 – 93.9 = 6.1%, which is the lowest amongst the values displayed in Exhibit 5.2; whereas the highest similarity is for sites s25 and s26: 100 – 13.7 = 86.3%. Checking back to the data in Exhibit 1.1 one can verify the similarity between sites s25 and s26, compared to the lack of similarity between s25 and s4.

[1] In fact, the Bray-Curtis dissimilarity can be computed on relative abundances, as we did for the chi-square distance, to take into account only "shape" differences – this point is discussed later. This version is often referred to as the relative Sørensen dissimilarity.

Exhibit 5.2: *Bray-Curtis dissimilarities, multiplied by 100, between the 30 samples of Exhibit 1.1, based on the count data for species* **a** *to* **e**. *Violations of the triangle inequality can be easily picked out: for example, from* s25 *to* s4 *the Bray-Curtis is 93.9, but the sum of the values "via* s6*" from* s25 *to* s6 *and from* s6 *to* s4 *is 18.6 + 69.2 = 87.8, which is shorter*

	s1	s2	s3	s4	s5	s6	· · ·	s24	s25	s26	s27	s28	s29
s2	45.7												
s3	29.6	48.1											
s4	46.7	55.6	46.7										
s5	47.7	34.8	50.8	78.6									
s6	52.2	22.9	52.2	69.2	41.9								
s7	45.5	41.5	49.1	87.0	21.2	50.9							
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱						
s25	70.4	39.3	66.7	93.9	52.9	18.6	· · ·	46.4					
s26	69.6	32.8	60.9	92.8	41.7	15.2	· · ·	39.3	13.7				
s27	63.6	38.1	63.6	93.3	38.2	21.5	· · ·	42.6	16.3	22.6			
s28	32.5	21.5	50.0	57.7	31.9	29.5	· · ·	30.9	41.8	47.5	34.4		
s29	43.4	35.0	43.4	54.5	31.2	53.6	· · ·	39.8	64.5	58.2	61.2	34.2	
s30	60.7	36.7	58.9	84.5	48.0	21.6	· · ·	40.8	18.1	25.3	23.6	37.7	56.8

Bray-Curtis dissimilarity versus chi-square distance

An ecologist would like some recommendation about whether to use Bray-Curtis or chi-square on a particular data set. It is not possible to make any absolute statement of which is preferable, but we can point out some advantages and disadvantages of each one. The advantage of the chi-square distance is that it is a true metric, while the Bray-Curtis dissimilarity violates the triangle inequality, which can be problematic when we come to analysing them later. The advantage of Bray-Curtis is that the scale is easy to understand: 0 means the samples are exactly the same, while 100 is the maximum difference that can be observed between two samples. The chi-square, on the other hand, has a maximum which depends on the marginal weights of the data set, and it is difficult to assign any substantive meaning to any particular value. If two samples have the same relative abundances, but different totals, then Bray-Curtis is positive, whereas chi-square is zero. As pointed out in a previous footnote in this chapter, Bray-Curtis dissimilarities can be calculated on the relative abundances (although conventionally the calculation is on raw counts), and in addition we could calculate chi-square distances on the raw counts, without "relativizing" them (although conventionally the calculation is on relative counts). This would make the comparison between the two approaches fairer.

So we also calculated Bray-Curtis on the relative counts and chi-square on the raw counts – Exhibit 5.3 shows parts of the four distance matrices, where the values in each triangular matrix have been strung out column-

wise (the column "site pair" shows which pair corresponds to the values in the rows).

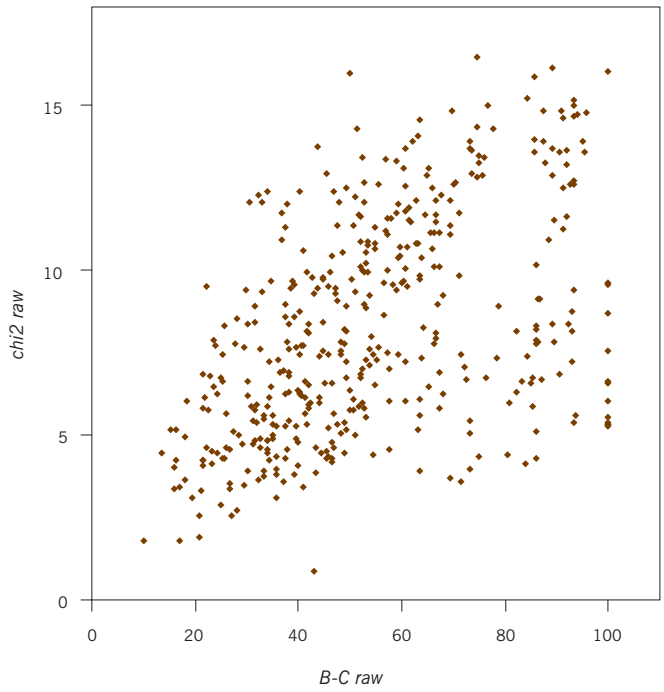
Site pair	*B-C raw*	*chi2 raw*	*B-C rel*	*chi2 rel*
(s2,s1)	45.679	7.398	48.148	1.139
(s3,s1)	29.630	3.461	29.630	0.855
(s4,s1)	46.667	4.146	50.000	1.392
(s5,s1)	47.692	5.269	50.975	1.093
(s6,s1)	52.212	10.863	53.058	1.099
(s7,s1)	45.455	4.280	46.164	1.046
(s8,s1)	93.333	5.359	92.593	2.046
(s9,s1)	33.333	5.462	40.741	0.868
(s10,s1)	40.299	6.251	36.759	0.989
(s11,s1)	35.714	4.306	36.909	1.020
(s12,s1)	37.500	5.213	39.762	0.819
(s13,s1)	57.692	5.978	59.259	1.581
(s14,s1)	63.265	5.128	59.091	1.378
(s15,s1)	20.755	1.866	20.513	0.464
(s16,s1)	85.714	13.937	80.960	1.700
(s17,s1)	100.000	5.533	100.000	2.258
(s18,s1)	56.897	11.195	36.787	0.819
(s19,s1)	16.923	1.762	11.501	0.258
(s20,s1)	33.333	3.734	31.987	0.800
⋮	⋮	⋮	⋮	⋮
(s23,s22)	34.400	7.213	25.655	0.688
(s24,s22)	61.224	9.493	35.897	0.897
(s25,s22)	23.567	7.855	25.801	0.617
s(24,s23)	34.177	4.519	16.401	0.340
s(25,s23)	37.681	11.986	37.869	1.001
(s25,s24)	56.757	13.390	44.706	1.142

Exhibit 5.3: *Various dissimilarities and distances between pairs of sites (count data from Exhibit 1.1).* B-C raw*: Bray Curtis dissimilarities on raw counts (usual definition and usage),* chi2 raw*: chi-square distances on raw counts,* B-C rel*: Bray-Curtis dissimilarities on relative counts,* chi2 rel*: chi-square distances on relative counts (usual definition and usage)*

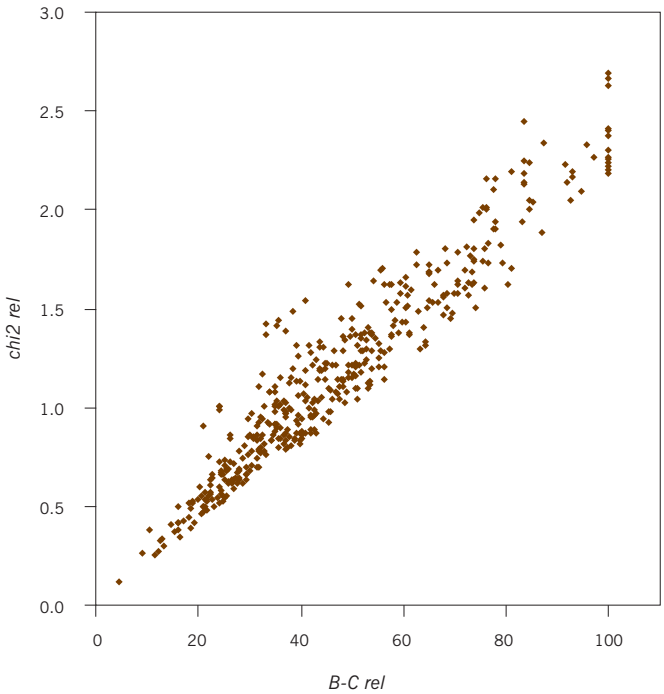
The scatterplots of the two comparable sets of measures are shown in Exhibit 5.4. Two features of these plots are immediately apparent: first, there is much better agreement between the two approaches when the counts have been relativized (plot (b)); and second, one can obtain 100% dissimilarity for the Bray-Curtis corresponding to different values of the chi-square distances: for example, in Exhibit 5.4(a) there are chi-square distances from approximately 5 to 16 corresponding to points above the tic-mark of 100 on the axis *B-C raw*.

Exhibit 5.4: *Graphical comparison of Bray-Curtis dissimilarities and chi-square distances for (a) raw counts, taking into account size and shape, and (b) relative counts, taking into account shape only*

(a)



(b)



This means that the measurement of shape is fairly similar in both measures, but the way they take size into account is quite different. A good illustration of this second feature is the measure between samples s1 and s17, which have counts as follows (taken from Exhibit 1.1):

	a	*b*	*c*	*d*	*e*	*Sum*
s1	0	2	9	14	2	*27*
s17	4	0	0	0	0	*4*

The Bray-Curtis dissimilarity is 100% because the two sets of counts are disjoint, whereas the chi-square distance is a fairly low 5.533 (see row (s17, s1) of Exhibit 5.3). This is because the absolute differences between the two sets are not large. If they were larger, say if we doubled both sets of counts, then the chi-square distance would increase accordingly whereas the Bray-Curtis would remain at 100%. It is by considering examples like these that researchers can obtain a feeling for the properties of these measures, in order to be able to choose the measure that is most appropriate for their own data.

L_1 distance (city-block)

When the Bray-Curtis dissimilarity is applied to relative counts, that is, row proportions $r_{ij} = n_{ij} / n_{i+}$, the row sums r_{i+} in the denominator of (5.2) are 1 for every row, so that the dissimilarity reduces to:

$$b_{ii'} = \frac{1}{2}\sum_{j=1}^{J}\left|r_{ij} - r_{i'j}\right| \tag{5.3}$$

The sum of absolute differences between two vectors is called the L_1 *distance,* or *city-block distance*. This is a true distance function since it obeys the triangle inequality, and as can be seen in Exhibit 5.4(b), agrees fairly well with the chi-square distance for the data under consideration. The reason why it is called the *city-block distance,* and also *Manhattan distance* or *"taxicab" distance,* can be seen in the two-dimensional illustration of Exhibit 5.5. Going from a point A to a point B is achieved by walking "around the block", compared to the Euclidean "straight line" distance. The city-block and Euclidean distances are special cases of the L_p distance, defined here between rows of a data matrix **X** (the Euclidean distance is obtained for $p = 2$):

$$d_{ii'}(p) = \left(\sum_{j=1}^{J}\left|x_{ij} - x_{i'j}\right|^{p}\right)^{1/p} \tag{5.4}$$

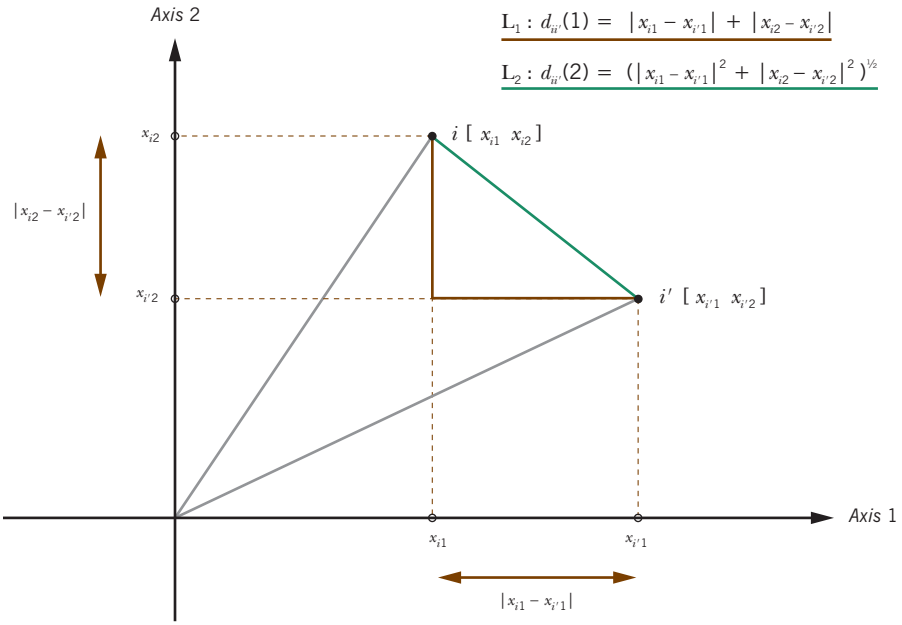


Exhibit 5.5: *Two-dimensional illustration of the L_1 (city-block) and L_2 (Euclidean) distances between two points i and i': the L_1 distance is the sum of the absolute differences in the coordinates, while the L_2 distance is the square root of the sum of squared differences*

Dissimilarity measures for presence—absence data

In Chapter 4 we considered the matching coefficient and the chi-square distance for categorical data in general, but there is a special case which is often of interest to ecologists: presence–absence, or dichotomous, data. When categorical variables have only two categories, there are a host of coefficients defined to measure inter-sample difference (see Bibliographical Appendix for references to this topic). Here we consider one example which is an alternative to the matching coefficient.

Exhibit 5.6 gives some data that we shall use again (in Chapter 7), concerning the presence–absence of 10 species in 7 samples. The inter-sample differences based on the matching coefficient would be obtained either by counting the matches or mismatches between the two samples. For example, between samples A and B there are 6 matches and 4 mismatches. Usually expressed relative to the number of variables (species) this would give a similarity value of 0.6 and a dissimilarity value of 0.4. But often in ecology it is possible to have very many species in the data set, up to 100 or more, and in each sample we find relatively few of these present. This makes the number of co-absences of species very high compared to the co-presences, but both count as matches. If co-absences are not informative, we can simply ignore them and calculate similarity in terms of co-presences. Furthermore, this co-presence count is expressed not relative to the total number of species but relative to the number of species present in at least one of the two

Exhibit 5.6:
Presence—absence data of 10 species in 7 samples

Samples	Species									
	sp1	*sp2*	*sp3*	*sp4*	*sp5*	*sp6*	*sp7*	*sp8*	*sp9*	*sp10*
A	1	1	1	0	1	0	0	1	1	1
B	1	1	0	1	1	0	0	0	0	1
C	0	1	1	0	1	0	0	1	0	0
D	0	0	0	1	0	1	0	0	0	0
E	1	1	1	0	1	0	1	1	1	0
F	0	1	0	1	1	0	0	0	0	1
G	0	1	1	0	1	1	0	1	1	0

samples under consideration. This is the definition of the *Jaccard index* for dichotomous data. Taking samples A and B of Exhibit 5.6 again, the number of co-presences is 4, we ignore the 2 co-absences, then we express 4 relative to 8, so the result is 0.5. In effect, the Jaccard index is the matching coefficient of similarity calculated for a pair of samples after eliminating all the species which are co-absent. The dissimilarity between two samples is – as before – 1 minus the similarity.

Here's another example, for samples C and D. This pair has 4 co-absences (for species 1, 7, 9 and 10), so we eliminate them. To get the dissimilarity we can count the mismatches – in fact, all the rest are mismatches – so the dissimilarity is $6/6 = 1$, the maximum that can be attained. Using the Jaccard approach we would say that samples C and D are completely different, whereas the matching coefficient would lead to a dissimilarity of 0.6 because of the 4 matched co-absences.

To formalize these definitions, the counts of matches and mismatches in a pair of samples are put into a 2×2 table as follows:

		Sample 2		
		1	0	
Sample 1	1	a	b	$a + b$
	0	c	d	$c + d$
		$a + c$	$b + d$	$a + b + c + d$

where a is the count of co-presences (1 and 1), b the count of mismatches where sample 1 has value 1 but sample 2 has value 0, and so on. The overall number of

matches is $a + d$, and mismatches $b + c$. The two measures of distance/dissimilarity considered so far are thus defined as:

$$\text{Matching coefficient dissimilarity:} \quad \frac{b+c}{a+b+c+d} = 1 - \frac{a+d}{a+b+c+d} \tag{5.5}$$

$$\text{Jaccard index dissimilarity:} \quad \frac{b+c}{a+b+c} = 1 - \frac{a}{a+b+c} \tag{5.6}$$

To give one final example, the correlation coefficient can be used to measure the similarity between two vectors of dichotomous data, and can be shown to be equal to:

$$r = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} \tag{5.7}$$

Hence, a dissimilarity can be defined as $1 - r$. Since $1 - r$ has a range from 0 (when $b = c = 0$, no mismatches) to 2 (when $a = d = 0$, no matches), a convenient measure between 0 and 1 is ½ $(1 - r)$.

Distances for mixed-scale data

When a data set contains different types of variables and it is required to measure inter-sample distance, we are faced with another problem of standardization: how can we balance the contributions of these different types of variables in an equitable way? We will demonstrate two alternative ways of doing this. The following is an example of mixed data (shown here are the data for four stations out of a set of 33:

Station	Continuous variables			Discrete variables	
	Depth	*Temperature*	*Salinity*	*Region*	*Substrate*
s3	30	3.15	33.52	Ta	Si/St
s8	29	3.15	33.52	Ta	Cl/Gr
s25	30	3.00	33.45	Sk	Cl/Sa
⋮	⋮	⋮	⋮	⋮	⋮
s84	66	3.22	33.48	St	Cl

Apart from the three continuous variables, depth, temperature and salinity there are the categorical variables of region (Tarehola, Skognes, Njosken and Storura), and substrate character (which can be any selection of clay, silt, sand, gravel and

stone). The fact that more than one substrate category can be selected implies that each category is a separate dichotomous variable, so that substrate consists of five different variables.

The first way of standardizing the continuous against the discrete variables is called *Gower's generalized coefficient of dissimilarity*. First we express the discrete variables as dummies and calculate the means and standard deviations of all variables in the usual way:

Station	Continuous variables			Sampled region				Substrate character				
	Depth	*Temperature*	*Salinity*	*Tarehola*	*Skognes*	*Njosken*	*Storura*	*Clay*	*Silt*	*Sand*	*Gravel*	*Stone*
s3	30	3.15	33.52	1	0	0	0	0	1	0	0	1
s8	29	3.15	33.52	1	0	0	0	1	0	0	1	0
s25	30	3.00	33.45	0	1	0	0	1	0	1	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s84	66	3.22	33.48	0	0	0	1	1	0	0	0	0
mean	58.15	3.086	33.50	0.242	0.273	0.242	0.242	0.606	0.152	0.364	0.182	0.061
sd	32.45	0.100	0.076	0.435	0.452	0.435	0.435	0.496	0.364	0.489	0.392	0.242

Notice that dichotomous variables (such as the substrate categories) are coded as a single dummy variable, not two, while polychotomous variables such as region are split into as many dummies as there are categories. The next step is to standardize each variable and multiply all the columns corresponding to dummy variables by $1/\sqrt{2} = 0.7071$, a factor which compensates for their higher variance due to the 0/1 coding:

Station	Continuous variables			Sampled region				Substrate character				
	Depth	*Temperature*	*Salinity*	*Tarehola*	*Skognes*	*Njosken*	*Storura*	*Clay*	*Silt*	*Sand*	*Gravel*	*Stone*
s3	–0.868	0.615	0.260	1.231	–0.426	–0.394	–0.394	–0.864	1.648	–0.526	–0.328	2.741
s8	–0.898	0.615	0.260	1.231	–0.426	–0.394	–0.394	0.561	–0.294	–0.526	1.477	–0.177
s25	–0.868	–0.854	–0.676	–0.394	1.137	–0.394	–0.394	0.561	–0.294	0.921	–0.328	–0.177
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
s84	0.242	1.294	–0.294	–0.394	–0.426	–0.394	1.231	0.561	–0.294	–0.526	–0.328	–0.177

Now distances are calculated between the stations using either the L_1 (city-block) or L_2 (Euclidean) metric. For example, using the L_1 metric and dividing the sum

of absolute differences by the total number of variables (12 in this example), the distances between the above four stations are given in the left hand table of Exhibit 5.7. Because the L_1 distance decomposes into parts for each variable, we can show the part of the distance due to the categorical variables, and the part due to the continuous variables. In this example the categorical variables are contributing more to the differences between the stations – the differences in the continuous variables are actually small if one looks at the original data, except for the distance between s84 and s25, where there is a bigger difference in the continuous variables, which then contribute almost the same (0.303) as the categorical ones (0.386).

Exhibit 5.7: *Distances between four stations based on the L_1 distance between their standardized and rescaled values, as described above. The distances are shown equal to the part due to the categorical (CAT.) variables plus the part due to the continuous (CONT.) variables*

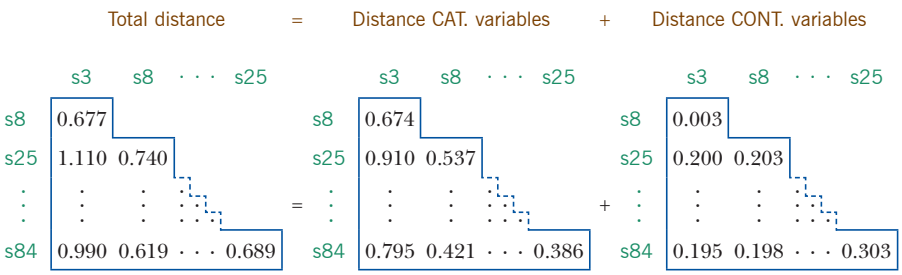


Exhibit 5.7 suggests the alternative way of combining different types of variables: first compute the distances which are the most appropriate for each set and then add them to one another. For example, suppose there are three types of data, a set of continuous variables, a set of categorical variables and a set of percentages or counts. Then compute the distance or dissimilarity matrices $\mathbf{D}_1$, $\mathbf{D}_2$ and $\mathbf{D}_3$ appropriate to each set of same-scale variables, and then combine these in a weighted average:

$$\mathbf{D} = \frac{w_1\mathbf{D}_1 + w_2\mathbf{D}_2 + w_3\mathbf{D}_3}{w_1 + w_2 + w_3} \tag{5.8}$$

Weights are a subjective but convenient inclusion, not only to account for the different scales in the distance matrices but also because there might be substantive reasons for down-weighting the distances for one set of variables, which might not be so important, or might suffer from high measurement error, for example. A default weighting system could be to make the variance of the distances the same in each matrix: $w_k = 1/s_k$, where s_k is the standard deviation of the distances in matrix $\mathbf{D}_k$.

A third possible way to cope with mixed-scale data such as these would be to fuzzy-code the continuous variables, as described in Chapter 3, and then apply a meas-

ure of dissimilarity appropriate to categorical data, with possible standardization as also discussed in Chapter 3. We shall make full use of this option in subsequent chapters and the two final case studies.

SUMMARY: Measures of distance between samples: non-Euclidean

1. The sum of absolute differences, or L_1 distance (or city-block distance), is an alternative to the Euclidean distance: an advantage of this distance is that it decomposes into contributions made by each variable (for the L_2 Euclidean distance, we would need to decompose the squared distance).
2. A well-defined distance function obeys the triangle inequality, but there are several justifiable measures of difference between samples that do not have this property: to distinguish these from true distances we often refer to them as dissimilarities.
3. The Bray-Curtis dissimilarity is frequently used by ecologists to quantify differences between samples based on abundance or count data. This measure is usually applied to raw abundance data, but can be applied to relative abundances just like the chi-square distance, in which case it is equivalent to the L_1, or city-block, distance. The chi-square distance can also be applied to the original abundances to include overall size differences in the distance measure.
4. A dissimilarity measure for presence–absence data is based on the Jaccard index, where co-absences are eliminated from the calculation, otherwise the measure resembles the matching coefficient.
5. Distances based on mixed-scale data can be computed after a process of standardization of all variables, using the L_1 or L_2 distances. Alternatively, distance matrices can be calculated for each set of same-scale variables and then these matrices can be linearly combined, optionally with user-defined weights.

CHAPTER 6

Measures of Distance and Correlation between Variables

In Chapters 4 and 5 we concentrated on distances between samples of a data matrix, which are usually the rows. We now turn our attention to the variables, usually the columns, and we can consider measures of distance and dissimilarity between these column vectors. More often, however, we measure the similarity between variables: this can be in the form of correlation coefficients or other measures of association. In this chapter we shall look at the geometric properties of variables, and various measures of correlation between them. In particular, we shall look at the geometric concept called a *scalar product,* which is highly related to the concept of Euclidean distance. The decision about which type of correlation function to use depends on the measurement scales of the variables, as we already saw briefly in Chapter 1. Finally, we also consider statistical tests of correlation, introducing the idea of permutation testing.

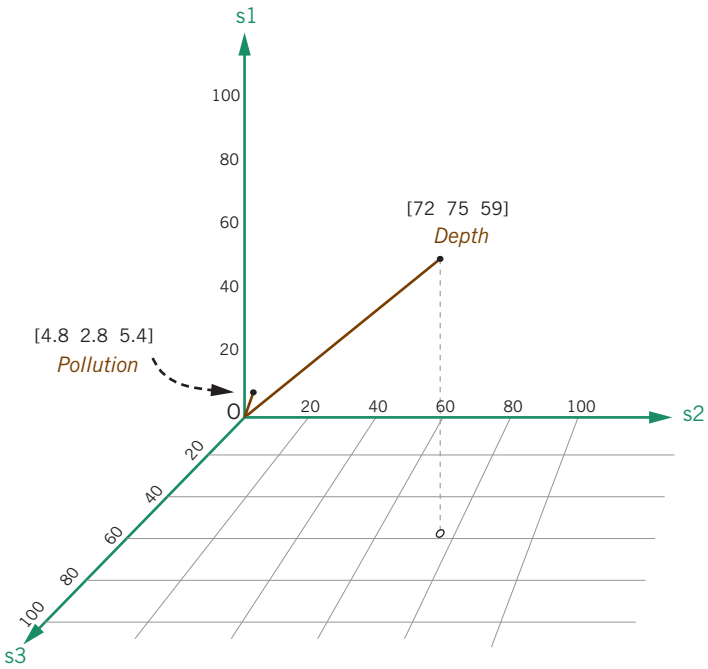
Contents

The geometry of variables

In Exhibits 4.3 and 5.5 in the previous chapters we have been encouraging the notion of samples being points in a multidimensional space. Even though we cannot draw points in more than three dimensions, we can easily extend the mathematical definitions of distance to samples for which we have J measurements, for any J. Now, rather than considering the samples, the rows of the data matrix,

Exhibit 6.1: *(a) Two variables measured in three samples (sites in this case), viewed in three dimensions, using original scales; (b) Standardized values; (c) Same variables plotted in three dimensions using standardized values. Projections of some points onto the "floor" of the* s2 – s3 *plane are shown, to assist in understanding the three-dimensional positions of the points*

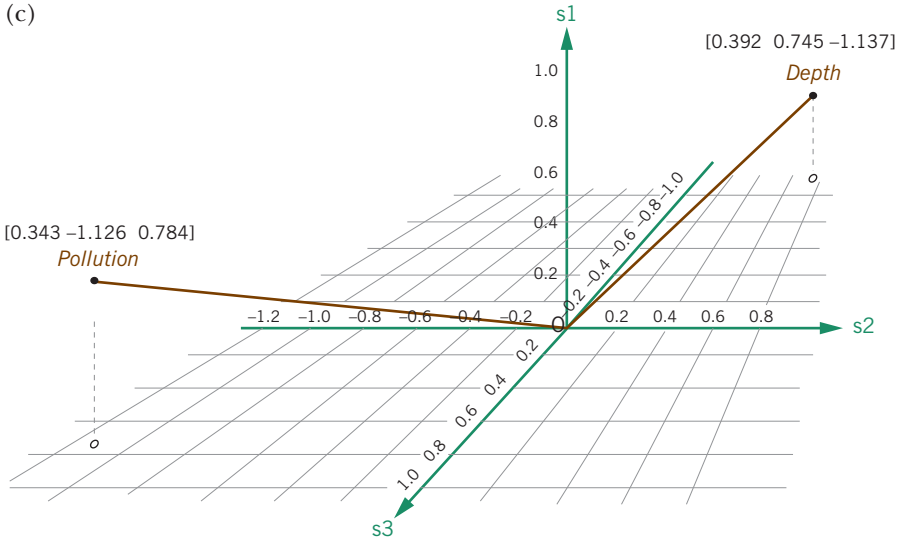
(a)



(b)

Site	*Depth*	*Pollution*
s1	0.392	0.343
s2	0.745	–1.126
s3	–1.137	0.784

(c)



we turn our attention to the variables (the columns of the data matrix) and their sets of observed values across the I samples. To be able to visualize two variables in I-dimensional space, we choose $I = 3$, since more than 3 is impossible to display or imagine. Exhibit 6.1(a) shows the variables depth and pollution according to the first three samples in Exhibit 1.1, with depth having values (i.e. coordinates) [72 75 59] and pollution [4.8 2.8 5.4]. Notice that it is the samples that now form the axes of the space. The much lower values for pollution compared to those for depth causes the distance between the variables to be dominated by this scale effect. Standardizing overcomes this effect – Exhibit 6.1(b) shows standardized values with respect to the mean and standard deviation of this sample of size 3 (hence the values here do not coincide with the standardized values in the complete data set, given in Exhibit 4.4). Exhibit 6.1(c) shows the two variables plotted according to these standardized values.

Correlation coefficient as an angle cosine

Exhibit 6.2 now shows the triangle formed by the two vectors in Exhibit 6.1(c) and the origin O, taken out of the three-dimensional space, and laid flat. From the coordinates of the points we can easily calculate the lengths of the three sides a, b and c of the triangle (where the sides a and b subtend the angle θ shown), so by using the cosine rule ($c^2 = a^2 + b^2 - 2ab\cos(\theta)$, which we all learned at school) we can calculate the cosine of the angle θ between the vectors, which turns out to be –0.798, exactly the correlation between pollution and depth (the angle is 142.9°). Notice that this is the correlation calculated in this illustrative sample of size 3, not in the original sample of size 30, where the estimated correlation is –0.396.

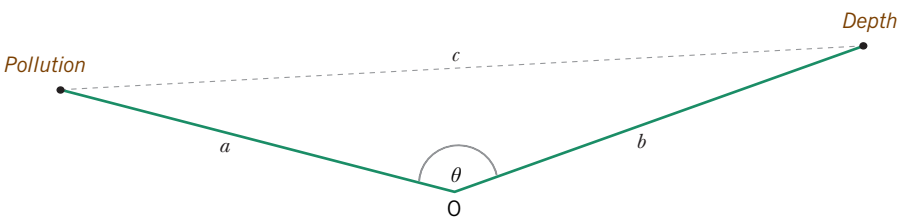


Exhibit 6.2: *Triangle of pollution and depth vectors with respect to origin* (O) *taken out of Exhibit 6.1(c) and laid flat*

Hence we have illustrated the result that the cosine of the angle between two standardized variables, plotted as vectors in the space of dimensionality I, the number of samples, is their correlation coefficient.

Correlation coefficient as a scalar product

But there is yet another way of interpreting the correlation coefficient geometrically. First we have to convert the standardized pollution and depth values to so-called *unit variables*. At present they are standardized to have variance 1, but a unit variable has sum of squares equal to 1 – in other words, its length is 1. Since the variance of I centred values is defined as $1/(I-1)$ times their sum

of squares, it follows that the sum of squares equals $(I-1)$ times the variance. By dividing the standardized values of pollution and depth in Exhibit 6.1(b) by $\sqrt{I-1}$, equal to $\sqrt{2}$ in this example, the standardized variables are converted to unit variables:

Site	*Depth*	*Pollution*
s1	0.277	0.242
s2	0.527	–0.796
s3	–0.804	0.554

[it can be checked that $0.277^2 + 0.527^2 + (-0.804)^2 = 0.242^2 + (-0.796)^2 + 0.554^2 = 1$]. The correlation coefficient then has the alternative definition as the sum of the products of the elements of the unit variables:

$$(0.242 \times 0.277) + (-0.796 \times 0.527) + (0.554 \times (-0.804)) = -0.798$$

i.e., the *scalar product:*

$$r_{jj'} = \sum_{i=1}^{I} x_{ij} x_{ij'} \tag{6.1}$$

where x_{ij} are the values of the unit variables.

The concept of a scalar product underlies many multivariate techniques which we shall introduce later. It is closely related to the operation of *projection*, which is crucial later when we project points in high-dimensional spaces onto lower-dimensional ones. As an illustration of this, consider Exhibit 6.3, which is the same as Exhibit 6.2 except that the sides *a* and *b* of the triangle are now shortened

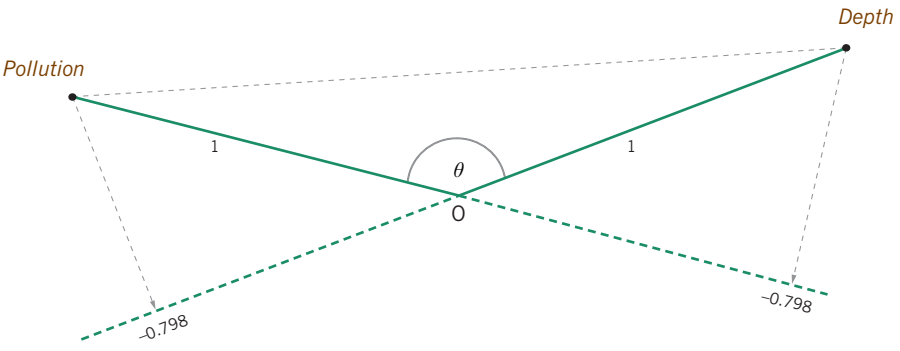


Exhibit 6.3: *Same triangle as in Exhibit 6.2, but with variables having unit length (i.e., unit variables. The projection of either variable onto the direction defined by the other variable vector will give the value of the correlation, cos(θ). (The origin O is the zero point – see Exhibit 6.1(c) – and the scale is given by the unit length of the variables.)*

to length 1 as unit variables (the subtended angle is still the same). The projection of either variable onto the axis defined by the other one gives the exact value of the correlation coefficient.

Distances based on correlation coefficients

When variables are plotted in their unit form as in Exhibit 6.3, the squared distance between the variable points is computed (again using the cosine rule) as $1 + 1 - 2\cos(\theta) = 2 - 2r$, where r is the correlation. In general, therefore, a distance $d_{jj'}$ between variables j and j' can be defined in terms of their correlation coefficient $r_{jj'}$ as follows:

$$d_{jj'} = \sqrt{2 - 2r_{jj'}} = \sqrt{2}\sqrt{1 - r_{jj'}} \tag{6.2}$$

where $d_{jj'}$ has a minimum of 0 when $r = 1$ (i.e., the two variables coincide), a maximum of 2 when $r = -1$ (i.e., the variables go in exact opposite directions), and $d_{jj'} = \sqrt{2}$ when $r = 0$ (i.e., the two variables are uncorrelated and are at right-angles to each other). For example, the distance between pollution and depth in Exhibit 6.3 is $\sqrt{2}\sqrt{1 - (-0.798)} = 1.896$.

An inter-variable distance can also be defined in the same way for other types of correlation coefficients and measures of association that lie between −1 and +1, for example the (*Spearman*) *rank correlation*. This so-called *nonparametric measure of correlation* is the regular correlation coefficient applied to the *ranks* of the data. In the sample of size 3 in Exhibit 6.1(a) pollution and depth have the following ranks:

Site	*Depth*	*Pollution*
s1	2	2
s2	3	1
s3	1	3

where, for example in the pollution column, the value 2.8 for site 2 is the lowest value, hence rank 1, then 4.8 is the next lowest value, hence rank 2, and 5.4 is the highest value, hence rank 3. The correlation between these two vectors is −1, since the ranks are indeed direct opposites – therefore, the distance between them based on the rank correlation is equal to 2, the maximum distance possible. Exhibit 6.4 shows the usual linear correlation coefficient, the Spearman rank correlation, and their associated distances, for the three variables based on their complete set of 30 sample values. This example confirms empirically that the results are more or less the same using ranks instead of the original values: that is, most of the correlation is in the ordering of the

Exhibit 6.4: *Correlations and associated distances between the three continuous variables of Exhibit 1.1: first the regular correlation coefficient on the continuous data, and second the rank correlation*

Correlation	*Depth*	*Pollution*	*Temperature*
Depth	1	−0.3955	−0.0034
Pollution	−0.3955	1	−0.0921
Temperature	−0.0034	−0.0921	1

Distance	*Depth*	*Pollution*	*Temperature*
Depth	0	1.6706	1.4166
Pollution	1.6706	0	1.4779
Temperature	1.4166	1.4779	0

Rank correlation	*Depth*	*Pollution*	*Temperature*
Depth	1	−0.4233	−0.0051
Pollution	−0.4233	1	−0.0525
Temperature	−0.0051	−0.0525	1

Distance	*Depth*	*Pollution*	*Temperature*
Depth	0	1.6872	1.4178
Pollution	1.6872	0	1.4509
Temperature	1.4178	1.4509	0

values rather than their actual numerical amounts. The rank correlation is also more *robust*, which means that it is less affected by unusual or extreme values in the data.

Distances between count variables

When it comes to the count data of Exhibit 1.1, the various distance measures considered in Chapter 5 can be used to measure distances between species. It makes little sense, however, to apply the chi-square distance or the Bray-Curtis dissimilarity to the raw data – these should be expressed as proportions, (i.e., relativized) with respect to their column sums. The two measures then turn out as in Exhibit 6.5, where the scatterplot shows them to be very similar, apart from their different scales, of course. The scatterplot is shown using the same horizontal and vertical scales as in Exhibit 5.4(b) in order to demonstrate that the spread of the distances between the columns is less than the corresponding spread between the rows.

Distances between categorical variables and between categories

Measures of distance between samples based on a set of dichotomous variables were defined on the basis of a 2×2 table of counts of matches and mismatches, and this same idea can be applied to the dichotomous variables based on their values across the samples: for example, the number of samples for which both variables were "present", and so on. Then the various measures of dissimilarity (5.5), (5.6) and (5.7) apply, in particular the one based on the correlation coefficient r. But after (5.7) we proposed that $1 - r$ would make a reasonable measure of dissimilarity (or $\frac{1}{2}(1 - r)$ to give it a range of 0 to 1). Now, based on our study of the geometry of variables in this chapter, a better choice would be $\sqrt{2}\sqrt{1-r}$ (or $\sqrt{1-r}/\sqrt{2}$ if again one prefers a value between 0 and 1), because this is a Euclidean distance and is therefore a true metric, whereas the previous definition turns out to be a squared Euclidean distance.

chi2	a	b	c	d
b	0.802			
c	1.522	1.407		
d	0.870	0.828	1.157	
e	1.406	1.550	1.855	1.430

B-C	a	b	c	d
b	28.6			
c	60.9	56.4		
d	32.9	33.5	41.4	
e	53.3	57.6	70.4	55.6

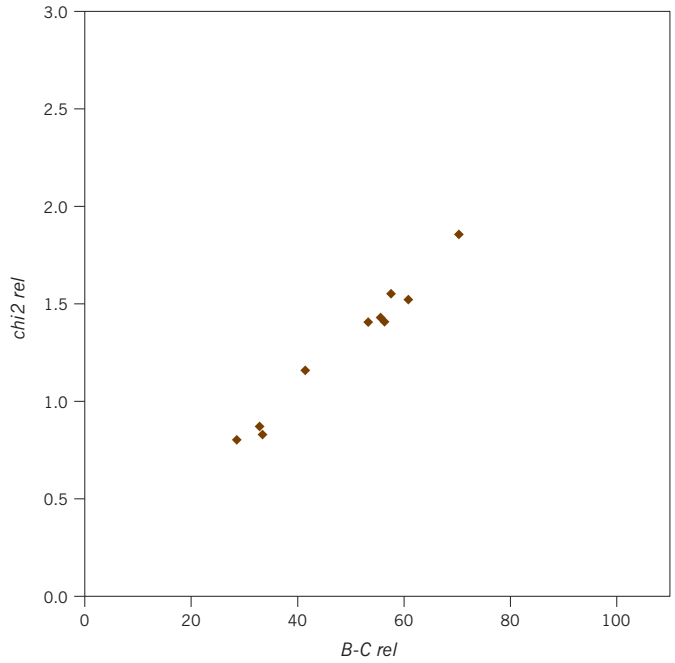


Exhibit 6.5: *Chi-square distances and Bray-Curtis dissimilarities between the five species variables, in both cases based on their proportions across the samples (i.e., removing the effect of different levels of abundances for each species). The two sets of values are compared in the scatterplot*

For categorical variables with more than two categories, there are two types of distances in question: distances between variables, and distances between categories of variables, both not easy to deal with. At the level of the variable, we can define a measure of similarity, or association, and there are quite a few different ways to do this. The easiest way is to use a variation on the chi-square statistic computed on the cross-tabulation of the pair of variables. In our introductory data of Exhibit 1.1 there is only one categorical variable, but let us categorize depth into three categories: low, medium and high depth, by simply cutting up the range of depth into three parts, so there are 10 sites in each category – this is the crisp coding of a continuous variable described in Chapter 3. The cross-tabulation of depth and sediment is then given in Exhibit 6.6 (notice that the counts of the depth categories are not exactly 10 each, because of some tied values in the depth data).

Exhibit 6.6: *Cross-tabulation of depth, categorized into three categories, and sediment type, for the data of Exhibit 1.1*

		Sediment		
		C	S	G
Depth	Low	6	5	0
	Medium	3	5	1
	High	2	1	7

The chi-square statistic for this table equals 15.58, but this depends on the sample size, so an alternative measure divides the chi-square statistic by the sample size, 30 in this case, to obtain the so-called *mean-square contingency coefficient*, denoted by $\phi^2 = 15.58/30 = 0.519$. We will rediscover ϕ^2 in later chapters, since it is identical to what is called the *inertia* in correspondence analysis, which measures the total variance of a data matrix.

Now ϕ^2 measures how similar the variables are, but we need to invert this measure somehow to get a measure of dissimilarity. The maximum value of ϕ^2 turns out to be one less than the number of rows or columns of the cross-tabulation, whichever is the smaller: in this case there are 3 rows and 3 columns, so one less than the minimum is 2. You can verify that if a 3×3 cross-tabulation has only one nonzero count in each row (likewise in each column), that is there is perfect association between the two variables, then $\phi^2 = 2$. So a dissimilarity could be defined as $2 - \phi^2$, equal to 1.481 in this example.

There are many alternatives, and we only mention one more. Since the maximum of ϕ^2 for an $I \times J$ cross-tabulation is $\min\{I-1, J-1\}$, we could divide ϕ^2 by this maximum. The so-called *Cramer's V coefficient* does this but also takes the square root of the result:

$$V = \sqrt{\frac{\phi^2}{\min\{I-1, J-1\}}} \tag{6.3}$$

This coefficient has the properties of a correlation coefficient, but is never negative because the idea of negative correlation for categorical variables has no meaning: variables are either not associated or have some level of (positive) association. Once again, subtracting V from 1 would give an alternative measure of dissimilarity.

Distances between categories

For a categorical variable such as sediment in Exhibit 1.1, measuring the distance between the categories C, S and G makes no sense at all, because they never co-occur in this data set. In this sense their correlations are always –1, and they

are all at maximum distance apart. We can only measure their similarity in their relation to other variables. For example, in Exhibit 6.6 the sediment categories are cross-tabulated with depth, and this induces a measure of distance between the sediment types. An appropriate measure of distance would be the chi-square distance between the column profiles of the table in Exhibit 6.6, which gives the following distances:

chi2	C	S
S	0.397	
G	1.525	1.664

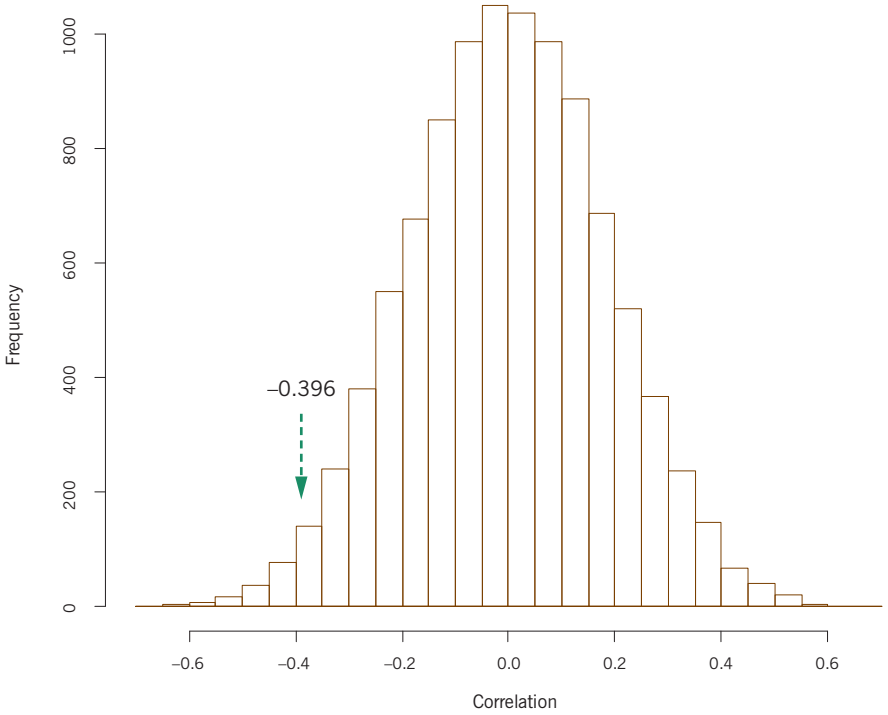
This shows that G is the most dissimilar to the other two sediment types, in terms of their respective relations with depth, which can be seen clearly in Exhibit 6.6.

Testing correlations: an introduction to permutation testing

Researchers usually like to have some indication of statistical significance of the relationships between their variables, so the question arises how to test the correlation coefficients and dissimilarity measures that have been described in this chapter. Tests do exist of some of these statistical quantities, for example there are several ways to test for the correlation coefficient, assuming that data are normally distributed, or with some other known distribution that lends itself to working out the distribution of the correlation. An alternative way of obtaining a *p*-value is to perform permutation testing, which does not rely on knowledge of the underlying distribution of the variables. The idea is simple, all that one needs is a fast computer and the right software, and this presents no problem these days. Under the null hypothesis of no correlation between the two variables, say between depth and pollution, the pairing of observations in the same sample is irrelevant, so we can associate any value of depth, say, with any value of pollution. Thus we generate many values of the correlation coefficient under the null hypothesis by permuting the values across the samples. This process generates what is called the *permutation distribution*, and the exact permutation distribution can be determined if we consider all the possible permutations of the data set. But even with a sample of size 30, the 30! possible permutations are too many to compute, so we estimate the distribution by using a random sample of permutations.

This is exactly what we did in Chapter 1 to estimate the *p*-value for the correlation between pollution and depth. A total of 9,999 random permutations were made of the 30 observations of one of the variables, say depth (with the order of pollution kept fixed), and Exhibit 6.7 is the histogram of the resulting correlations, with the actually observed correlation of −0.396 indicated. The *p*-value is the probability of the observed result and any more extreme ones, and since this is a two-sided

Exhibit 6.7: *Estimated permutation distribution for the correlation between pollution and depth (data from Exhibit 1.1), for testing the null hypothesis that the correlation is zero. The observed value of –0.396 is shown, and the permutation test consists in counting how many of the simulated correlations have an absolute value greater than or equal to 0.396*



testing problem, we have to count how many of the 10,000 permutations (including the observed one, this is why we generate 9,999) are equal or more extreme than 0.396 in absolute value. It turns out there are 159 values more extreme on the negative side (≤ -0.396) and 137 on the positive side (≥ 0.396), giving an estimated p-value of $296/10{,}000 = 0.0296$. This is very close to the p-value of 0.0305, which is calculated from the classical t-test for the correlation coefficient:

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = -2.279\,,$$

corresponding to a two-sided p-value of 0.0305, using the t-distribution with $n-2$ degrees of freedom ($n = 30$ here).

SUMMARY: Measures of distance and correlation between variables

1. Two variables that have been centred define two directions in the multidimensional space of the samples.
2. The cosine of the angle subtended by these two direction vectors is the classic linear correlation coefficient between the variables.

3. There are advantages in having the set of observations for each variable of unit length. This is obtained by dividing the standardized variables by $\sqrt{I-1}$, where I is the sample size, so that the sum of squares of their values is equal to 1. These are then called *unit variables*.

4. The distance d between the points defined by the unit variables is $d = \sqrt{2}\sqrt{1-r}$, where r is the correlation coefficient. Conversely, the correlation is $r = 1 - \frac{1}{2}d^2$.

5. Distances between count variables can be calculated in a similar way to distances between samples for count data, with the restriction that the variables be expressed as profiles, that is as proportions relative to their total across the samples.

6. Distances between dichotomous categorical variables can be calculated as before for distances between samples based on dichotomous variables.

7. Distances between categories of a polychotomous variable can only be calculated in respect of the relation of this variable with another variable.

8. Permutation tests are convenient computer-based methods of arriving at p-values for quantifying the significance of relationships between variables.

VISUALIZING DISTANCES AND CORRELATIONS

CHAPTER 7

Hierarchical Cluster Analysis

In Part 2 (Chapters 4 to 6) we defined several different ways of measuring distance (or dissimilarity as the case may be) between the rows or between the columns of the data matrix, depending on the measurement scale of the observations. As we remarked before, this process often generates tables of distances with even more numbers than the original data, but we will now show how this step actually simplifies our understanding of the data. Distances between objects can be visualized in many simple and evocative ways. In this chapter we shall consider a graphical representation of a matrix of distances which is perhaps the easiest to understand – a dendrogram, or tree – where the objects are joined together in a hierarchical fashion from the closest, that is most similar, to the furthest apart, that is the most different. The method of hierarchical cluster analysis is best explained by describing the algorithm, or set of instructions, which creates the dendrogram result. In this chapter we demonstrate the application of hierarchical clustering on a small example and then list the different variants of the method that are possible.

Contents

The algorithm for hierarchical clustering

As an example we shall consider again the small data set in Exhibit 5.6: seven samples on which 10 species are indicated as being present or absent. In Chapter 5 we discussed two of the many dissimilarity coefficients that are possible to define between the samples: the first based on the matching coefficient and the second based on the Jaccard index. The latter index counts the number of "mismatches" between two samples after eliminating the species that do not occur in either of the pair. Exhibit 7.1 shows the complete table of inter-sample dissimilarities based on the Jaccard index.

Exhibit 7.1: *Dissimilarities, based on the Jaccard index, between all pairs of seven samples in Exhibit 5.6. Both the lower and upper triangles of this symmetric dissimilarity matrix are shown here (the lower triangle is outlined as in previous tables of this type)*

Samples	A	B	C	D	E	F	G
A	0.000	0.500	0.429	1.000	0.250	0.625	0.375
B	0.500	0.000	0.714	0.833	0.667	0.200	0.778
C	0.429	0.714	0.000	1.000	0.429	0.667	0.333
D	1.000	0.833	1.000	0.000	1.000	0.800	0.857
E	0.250	0.667	0.429	1.000	0.000	0.778	0.375
F	0.625	0.200	0.667	0.800	0.778	0.000	0.750
G	0.375	0.778	0.333	0.857	0.375	0.750	0.000

The first step in the hierarchical clustering process is to look for the pair of samples that are the most similar, that is the closest in the sense of having the lowest dissimilarity – this is the pair B and F, with dissimilarity equal to 0.2.[1] These two samples are then joined at a level of 0.2 in the first step of the dendrogram, or clustering tree (see the first diagram of Exhibit 7.3, and the vertical scale of 0 to 1 which calibrates the level of clustering). The point at which they are joined is called a *node*.

We are basically going to keep repeating this step, but the only problem is how to calculate the dissimilarity between the merged pair (B,F) and the other samples. This decision determines what type of hierarchical clustering we intend to perform, and there are several choices. For the moment, we choose one of the most popular ones, where the dissimilarity between the merged pair and the others will be the maximum of the pair of dissimilarities in each case. For example, the dissimilarity between B and A is 0.500, while the dissimilarity between F and A is 0.625. Hence we choose the maximum of the two, 0.625, to quantify the dissimilarity between (B,F) and A. Continuing in this way we obtain a new dissimilarity matrix Exhibit 7.2.

The process is now repeated: find the smallest dissimilarity in Exhibit 7.2, which is 0.250 for samples A and E, and then cluster these at a level of 0.25, as shown in the second figure of Exhibit 7.3. Then recompute the dissimilarities between the merged pair (A,E) and the rest to obtain Exhibit 7.4. For example, the dissimilarity between (A,E) and (B,F) is the maximum of 0.625 (A to (B,F)) and 0.778 (E to (B,F)).

[1] Recall what this value means: five species occurred in at least one of the samples B and F, four occurred in both, while one was present in B but not in F, so the Jaccard index of dissimilarity is $1/5 = 0.2$.

Samples	A	(B,F)	C	D	E	G
A	0.000	0.625	0.429	1.000	0.250	0.375
(B,F)	0.625	0.000	0.714	0.833	0.778	0.778
C	0.429	0.714	0.000	1.000	0.429	0.333
D	1.000	0.833	1.000	0.000	1.000	0.857
E	0.250	0.778	0.429	1.000	0.000	0.375
G	0.375	0.778	0.333	0.857	0.375	0.000

Exhibit 7.2: *Dissimilarities calculated after* B *and* F *are merged, using the "maximum" method to recompute the values in the row and column labelled* (B,F)

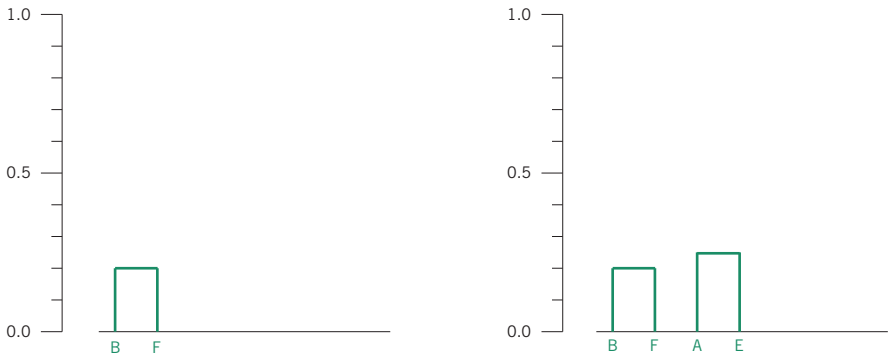


Exhibit 7.3: *First two steps of hierarchical clustering of Exhibit 7.1, using the "maximum" (or "complete linkage") method*

Samples	(A,E)	(B,F)	C	D	G
(A,E)	0.000	0.778	0.429	1.000	0.375
(B,F)	0.778	0.000	0.714	0.833	0.778
C	0.429	0.714	0.000	1.000	0.333
D	1.000	0.833	1.000	0.000	0.857
G	0.375	0.778	0.333	0.857	0.000

Exhibit 7.4: *Dissimilarities calculated after* A *and* E *are merged, using the "maximum" method to recompute the values in the row and column labelled* (A,E)

In the next step the lowest dissimilarity in Exhibit 7.4 is 0.333, for C and G – these are merged, as shown in the first diagram of Exhibit 7.6, to obtain Exhibit 7.5. Now the smallest dissimilarity is 0.429, between the pair (A,E) and (B,G), and they are shown merged in the second diagram of Exhibit 7.6. Exhibit 7.7 shows the last two dissimilarity matrices in this process, and Exhibit 7.8 the final two steps of the construction of the dendrogram, also called a *binary tree* because at each step two objects (or clusters of objects) are merged. Because there are 7 objects to be clustered, 6 nodes are formed in the sequential process (i.e., one

Exhibit 7.5: *Dissimilarities calculated after* C *and* G *are merged, using the "maximum" method to recompute the values in the row and column labelled* (C,G)

Samples	(A,E)	(B,F)	(C,G)	D
(A,E)	0.000	0.778	0.429	1.000
(B,F)	0.778	0.000	0.778	0.833
(C,G)	0.429	0.778	0.000	1.000
D	1.000	0.833	1.000	0.000

Exhibit 7.6: *The third and fourth steps of hierarchical clustering of Exhibit 7.1, using the "maximum" (or "complete linkage") method. The point at which objects (or clusters of objects) are joined is called a node*

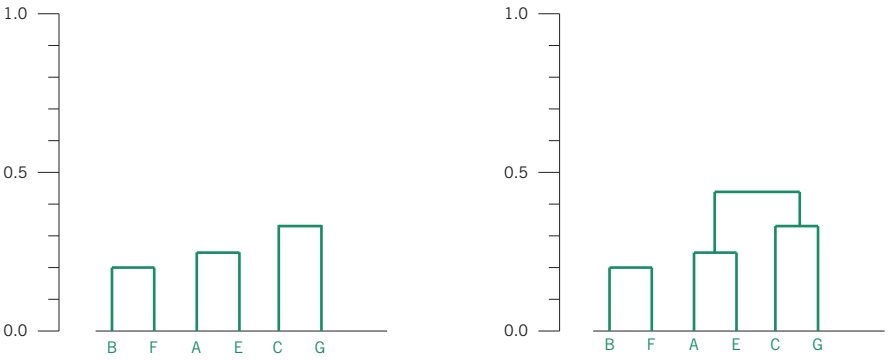


Exhibit 7.7: *Dissimilarities calculated after* C *and* G *are merged, using the "maximum" method to recompute the values in the row and column labelled* (C,G)

Samples	(A,E,C,G)	(B,F)	D
(A,E,C,G)	0.000	0.778	1.000
(B,F)	0.778	0.000	0.833
D	1.000	0.833	0.000

Samples	(A,E,C,G,B,F)	D
(A,E,C,G,B,F)	0.000	1.000
D	1.000	0.000

less than the number of objects) to arrive at the final tree where all objects are in a single cluster.

Cutting the tree

The final dendrogram on the right of Exhibit 7.8 is a compact visualization of the dissimilarity matrix in Exhibit 7.1, computed on the presence-absence data of Exhibit 5.6. Interpretation of the structure of data is made much easier now – we can see that there are three pairs of samples that are fairly close, two of these pairs [(A,E) and (C,G)] are in turn close to each other, while the sin-

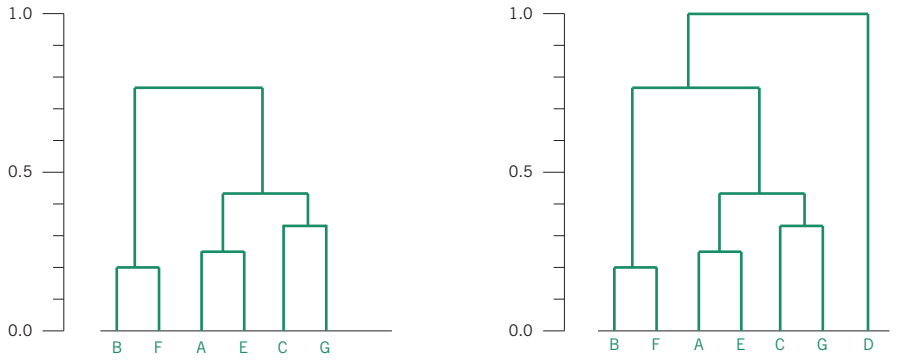


Exhibit 7.8:
The fifth and sixth steps of hierarchical clustering of Exhibit 7.1, using the "maximum" (or "complete linkage") method. The dendrogram on the right is the final result of the cluster analysis

gle sample D separates itself entirely from all the others. Because we used the "maximum" method, all samples clustered below a particular level of dissimilarity will have inter-sample dissimilarities less than that level. For example, 0.5 is the point at which samples are exactly as similar to one another as they are dissimilar, so if we look at the clusters of samples below 0.5 – i.e., (B,F), (A,E,C,G) and (D) – then within each cluster the samples have more than 50% similarity, in other words more than 50% co-presences of species. The level of 0.5 also happens to coincide in the final dendrogram with a large jump in the clustering levels: the node where (A,E) and (C,G) are clustered is at level of 0.429, while the next node where (B,F) is merged is at a level of 0.778. This is thus a very convenient level to *cut* the tree to define clusters. If the branches are cut at 0.5, we are left with the three clusters of samples (B,F), (A,E,C,G) and (D), which can be labelled types 1, 2 and 3 respectively. In other words, we have created a categorical variable, with three categories, and the samples are classified as follows:

A	B	C	D	E	F	G
2	1	2	3	2	1	2

Checking back to Chapter 2, this is exactly the objective which we described in the lower right hand corner of the multivariate analysis scheme (Exhibit 2.2) – to reveal a categorical latent variable which underlies the structure of a data set.

Maximum, minimum and average clustering

Two crucial choices are necessary when deciding on a cluster analysis algorithm. The first is to decide how to quantify dissimilarities between two clusters: in the above illustration the Jaccard index was used. The second choice is how to update the matrix of dissimilarities at each step of the clustering:

in the algorithm described above the maximum value of the between-cluster dissimilarities was chosen. This is called the *maximum* method, also known as *complete linkage* cluster analysis, because a cluster is formed when all the dissimilarities ("links") between pairs of objects in the cluster are less then a particular level. There are several alternatives to complete linkage as a clustering criterion, and we only discuss two of these: minimum and average clustering.

The *minimum* method goes to the other extreme and forms a cluster when only one pair of dissimilarities (not all) is less than a particular level – this is known as *single linkage* cluster analysis. So at every updating step we choose the minimum of the two distances and two clusters of objects can be merged when there is a single close link between them, irrespective of the other inter-object distances. In general, this is not a suitable choice for most applications, because it can lead to clusters that are quite heterogeneous internally, and the usual object of clustering is to obtain homogeneous clusters.

The *average* method is an attractive compromise where at each step the dissimilarity between two clusters is the average of all the pairwise dissimilarities between the clusters. This is also dubbed UPGMA clustering in the literature, a rather laborious abbreviation for "Unweighted Pair-Group Method using Averages". Notice that this is not merely taking the arithmetic average of the two dissimilarity values available at each step for the updating, but rather taking into account the size of each cluster as well. For example, if one cluster contains two cases and another three, then there are six dissimilarities on which the (unweighted) arithmetic average needs to be computed – this is equivalent to weighting the clusters by their sample sizes in the updating of the average dissimilarity.

Validity of the clusters

If a cluster analysis is performed on a data matrix, a set of clusters can always be obtained, even if there is no actual grouping of the objects, in this case the samples. So how can we evaluate whether the three clusters in this example are not just any three groups which we might have obtained on random data with no underlying structure? We shall consider this question more closely in Chapter 17 when we deal with statistical inference and give one possible answer to this problem using a permutation test. Apart from this statistical issue there is also the substantive issue of where to cut the tree. In this example, the three clusters established by complete linkage were such that within each cluster all inter-sample dissimilarities were all less than 0.5. It would be difficult to justify cutting the tree at a higher level, because that would mean that some pairs of samples in a cluster would be more dissimilar than similar. But this substantive cut-off level of 0.5 is particular to the Jaccard index and

to measures like the Bray-Curtis that have scales with a clear interpretation. If one uses an Euclidean distance, or chi-square distance, for example, their scales are not clearly interpretable and we have to resort to deciding on the number of clusters by an inspection of the tree, cutting the branches where there is a big jump in the level of successive nodes, or by a statistical criterion described in Chapter 17.

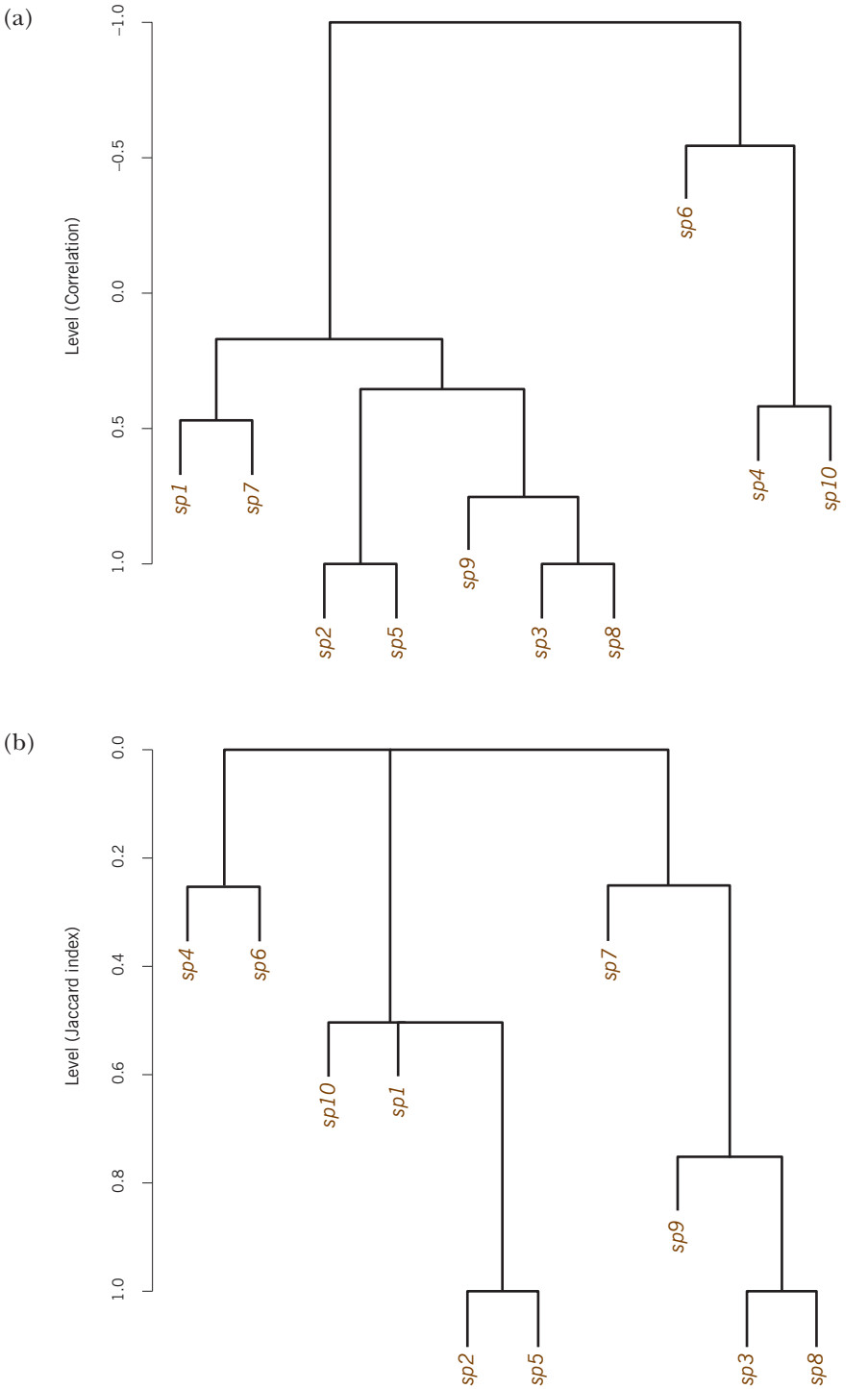
Clustering correlations on variables

Just like we clustered samples, so we can cluster variables in terms of their correlations. In this case it may be more intuitive to show cluster levels as the correlation, the measure of similarity, rather than the reverse measure of dissimilarity. The similarity based on the Jaccard index can also be used to measure association between species – the index counts the number of samples that have both species of the pair, relative to the number of samples that have at least one of the pair. Exhibit 7.9 shows the cluster analyses based on these two alternatives, for the columns of Exhibit 5.6. There are two differences here compared to previous dendrograms: first, the vertical scale is descending as the tree is being constructed from the bottom up, since the clustering is on similarities, and second, the R function `hclust` used to perform the clustering places the species labels at a constant distance below their initial clustering levels. The fact that these two trees are so different is no surprise: the first one based on the correlation coefficient takes into account the co-absences, which strengthens the correlation, while the second does not. Both have the pairs (*sp2*,*sp5*) and (*sp3*,*sp8*) at maximum similarity of 1 because these are identically present and absent across the samples. Species *sp1* and *sp7* are similar in terms of correlation, due to co-absences – *sp7* only occurs in one sample, sample E, which also has *sp1*, a species which is absent in four other samples. Notice in Exhibit 7.9(b) how species *sp10* and *sp1* both join the cluster (*sp2*,*sp5*) at the same level (0.5).

Clustering a large data set

The more objects there are to cluster, the more complex becomes the result, and we would not generally apply this method to a set of more than 100 objects, say. In Exhibit 4.5 we showed part of the matrix of standardized Euclidean distances between the 30 sites of Exhibit 1.1, and Exhibit 7.10 shows the hierarchical clustering of this distance matrix, using compete linkage. There are two obvious places where we can cut the tree, at about level 3.4, which gives four clusters, or about 2.7, which gives six clusters. Which one we should choose depends on substantive as well as statistical grounds. For example, the six-cluster solution splits a large group on the right hand side of the dendrogram into two; if this is usefully interpreted as two different sets of sites in the context of the study, then the six-cluster solution would be preferred. But there is also the statistical issue about whether that split can be considered random or not, which is what we will deal with in Chapter 17.

Exhibit 7.9: *Complete linkage cluster analyses of similarities between species: (a) r, the correlation coefficient between species); (b) Jaccard similarity index between species. The* R *function* `hclust` *which calculates the dendrograms places the object (species) labels at a constant distance below its clustering level*



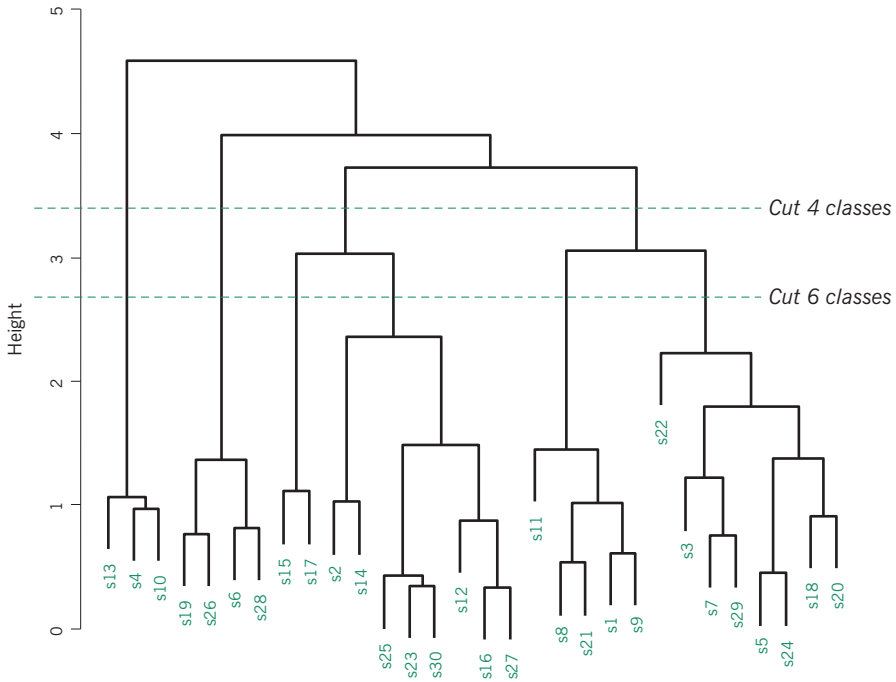


Exhibit 7.10: *Complete linkage cluster analyses of the standardized Euclidean distances of Exhibit 4.5*

If the data set consists of a very large set of objects, say in the thousands, then nonhierarchical clustering can be used, as described in the next chapter, but the number of clusters desired has to be specified in advance. A hybrid approach could be to initially reduce (using nonhierarchical clustering) the very large set into a large number of very compact clusters, for example reducing a set of 1,000 objects to about 100 clusters with an average of 10 objects each, and then performing hierarchical clustering on the 100 clusters.

SUMMARY: Hierarchical cluster analysis

1. Hierarchical cluster analysis of n objects is defined by a stepwise algorithm performed on a matrix of appropriately chosen dissimilarities or distances previously computed between the objects. Two objects are merged at each step, the two which have the least dissimilarity or distance.

2. As the algorithm proceeds, objects become clusters of objects, so we need to decide how to measure dissimilarity/distance between clusters. Some standard options are the maximum dissimilarity (complete linkage) between the objects of each cluster, the minimum dissimilarity (single linkage) or the average dissimilarity (average linkage).

3. The results of a hierarchical clustering are graphically displayed in the form of a dendrogram (or binary tree), with $n - 1$ nodes.

4. In order to form discrete groups, the branches of this tree are cut at a level where there is a lot of "space", that is where there is a relatively large jump in levels of two consecutive nodes.

5. Either rows or columns of a matrix can be clustered – in each case we choose the appropriate dissimilarity measure. It is more intuitive to show the results of clustering of variables in terms of their similarity measure, for example their correlations.

CHAPTER 8

Ward Clustering and *k*-means Clustering

This chapter continues the theme of cluster analysis, first with a popular alternative to the hierarchical clustering methods presented in Chapter 7 – Ward clustering. This method is based on the same concepts as analysis of variance (ANOVA), so we shall give a brief introduction to ANOVA, specifically to the definitions of between-group and within-group variance, to motivate Ward clustering. Exactly the same concepts are used in a different clustering algorithm called *k-means clustering*. This form of clustering, which is an example of "nonhierarchical" clustering, is particularly useful when a very large number of objects need to be clustered, where the dendrogram would be so big that it becomes too burdensome to visualize and interpret. In this situation, all we really want is a partitioning of the objects into a set of groups. Nonhierarchical clustering algorithms such as *k*-means do not result in a dendrogram – the user specifies in advance how many groups are being sought (the *k* of *k*-means) and the final result is the allocation of each object to a group so that the groups are as internally homogeneous as possible. This measure of internal homogeneity is the same as in Ward clustering, hence our treatment of these two methods together in this chapter.

Contents

Analysis of variance (ANOVA)

To introduce the concepts inherent in Ward and nonhierarchical clustering, it is worthwhile to recall analysis of variance, abbreviated as ANOVA. ANOVA is concerned with testing the difference between means of a continuous variable observed in different groups. As an example, we can use

the continuous variable "pollution" and the categorical variable "sediment" from Exhibit 1.1, where sediment divides the sample into three groups: clay (11 sites), sand (11 sites) and gravel (8 sites). In the middle of Exhibit 1.5 the box-and-whisker plot for pollution shows the medians of the three groups, or subsamples, and their dispersions between first and third quartiles – here we shall be concerned with the means and variances in each subsample. Exhibit 8.1 shows an alternative graphical display of the same data, where each value is shown on the pollution scale and is coded according to its respective sedimentary group. The means of each group are indicated as well as the mean pollution of all 30 sites. In ANOVA, the separateness of the three groups is measured by how far the means are away from the overall mean, taking into account the size of the groups, the so-called *between-group sum of squares* BSS:

$$\mathrm{BSS} = \sum_{g=1}^{G} n_g (\bar{x}_g - \bar{x})^2 \tag{8.1}$$

where G = number of groups, n_g = the sample size in the g-th group, $\bar{x}_g$ is the g-th group mean and $\bar{x}$ is the overall mean. In this particular case the calculation gives a value of BSS = 37.6. In isolation this value tells nothing about how separate the groups are, because if the three groups of points were more tightly dispersed about their respective means, we would get the same value of BSS even though the groups appear more separate. The dispersion of the observations around their respective group means thus needs to be taken into account, and this is calculated by the *within-group sum of squares* WSS:

$$\mathrm{WSS} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} (x_{ig} - \bar{x}_g)^2 \tag{8.2}$$

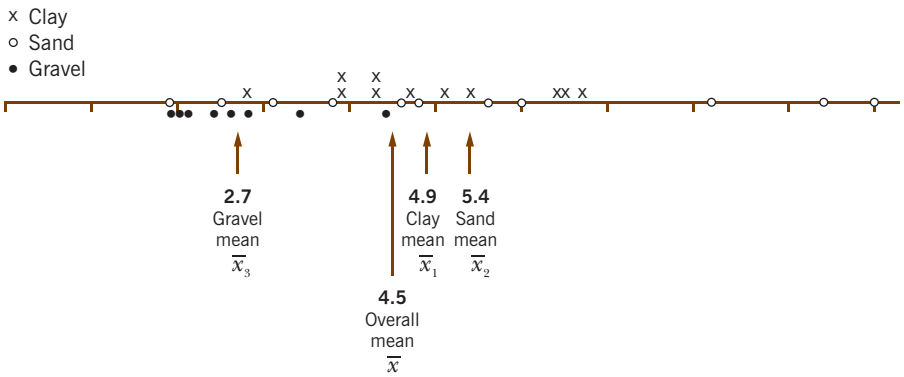


Exhibit 8.1: *Representation of the 30 values of pollution (see Exhibit 1.1), coded for the three sediment types. The means (to one decimal place) of the three subsets of data are indicated, as well as the overall mean (compare this graphical representation with that of the middle plot of Exhibit 1.5, where the medians and quartiles are displayed)*

which in this example is equal to WSS = 95.4. The beauty of using sums of squares is that BSS and WSS add up to the *total sum of squares*, TSS:

$$\text{TSS} = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{ig} - \bar{x})^2 \tag{8.3}$$

that is,

$$\text{BSS} + \text{WSS} = \text{TSS} \tag{8.4}$$

in this case: 37.6 + 95.4 = 133.0. TSS in (8.3) is the sum of squared deviations of all the observations from the overall mean, which measures the dispersion of all the data points around the overall mean. WSS in (8.2), on the other hand, is measuring the dispersion of the observations in the groups around their own respective group means, so WSS must be less than TSS. Notice that TSS divided by $n - 1$ (= 29) is the usual sample variance (of all the observations), while each of the G summations in (8.2), divided by its respective $n_g - 1$, is the variance of the g-th group. Furthermore, (8.1) divided by $G - 1$ is the variance of the $G = 3$ means weighted by their respective group sizes. The term *analysis of variance* derives from the fact that (8.4) implies this decomposition of variance into parts between and within the groups. In order to test whether there is significant separation of the groups, or whether the observed separation is compatible with random variation in the data, the values of BSS and WSS are combined into the classic F-statistic.[1] This *F-test* gives a p-value of 0.0112, indicating significant differences between the sediment groups in terms of pollution. We could also perform a permutation test, to be described in Chapter 17, which estimates the p-value as 0.0106, very close to that of the F-test.

Looking for the optimal solution

In ANOVA the grouping variable is prescribed (sediment type in the above example), but in cluster analysis we are looking for a grouping variable in the data. In the one-dimensional example of Exhibit 8.1, suppose we have no classification of the 30 values, what would be the optimal clustering of the data into three groups? Optimality could be defined as maximizing the ratio BSS/TSS, which is equivalent to optimizing any increasing function of that ratio, for example BSS itself (since TSS is fixed), or BSS/WSS, or the F-statistic defined in the footnote. Because there is only one variable and a fairly small sample size, we can investigate every pair of cutpoints that separates the data set into three groups (clearly,

[1] The classical test in ANOVA for testing differences between means is the F-test, where $F = \frac{BSS/(G-1)}{WSS/(n-G)}$ has the F-distribution with $G - 1$ and $n - G$ "degrees of freedom". The observed value $F = (37.6/2)/(95.4/27) = 5.32$ has an associated p-value of 0.0112, which is very close to the p-value of 0.0106 of the permutation test.

for maximal separation of the three groups, each group is defined as a contiguous interval on the pollution scale). There are $29 \times 28/2 = 406$ pairs of cutpoints, and the maximum BSS/TSS turns out to be 0.867 obtained when the three groups are defined as (i) the first 14 values; (ii) the next 13 values; (iii) the last 3 values, as shown in green in the lower part of Exhibit 8.2. This exhaustive way of looking for a given number of clusters is actually the nonhierarchical clustering treated in a later section, but we do it here to contrast with Ward clustering, which is the hierarchical version of this search.

Ward clustering in one dimension

Ward clustering also tries to maximize BSS (or, equivalently, minimize WSS) but does it at each step of a hierarchical clustering like the ones described in Chapter 7. So we start with single objects and look for the pair that are the "closest", in terms of keeping the WSS as small as possible (and thus the BSS as large as possible), and proceed stepwise in this way until a dendrogram is constructed. Exhibit 8.2 shows the dendrogram constructed by Ward clustering, and the associated three-cluster solution using a cutting of the tree at about level 10, giving clusters of 9, 14 and 7 sites. Notice that the Ward procedure does not necessarily find the optimal solution – this is because the hierarchical clustering is stepwise and every merging of clusters depends on what has happened previously. For example, the values 6.0; 6.4; 6.5 and 6.7 join the smaller cluster on the right formed by the three top values 8.2; 9.4 and 10.0, whereas in the optimal solution these three top values form a

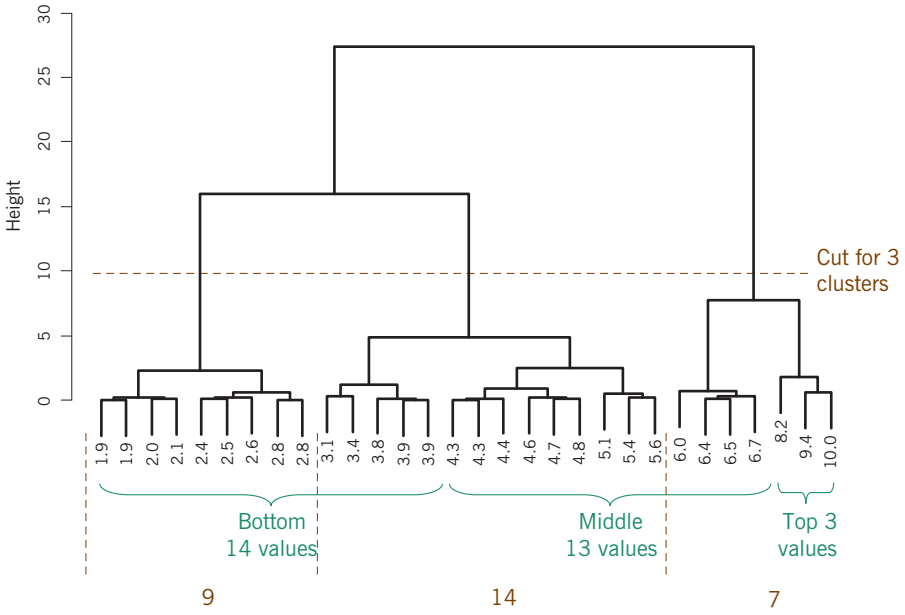


Exhibit 8.2: *Ward clustering of the 30 sites in Exhibit 1.1 according to the single variable "pollution", showing the cutpoint for a 3-cluster solution (partitioning of 9; 14 and 7 values, shown by vertical dashed lines), with between-to-total sum of squares ratio, BSS/TSS = 0.825. The sites are labelled by their pollution values. The curly brackets show the globally optimal 3-cluster solution (partitioning of 14; 13 and 3 values) for which BSS/TSS = 0.867*

cluster alone. The Ward clustering solution, with BSS/TSS = 0.825, is actually quite far from the optimal partitioning of the 30 sites, where BSS/TSS = 0.867, computed above.

Ward clustering in several dimensions

The type of exhaustive search that we could do above in one dimension, looking at all possible cutpoints, becomes much more difficult when the data are multidimensional: for example, for the three-dimensional (continuous) environmental data of Exhibit 1.1, we would have to consider all pairs of planes dividing the sample into three subsamples. Hierarchical Ward clustering, however, is still very simple to execute, even though it is unlikely to find the optimal solution. The algorithm proceeds in the same way as for the unidimensional case, with the BSS and TSS measures using squared distances in multidimensional space, which are the natural generalizations of the squared differences in one dimension. For example, BSS in (8.1) becomes, in the multidimensional version:

$$\text{BSS} = \sum_{g=1}^{G} n_g d(\bar{\mathbf{x}}_g, \bar{\mathbf{x}})^2 \tag{8.5}$$

where $\bar{\mathbf{x}}_g$ and $\bar{\mathbf{x}}$ are now the g-th mean vector and overall mean vector, respectively. When there are more than one variable, then the issue of standardization becomes important when defining the distance, as explained in Chapter 4. Exhibit 8.3 shows the Ward clustering of the 30 samples based on Euclidean distance using the three standardized variables (depth, pollution and temperature) – part of the distance matrix has been given in Exhibit 4.5.

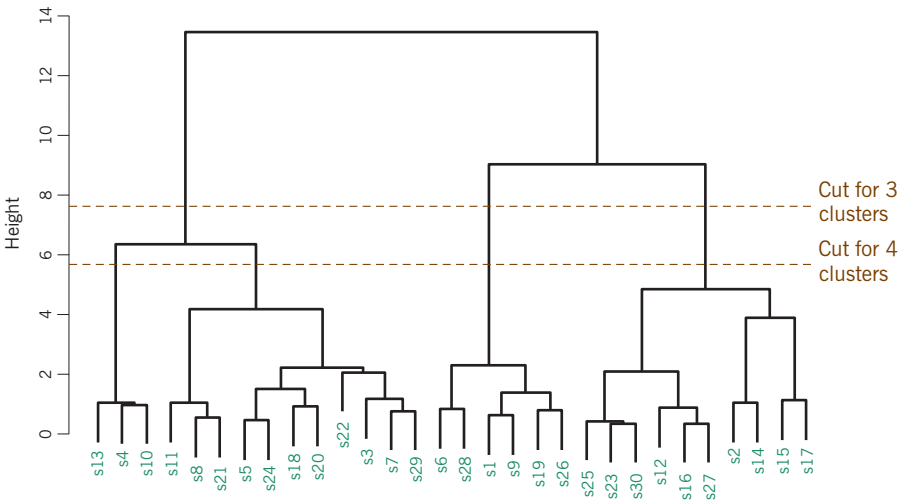


Exhibit 8.3: *Ward clustering of the 30 sites in Exhibit 1.1 according to the three variables depth, pollution and temperature, using standardized Euclidean distances (Exhibit 4.5). Cuts are shown which give three and four clusters*

Notice how different this result is in appearance from the complete linkage clustering of Exhibit 7.10, although on the left in both dendrograms we can see that the cluster of three sites (s13,s4,s10) corresponding to the three highest pollution values remains a separate cluster in both analyses. In the next section we shall show that the cluster solutions are, in fact, very similar. We can again perform a permutation test for clusteredness, to be described more fully in Chapter 17. For example, for the four-cluster solution, the permutation test estimates a p-value of 0.465, so there is no evidence of clustering in these data, and the analysis simply serves to partition the sites into four groups in their three-dimensional spatial continuum.

Comparing cluster solutions

One cluster analysis, which yields p clusters, can be compared to another cluster analysis on the same data, giving q clusters, by cross-tabulating the categories from the two solutions. For example, let us compare the four-category complete linkage solution from Exhibit 7.10 ($p = 4$) with the four-cluster Ward solution from Exhibit 8.3 ($q = 4$), leading to the following cross-tabulation of the 30 sites:

		Ward clustering			
		Cluster 1	*Cluster 2*	*Cluster 3*	*Cluster 4*
Complete linkage clustering	*Cluster 1*	2	0	11	0
	Cluster 2	0	10	0	0
	Cluster 3	0	0	0	3
	Cluster 4	4	0	0	0

Apart from the two sites in the first cell of the table, all the sites fall into the same clusters in both solutions – the two solutions would agree perfectly if these two (identified as sites s1 and s9) were either in cluster 4 of the complete linkage solution, or in cluster 3 of the Ward solution. There are several ways to measure the agreement between the two solutions, as summarized in such a cross-tabulation; for example, Cramer's V statistic given in (6.3) – which is equal to 1 for perfect agreement, is equal to 0.925 in this example.

Nonhierarchical clustering by k-means

Instead of constructing a dendrogram, nonhierarchical clustering searches for a prescribed number of clusters in the data. We shall describe the most popular of the nonhierarchical algorithms, called *k-means clustering*. The k refers to the specified number of groups we are looking for in the data set, and *means* refers to the fact that in each iteration of the algorithm objects are allocated to the closest group mean. The k-means algorithm proceeds as follows, where n objects need to be clustered into k groups, and we have a distance function between any pair of

objects (which should be of the Euclidean type, weighted or unweighted, for the decomposition (8.4) to be valid):

1. Choose k objects at random as starting *seeds*; or use k prespecified objects as seeds.

2. Calculate the distances of all n objects to the k seeds, and allocate each object to its closest seed – this gives a first clustering of the objects into k groups.

3. Calculate the means of the k clusters, used as seeds for the next iteration.

4. Repeat steps 2. and 3. until convergence, that is when there is no change in the group allocation from one iteration to the next.

It can be proved that when the distance function is of the Euclidean type, optionally weighted, then the value of the between-groups sum of squares (BSS) must increase from one iteration to the next. Since BSS cannot be higher than TSS, the algorithm must converge, but there is no guarantee that the convergence is at the global optimum – we say that the algorithm converges at a local optimum, which could be the global one, we simply do not know.

The k-means algorithm is very fast and is generally used for very large data sets, typically found in the social sciences, for example looking for clusters of attitudes in a political survey of thousands of people. If a large ecological data set is encountered, this is a way to find some simple structure in the sample units. In the clustering of the 30 sites described previously, in terms of the single variable pollution, we did know the global optimum (BSS/TSS = 0.867, which is the highest possible value for this example – see Exhibit 8.2) because we could do an exhaustive search of all three-cluster solutions. We already saw that Ward clustering gave a nonoptimal solution, with BSS/TSS = 0.825. Even though k-means is usually used for much bigger data sets and many variables, we applied it to the same example, specifying three clusters as before. The result was BSS/TSS = 0.845, which is an improvement over the hierarchical clustering solution, but still not the global optimum. In k-means clustering the starting set of seeds is quite crucial to the result – unless we have some prior information on what constitutes good seeds to start growing clusters, the initial seeds are chosen randomly. So it is recommended to use several random sets of starting seeds and then take the best result. When we repeated k-means clustering using 10 different random starts, we did indeed find the optimal solution with BSS/TSS = 0.867.

Similarly, we can compare the k-means result, after several random starts, with the Ward clustering on the three-dimensional data. The four-cluster solution

obtained in the latter result, shown in Exhibit 8.3, gives a BSS/TSS ratio of 0.637. The best *k*-means solution, after 10 random starts, has an improved BSS/TSS equal to 0.648. So it seems, just after these few examples, that if one is interested in obtaining only a partition of the objects, then *k*-means clustering with several random starts does perform better than hierarchical Ward clustering. It does no harm, however, to do both and check which gives the better solution.

Weighting the objects in Ward and *k*-means clustering

The central concepts of Ward and *k*-means clustering are the measures BSS, WSS and TSS, where the objective is to maximize BSS, or equivalently minimize WSS because they add up to TSS, which is a constant. In the definitions (8.1), (8.2) and (8.3) of these measures, each object is counted, or weighted, equally. But in some situations (we shall see one of these when we treat correspondence analysis) we would like to count some objects differently from others, that is weight the objects differentially. If $w_1, w_2, \ldots, w_n$ denote positive weights assigned to the n objects, then (8.1)–(8.3) can be generalized as:

$$\text{BSS} = \sum_{g=1}^{G} w_g (\bar{x}_g - \bar{x})^2 \tag{8.6}$$

where w_g is the total weight of the objects in the g-th group: $w_g = \sum_{i=1}^{n_g} w_i$.

$$\text{WSS} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} w_i (x_{ig} - \bar{x}_g)^2 \tag{8.7}$$

$$\text{TSS} = \sum_{g=1}^{G} \sum_{i=1}^{n_g} w_i (x_{ig} - \bar{x})^2 \tag{8.8}$$

The equally weighted versions used before are thus a simple case when $w_i = 1$. The multidimensional equivalents – for example, BSS in (8.5) – are generalized in a similar fashion.

SUMMARY: Ward clustering and *k*-means clustering

1. Ward clustering is a hierarchical cluster analysis where the criterion for merging two clusters at each node of the tree is to maximize the separation of the new cluster's mean from the means of the other clusters. The separation between clusters is measured by the between-group sum of squares (BSS).
2. Equivalently, the criterion is based on minimizing the dispersion within the newly combined cluster. The dispersion within clusters is measured by the within-group sum of squares (WSS).
3. BSS and WSS sum to a constant, the total sum of squares (TSS). Thus, maximization of BSS is equivalent to minimization of WSS.

4. *k*-means clustering is a nonhierarchical cluster analysis based on exactly the same criteria as Ward clustering, with the difference that a solution is sought by an iterative procedure which successively allocates the set of observations to a set of *k* seeds, where *k* is the number of clusters specified by the user.

5. The initial seeds are *k* observations randomly chosen, or specified by the user, from which the algorithm can start to allocate observations to their nearest seed, providing a first clustering of the observations. Mean points in each cluster are calculated, which provide the seeds for the next iteration, and this process is repeated until there is no change in the clustering from one iteration to the next.

6. If interest is just in finding a set of clusters rather than visualizing the complete clustering process, then *k*-means clustering seems to find better solutions, but the analysis should be repeated several times with different random sets of initial seeds.

7. Both Ward clustering and *k*-means can be generalized to include observation weights, which give observations varying importance in the cluster analysis.

CHAPTER 9

Multidimensional Scaling

In Chapter 7 a square matrix of distances or dissimilarities was visualized in the form of a dendrogram, trying to establish groups of points with relatively small within-group distances but relatively large between-group distances. Samples are not usually clustered naturally but are more often spread around continuously in the multidimensional space of the variables. Strongly correlated variables imply a certain amount of redundancy in the data, which means that less dimensions than the number of variables are required to describe the sample positions. Multidimensional scaling (MDS) is an alternative way of visualizing a distance or dissimilarity matrix, with the objective of representing the samples in a low-dimensional space, usually two- or three-dimensional, reproducing as closely as possible the inter-sample proximities (either distances or dissimilarities). The method is thus attempting to make a spatial map of the data to facilitate interpretation of the positions of the samples relative to one another. Since our intuitive understanding of a map is through the physical concept of a Euclidean distance, it will be an issue whether the sample proximities are Euclidean or not. Usually Euclidean-type distances will be mapped by so-called *metric* scaling methods, for example classical (metric) MDS, while non-Euclidean ones will be mapped by *nonmetric* methods.

Contents

A simple exact distance matrix

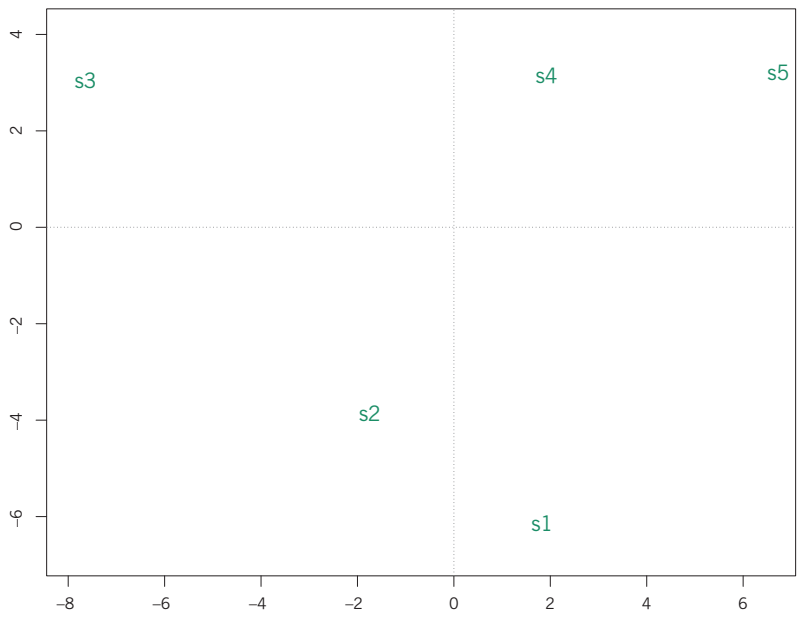
To introduce how MDS works in a simple way, consider the following matrix **D** of distances between five samples, numbered s1 to s5:

$$\mathbf{D} = \begin{array}{c} \\ \text{s1} \\ \text{s2} \\ \text{s3} \\ \text{s4} \\ \text{s5} \end{array} \begin{array}{c} \begin{array}{ccccc} \text{s1} & \text{s2} & \text{s3} & \text{s4} & \text{s5} \end{array} \\ \begin{bmatrix} 0 & 4.24 & 13.23 & 9.33 & 10.58 \\ 4.24 & 0 & 9.11 & 7.94 & 11.05 \\ 13.23 & 9.11 & 0 & 9.59 & 14.39 \\ 9.33 & 7.94 & 9.59 & 0 & 4.80 \\ 10.58 & 11.05 & 14.39 & 4.80 & 0 \end{bmatrix} \end{array} \tag{9.1}$$

First, we do not have an immediate way of telling whether this is a true distance matrix, but we can easily check the first two properties of a distance [see the properties in (5.1)], so at least we know that this is a dissimilarity matrix. A more laborious exercise would be to go through all triplets of samples (and there are 10 of these) to satisfy ourselves that the triangular inequality is indeed satisfied, for example, for samples s1, s2 and s3, with inter-sample values of 4.24; 13.23 and 9.11, the sum of any two is always greater than the third. In fact, we will see now that these five samples can be perfectly displayed in a two-dimensional map, reproducing exactly the above distance matrix.

Performing a so-called *classical MDS* (to be explained below) on the matrix **D**, the following map of the five samples in Exhibit 9.1 is obtained.

Exhibit 9.1: *Classical multidimensional scaling solution in two dimensions of the matrix* **D**, *using the* R *function* `cmdscale`



It can be verified that the distances between the samples are exactly those in the distance matrix **D**. Putting this another way, if we were given the coordinates of the five points in Exhibit 9.1 we could quite easily compute the distance matrix **D**, but MDS does the reverse, it starts with a distance matrix and constructs the map. The output of the MDS analysis is (1) the set of coordinates of the points i on each of the dimensions k of the solution, which we denote by f_{ik}, gathered in a matrix **F**; and (2) the parts of variance on each dimension, denoted by λ_k for the k-th dimension. For this example here are those results:

$$\mathbf{F} = \begin{array}{c} \\ s1 \\ s2 \\ s3 \\ s4 \\ s5 \end{array} \begin{array}{c} \begin{array}{cc} dim1 & dim2 \end{array} \\ \begin{bmatrix} 1.618 & -6.042 \\ -1.955 & -3.755 \\ -7.880 & 3.166 \\ 1.711 & 3.285 \\ 6.506 & 3.345 \end{bmatrix} \end{array} \qquad [\lambda_1 \quad \lambda_2] = \begin{array}{c} \begin{array}{cc} dim1 & dim2 \end{array} \\ [11.380 \quad 8.260] \end{array} \tag{9.2}$$

Computing the Euclidean distances between the rows of **F** will lead to exactly the same matrix of distances in **D**.

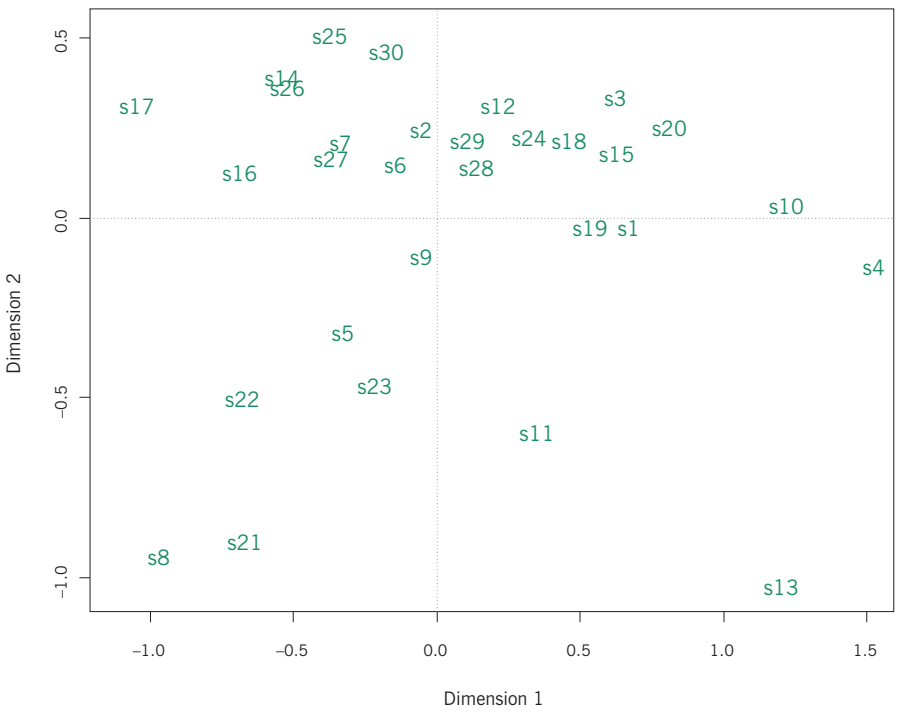
Classical MDS

The classical MDS procedure used to pass from the distance matrix **D** to the map in Exhibit 9.1 works as follows. For n points, the maximum dimensionality is one less, $n - 1$. For example, two points lie on a line (one-dimensional), three points lie in a plane (two-dimensional), but the three points could also lie on a line and be one-dimensional, and so on. In their $(n - 1)$–dimensional space, using an eigenvalue-eigenvector routine, we can identify the principal dimensions of the set of points, in descending order of importance. This order is determined by the eigenvalues, with the highest eigenvalue and associated eigenvector indicating dimension 1, the second dimension 2 and so on. In fact, the eigenvalues quantify the variance explained by each dimension. In the above example of five points, the dimensionality is at most 4, but it turns out that the third and fourth eigenvalues are zero, which means that the points are exactly two-dimensional. The total variance of the points, that is the quantification of their dispersion in the two-dimensional space, is the sum of the two eigenvalues, 11.38 + 8.26 = 19.64, with the first dimension accounting for 11.38/19.64 = 0.579, or 57.9%, of the total, and the second dimension the remaining 42.1%.

In practice: more complex proximity matrices

From now on we shall often call distance and dissimilarity matrices collectively as *proximity* matrices, and distinguish between them where necessary. In practice, you will never have a proximity matrix that is so simple as to be exactly two-dimensional. Let's take two examples of matrices we have encountered already, first the matrix of chi-square distances between the 30 samples in Chapter 4 – see Exhibit 4.7. Applying classical MDS to this matrix, the following eigenvalues are obtained: 12.37; 5.20; 3.83; 2.23, and all the remaining ones are zeros. This indicates that the distances are four-dimensional,[1] while in theory they could have any dimensionality up to 29. Exhibit 9.2 shows the MDS map of the chi-square distances between the 30 samples with respect to the first two dimensions. According to the eigenvalues, the total variance is 12.37 + 5.20 + 3.83 + 2.23 = 23.63, of which 12.37/23.37, or 52.4%, is accounted for by dimension 1 and 5.20/23.63, or 22.0%, is accounted for by dimension

Exhibit 9.2: *Classical multidimensional scaling solution in two dimensions of the matrix of chi-square distances of Exhibit 4.7. The percentages of variance on the horizontal and vertical axes are 52.4% and 22.0% respectively*



[1] The chi-square distances were computed on only five abundance values per sample, and Euclidean-type distances in this situation would usually be of dimensionality 5. So then why is the dimensionality equal to 4? This is because the chi-square distances are based on the relative abundances and since the relative abundances for each sample always add up to fixed value, 1, only four values are *free* and the fifth is one minus the sum of the other four. This is one of the properties inherent in correspondence analysis, treated in Chapter 13.

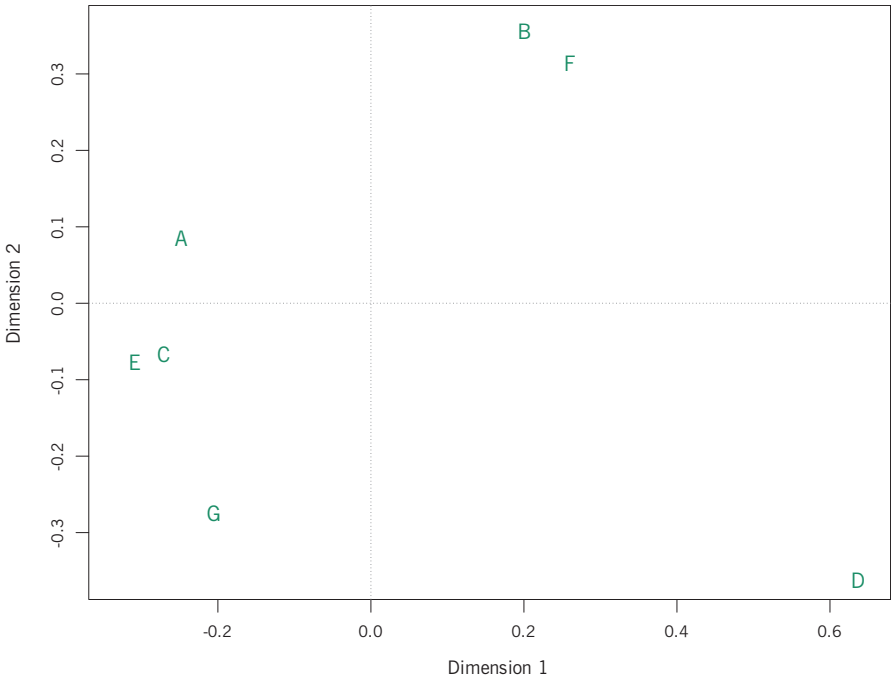
2, totalling 74.4% for the map as a whole. It can be seen in Exhibit 9.2 that the first dimension shows more variance than the second.

Exhibit 9.2 is an approximate display of the matrix of chi-square distances, but it is the best one can do in a two-dimensional display according to the optimization criterion inherent in classical scaling, namely to maximize explained distance variance. Each dimension of the display can be thought of as a variable trying to explain the variance in the distance matrix, like independent variables in a regression. These dimensions are uncorrelated, so their separate percentages of variance can be simply added to give the cumulative percentage of variance explained for by the set of dimensions. Hence, the percentage 74.4% of variance explained can be interpreted just like an R^2 in regression analysis.

Now as a second example we apply the same procedure to the matrix of Jaccard proximity values in Exhibit 7.1 between seven samples, which in theory has a dimensionality no higher than 6. The eigenvalues that emerge are: 0.786; 0.452; 0.148; 0.037; 0.000; –0.002; –0.030. Indeed, six of them are nonzero, but there are two negative eigenvalues, indicating that this proximity matrix is not Euclidean (hence the inclusion of the Jaccard index in Chapter 5 on non-Euclidean dissimilarity functions). So in this case it is impossible to represent these proximities in any Euclidean space, and the negative eigenvalues give an idea of how much of the variance cannot be displayed. The part that is Euclidean is the sum of the positive eigenvalues: 0.786 + 0.452 + 0.148 + 0.037 = 1.423, while the part that cannot be displayed is the sum of the absolute values of the negative eigenvalues: 0.002 + 0.030 = 0.032, which is quite small compared to 1.423. Exhibit 9.3 shows the classical MDS display in two dimensions of the samples.

Before interpreting this display, how do we quantify the variance explained in this case? There are two ways to do it, depending on whether the non-Euclidean part is included in the total variance or not. The first two eigenvalues, 0.786 and 0.452, can be expressed relative to the sum of the positive eigenvalues, 1.423, or the sum of the absolute values of all the eigenvalues, 1.423 + 0.032 = 1.455. In the former case the percentages would be 56.5% and 32.5%, totalling 89.0%, while in the latter case they would be slightly lower, 55.3% and 31.8%, totalling 87.1%. The non-Euclidean part is quite small in this case, hence the small differences between the two options. An acceptable way of reporting the results would be to say that 2.1% (i.e., 0.032/1.455 = 0.021) of the total variance, is non-Euclidean, and that, of the Euclidean part of the variance, 89.0% (i.e., (0.786 + 0.452)/1.423 = 0.890) is displayed in Exhibit 9.3.

Exhibit 9.3: *Classical multidimensional scaling solution in two dimensions of the matrix of Jaccard dissimilarities of Exhibit 7.1.The percentages of variance on the horizontal and vertical axes are 56.5% and 32.5% respectively (expressed relative to the four-dimensional Euclidean part of the variance)*



Comparing the MDS map in Exhibit 9.3 with the dendrogram in Exhibit 7.8, the separation of the groups {B,F}, {A,C,E,G} and {D} is verified here. The only noticeable difference between the two results is that in the clustering A was joined with E and C with G before they all joined together in a cluster of four, whereas in Exhibit 9.3 C and E look like they should have been clustered first. This is because C and E are, in fact, quite far apart in the third dimension, with coordinates of −0.24 and 0.18 respectively. This large difference in the third dimension can not be seen in Exhibit 9.3 but the original data in Exhibit 7.1 show C indeed further from E compared to the dissimilarity between A and E or C and G.

Nonmetric MDS

The type of MDS described above is included in the family of *metric* MDS methods, since the observed proximities are accepted as quantitative measures of difference between the samples, and the error in the solution is quantified by calculating actual differences between the observed values and those that are displayed in the solution map. Especially when it is known that the proximity matrix is non-Euclidean, an alternative form of MDS may be used, called *nonmetric* MDS, which has a more relaxed way of measuring the quality of the solution. In nonmetric MDS we are not interested in reproducing the proximities themselves, but rather their ordering, that is if we sorted all the observed proximities from

smallest to largest, and we did the same for all the interpoint distances in the solution, then a perfect solution would be if the two orderings were identical. In the matrix of Jaccard dissimilarities of Exhibit 7.1 there are ½ × 7 × 6 = 21 values, ordered in the first set of columns of Exhibit 9.4 (notice the tied ranks). In the second set of columns, the distances between points in Exhibit 9.3 are ordered (there are usually no tied ranks in such fitted distances). Corresponding pairs of samples are linked: if all these links were horizontal, then the distances would be perfectly in order.

The objective of nonmetric MDS would be to get a better correspondence in the orderings of the points. The result for this small data set is surprising, because of the inherent clustering of the samples, shown in Exhibit 9.5.

A comparison of the observed and fitted distances in Exhibits 9.3 and 9.5 clarifies what has happened – see the two plots in Exhibit 9.6. The objective

(a) Observed Jaccard indices

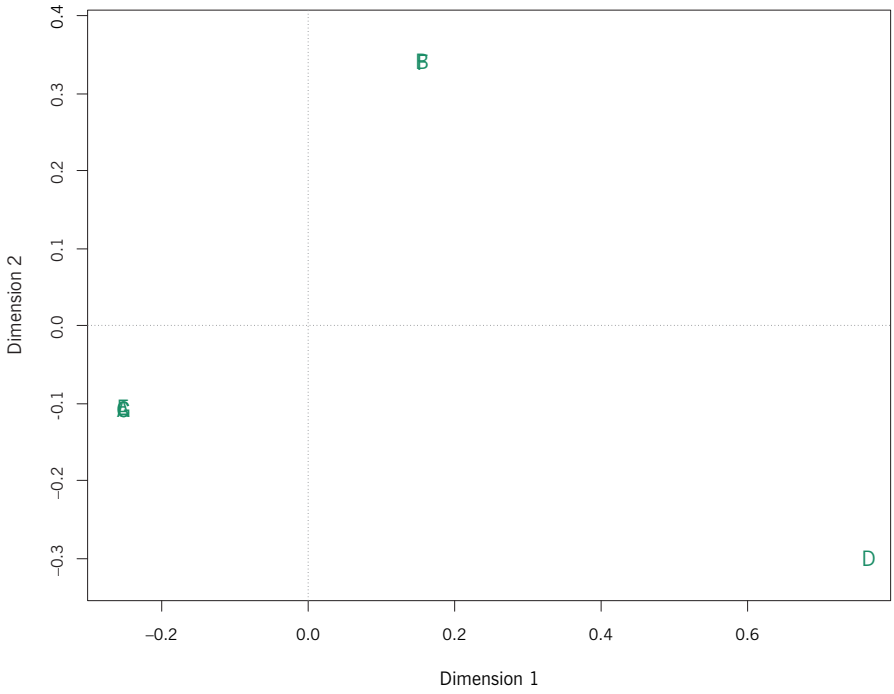
Rank	Jaccard	Pair
1	0.200	(B,F)
2	0.250	(A,E)
3	0.333	(C,G)
4.5	0.375	(E,G)
4.5	0.375	(A,G)
6.5	0.429	(A,C)
6.5	0.429	(C,E)
8	0.500	(A,B)
9	0.625	(A,F)
10.5	0.667	(B,E)
10.5	0.667	(C,F)
12	0.714	(B,C)
13	0.750	(F,G)
14.5	0.778	(B,G)
14.5	0.778	(E,F)
16	0.800	(D,F)
17	0.833	(B,D)
18	0.857	(D,G)
20	1.000	(A,D)
20	1.000	(C,D)
20	1.000	(D,E)

(b) Distances in metric MDS

Rank	Distance	Pair
1	0.040	(C,E)
2	0.074	(B,F)
3	0.154	(A,C)
4	0.175	(A,E)
5	0.219	(C,G)
6	0.223	(E,G)
7	0.365	(A,G)
8	0.524	(A,B)
9	0.559	(A,F)
10	0.633	(B,C)
11	0.655	(C,F)
12	0.669	(B,E)
13	0.692	(E,F)
14	0.752	(B,G)
15	0.754	(F,G)
16	0.775	(D,F)
17	0.842	(B,D)
18	0.848	(D,G)
19	0.955	(C,D)
20	0.988	(D,E)
21	0.993	(A,D)

Exhibit 9.4: *Ordering of the original Jaccard dissimilarities, from lowest to highest, and ordering of the interpoint distances in the metric MDS of Exhibit 9.3*

Exhibit 9.5: *Nonmetric MDS of the Jaccard dissimilarities of Exhibit 7.1. The samples agglomerate into three groups, identical to the clustering in Exhibit 7.8*



of the nonmetric MDS is to find a configuration of the points such that the interpoint distances are close to the ordering of the original distances. In each plot a monotonically increasing function is shown (i.e., a function that never decreases) which best fits the interpoint distances in the map – this function is obtained by a procedure called *monotonic regression*. The error is quantified by the sum of squared deviations between the fitted distances (green circles) and the monotonic regression line. If the sequence of fitted points were always ascending then the monotonic regression line would simply join the points and the error would be zero. Clearly, the upper plot of Exhibit 9.6 shows that there are relatively large deviations of the points from the best-fitting monotonic regression line compared to the near zero deviations in the lower plot. In the lower plot it looks like a perfect fit, but the enlargement of the first seven points shows that there are indeed very small deviations). To explain what has happened there, notice that the interpoint distances between B and F and among all pairs of points in the set {A,C,E,G} are the smallest in the original dissimilarity matrix. Hence, the nonmetric approach puts them all at near-zero distance from one another, and all their values can thereby be reduced to near zero. This maintains their ordering, with very little error from a monotonically increasing relationship, as shown in the enlargement.

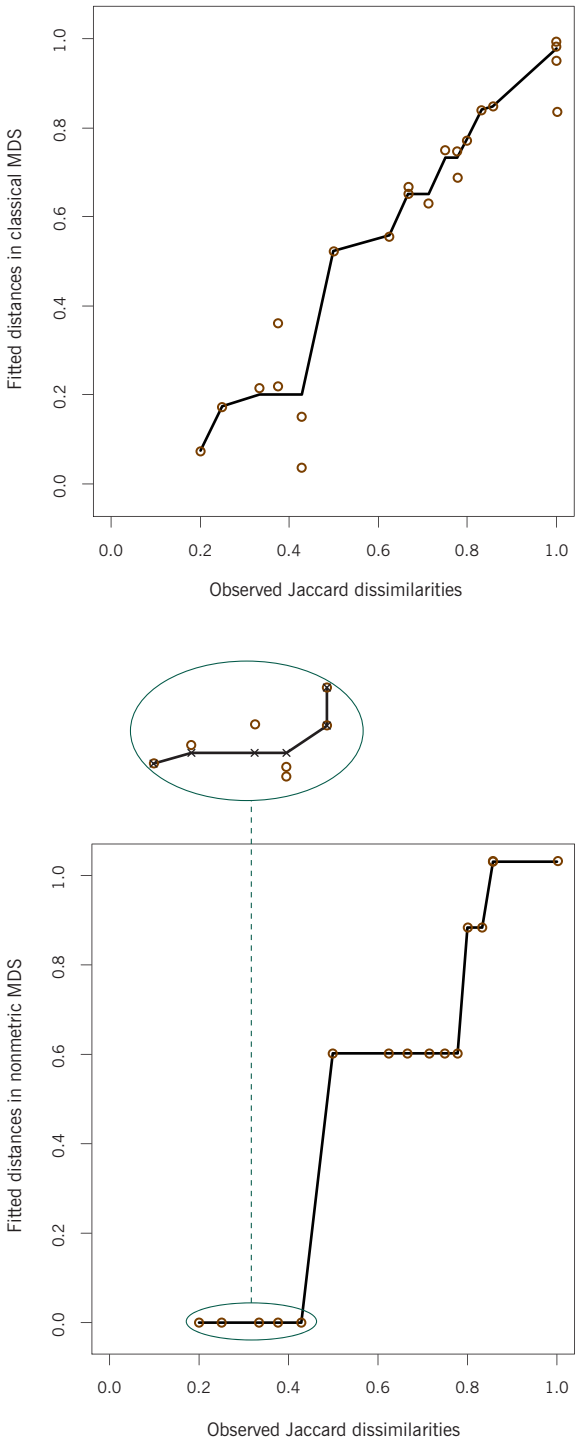


Exhibit 9.6:
The horizontal axis shows the observed dissimilarities from Exhibit 7.1, and the vertical axes show the fitted interpoint distances from Exhibits 9.3 and 9.5 respectively. In both plots the closest fitting monotonically increasing function is shown. The vertical scale of the first seven points in the nonmetric MDS (see lower plot) is expanded considerably to show the small lack of fit for those points

The actual measure of error in a nonmetric MDS is a normalized version of the sum of squared errors, called *stress*. The most popular one is known as *Kruskal's stress formula 1:*

$$\text{stress} = \left(\frac{\sum_{i<j} (\hat{d}_{ij} - d_{ij})^2}{\sum_{i<j} d_{ij}^2} \right)^{1/2} \qquad (9.1)$$

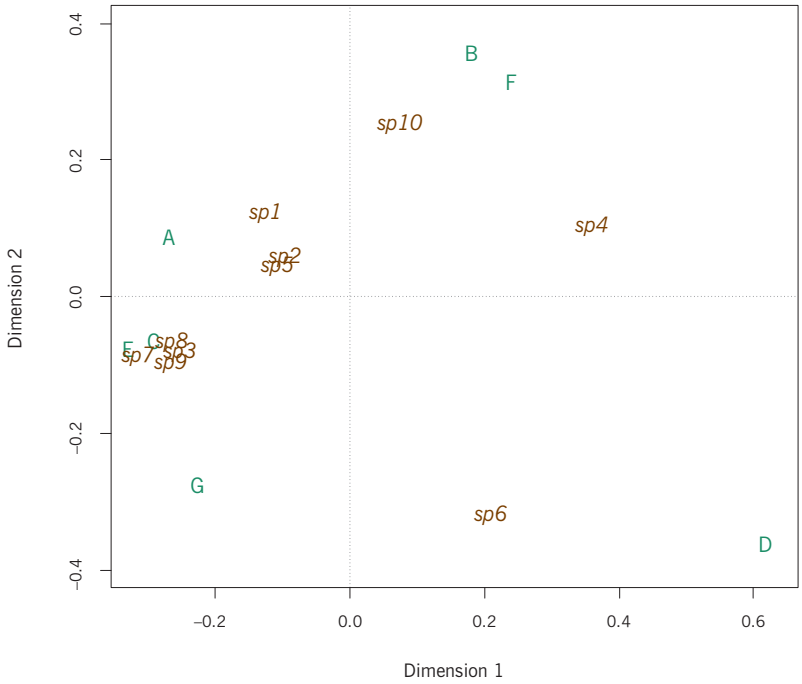
where d_{ij} is the distance between points i and j in the MDS map and $\hat{d}_{ij}$ is the corresponding value on the monotonic regression line (hence the values $|\hat{d}_{ij} - d_{ij}|$ are the vertical discrepancies between points and the line in Exhibit 9.6). This stress measure is often multiplied by 100 and considered as a percentage error: using this convention the metric MDS in Exhibit 9.3 would have a stress of 8.3% while the nonmetric MDS in Exhibit 9.5 would have a stress near zero equal to 0.008% (very low stress values are typical for small data sets like this one – later we will show the result for a larger data set).

Adding variables to MDS maps

The MDS maps in Exhibits 9.3 and 9.5 show the samples only, but the Jaccard dissimilarity matrix was constructed on a samples-by-species data matrix (Exhibit 5.6). Since this is simple presence-absence data, the species can be shown in the MDS maps at positions near the samples that contain them. Usually, they would be situated at the average spatial position of the corresponding samples, shown in Exhibit 9.7 for the two MDS maps. For example, species sp6 is present in samples D and G, so is at an average position halfway between them, while species sp4 is present in samples B, D and F, and is thus positioned at an average position of these three samples.

MDS of Bray-Curtis dissimilarities

In Chapter 5 we computed the Bray-Curtis dissimilarities between 30 samples, s1 to s30, based on the abundances of five species, *a* to *e* – see Exhibit 5.2, where we also pointed out that this measure violated the triangle inequality and was therefore not a metric. For this reason, nonmetric MDS is usually used to map Bray-Curtis indices, but first let us see how metric MDS would handle the display of Exhibit 5.2. The maximum dimensionality of this set of 30 samples is 29, and, as expected, we obtain several negative eigenvalues in the classical MDS: in fact, 14 eigenvalues are positive, with a sum of 57,729, while 15 are negative, with a sum of absolute values equal to 8,176 (Exhibit 9.8). This latter amount quantifies how much variance is impossible to display in a Euclidean space. The first two eigenvalues are 19,102 and 14,825, so the variance explained by the two-dimensional solution, relative to the Euclidean part, is (19,102 + 14,825)/57,729, or 58.8%, that is an error of 41.2%. Computing the stress on this solution, however, gives a value of 16.3%, showing again that the stress criterion always appears more optimistic than the explained variance one.



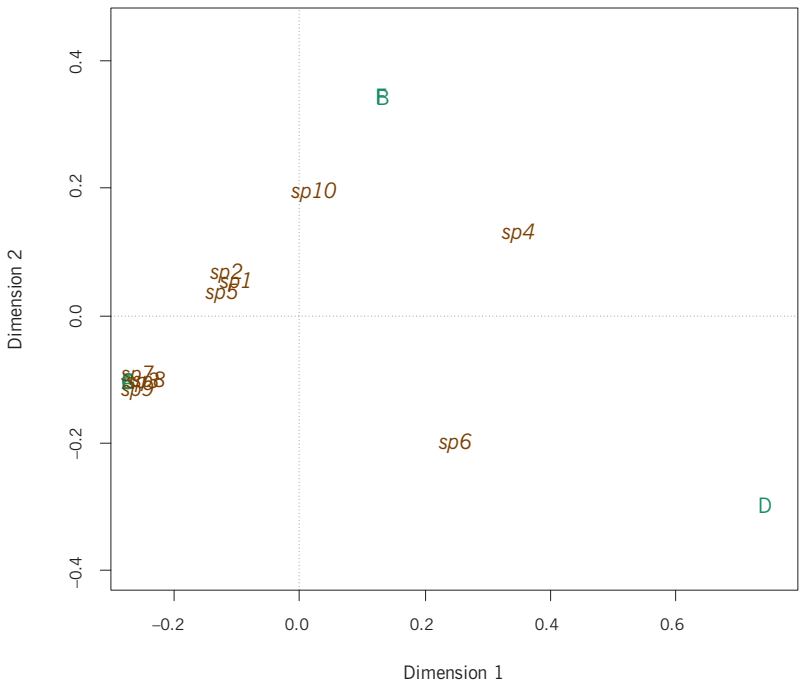
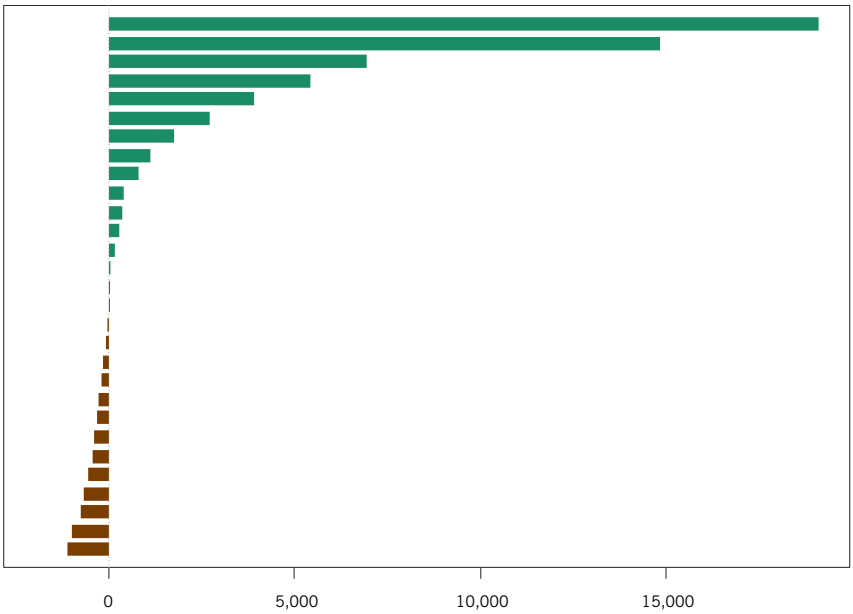


Exhibit 9.7:
The MDS maps of Exhibits 9.3 and 9.5 with the species added at the average positions of the samples that contain them

Exhibit 9.8:
The eigenvalues in the classical MDS of the Bray-Curtis dissimilarity indices of Exhibit 5.2, showing positive eigenvalues in green and negative ones in brown

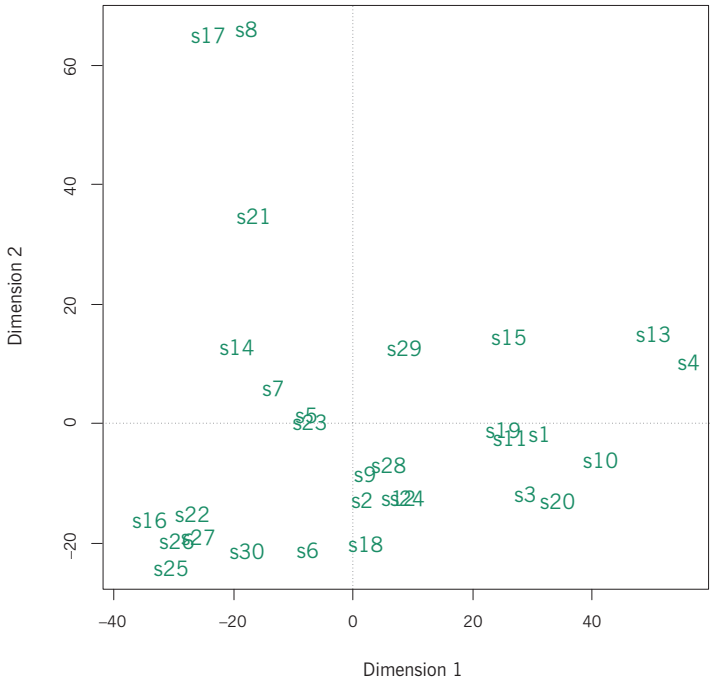


Performing nonmetric MDS on the same data gives a stress value of 13.5%, which is not a big improvement on the 16.3%, suggesting that the two resulting maps will not be as different as we found for the smaller data set of Jaccard indices. This is indeed the case, as shown by the quite similar maps in Exhibit 9.9.

In our experience, when there is a large number of samples (and by "large" we mean, as most statisticians do, 30 or more, as in this example), the metric and nonmetric approaches generally agree in their solutions. Where they disagree is in the quantification of the success of their results, with the stress measure always giving a more optimistic value because it does not measure the recovery of the proximities themselves, but their ordering in the map.

Adding count variables to MDS maps

The maps in Exhibit 9.9 emanate originally from abundance data on five species, so the question now is how to include these species on the map. We shall consider alternative ways of doing this in future chapters, but for the moment let us use the same approach as in Exhibit 9.7 when the species were positioned at the averages of the samples that contained them. The difference here is that we have abundance counts for the species across the samples, so what we can do is to position each species at their weighted average across the samples. For example, species *a* has abundances of 0, 26, 0, 0, 13, etc., and a total abundance of 0 + 26 + 0 + 0 + 13 + ... = 404, so the position of *a* is at a weighted average position of the 30 species, with weights 26/404 = 0.064 on sample s2, 13/404 = 0.032 on sample s5, and



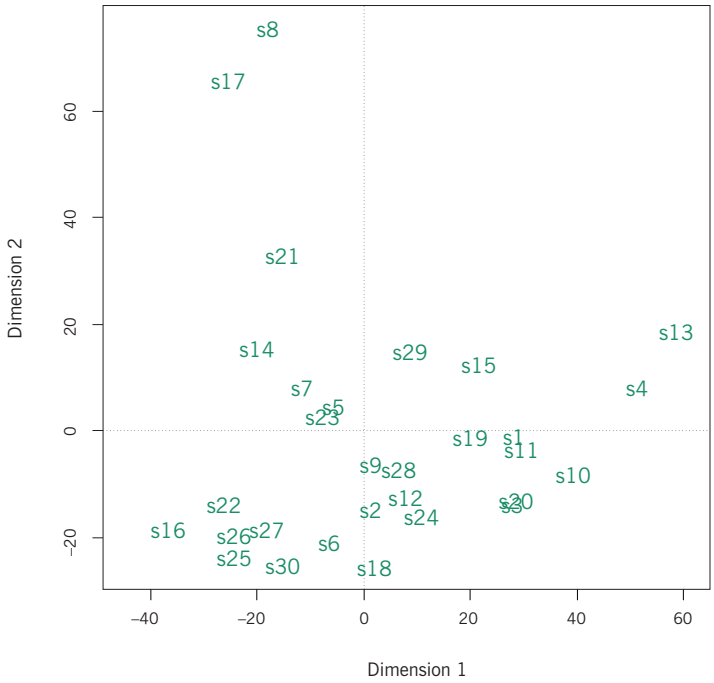
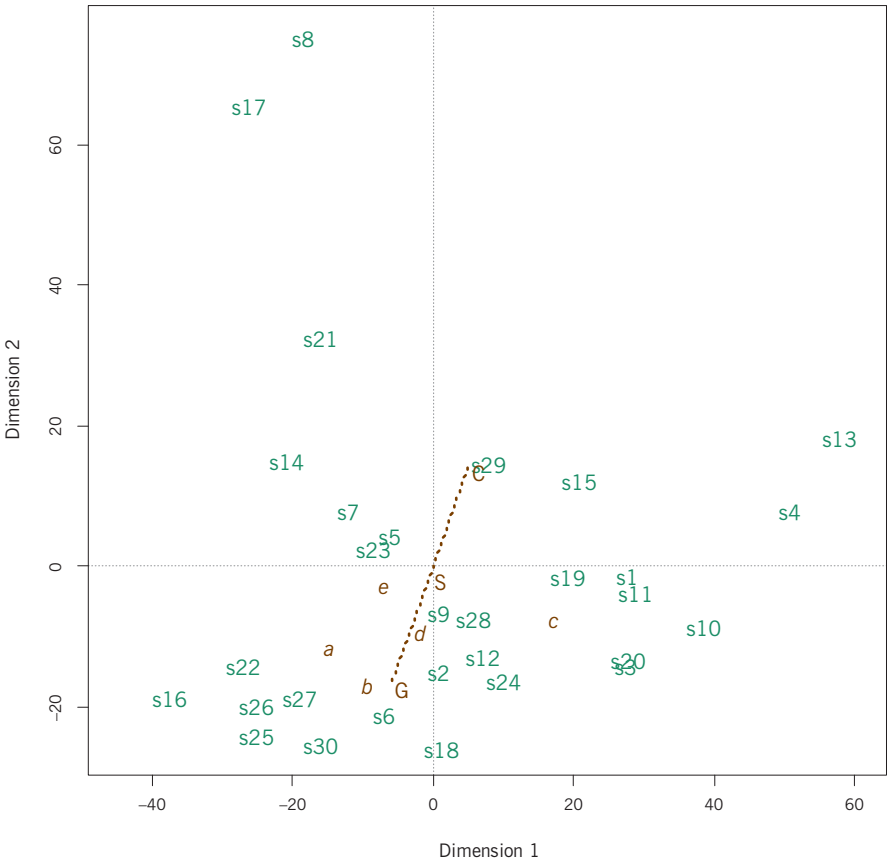


Exhibit 9.9:
Classical MDS map (upper) and nonmetric MDS map (lower) of the Bray-Curtis dissimilarities of Exhibit 5.2

Exhibit 9.10: *Nonmetric MDS solution (right hand map in Exhibit 9.9) with species* a *to* e *added by weighted averaging of sample points, and sediment types* C, S *and* G *by averaging*



so on. Exhibit 9.10 shows the species positions on the nonmetric MDS solution, showing, for example, that species *a* and *b* are relatively more abundant in the samples at lower left, while *c* is more associated with samples on the right. Similarly, even though the ordinal sediment types C (clay), S (sand) and G (gravel) have not been used in the mapping, they can be depicted at the averages of the subsets of samples corresponding to them. The samples thus appear to follow a trend from top right (more clay) to bottom left (more gravel).

SUMMARY: Multidimensional scaling

1. Multidimensional scaling (MDS) is a method that attempts to make a spatial map of a matrix of proximities, either distances or dissimilarities defined between sample units, so that the interpoint distances in the map come as close as possible to the given proximities according to the chosen fit criterion.

2. The fit criterion in metric MDS involves approximating the actual proximity values by the mapped distances, for example by least-squares.

3. Classical MDS is a particular form of metric MDS that relies on the eigenvalue-eigenvector decomposition of a square matrix. The eigenvalues give convenient measures of variance explained on each axis, and the dimensions of the solution are uncorrelated.

4. Nonmetric MDS has a more relaxed fit criterion in that it strives to match only the ordering of the proximities to the ordering of the mapped distances.

5. The error in classical MDS is quantified by the percentage of unexplained variance, while in nonmetric MDS the error is quantified by the stress.

6. The stress measure always gives a more optimistic result, because of the relaxation of approximating the proximity values in the map in favour of their rank ordering.

7. In most cases, however, when the size of the proximity data matrix is quite large, say for at least 30 sample units, the results of the two approaches will be essentially the same.

8. When the proximities are of a Euclidean type, it will be more useful to use the metric scaling approach because of the connection with methods such as principal component analysis (Chapter 12) and correspondence analysis (Chapter 13). There would be little advantage, for example, in applying nonmetric scaling to a matrix of chi-square distances.

9. When the proximities are non-Euclidean, the nonmetric approach avoids the dilemma that the triangle inequality is violated by concentrating on ordering of proximities rather than their actual values.

REGRESSION AND PRINCIPAL COMPONENT ANALYSIS

CHAPTER 10

Regression Biplots

In the previous chapter, displays of samples were obtained in a scatterplot with spatial properties (hence often called a *map),* approximating given distance or dissimilarity matrices. Then some types of variables were added to the display, specifically zero/one categorical variables (e.g., presences of species, sediment categories) and count variables (e.g., species abundances). In this chapter we continue with this theme of adding variables to a plot of samples, including continuous variables in their original form or in fuzzy-coded form. When samples and variables are displayed jointly in such a scatterplot, it is often called a *biplot.* This designation implies that a certain property holds between the two sets of points in the display in terms of the scalar products between the samples and variables. In this chapter we consider the simplest form of biplot, the regression biplot, which will serve two purposes: first, to give a different geometric interpretation of multiple regression; and second, to give a basic understanding of all the joint displays of samples and variables that will appear in the rest of this book.

Contents

Algebra of multiple linear regression

The multiple linear regression model postulates that the expected value of a response variable Y (i.e., the mean of Y) is a linear combination of several explanatory variables $x_1, x_2, \ldots, x_p$:

$$\mathrm{E}(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_p x_p \qquad (10.1)$$

For example, using the data of Exhibit 1.1, consider the regression of species labelled *d* on depth, pollution and temperature. The model is estimated as:

$$E(d) = 6.271 + 0.148 \times \text{depth} - 1.388 \times \text{pollution} - 0.043 \times \text{temperature} \quad (10.2)$$

Notice that, for the moment, we do not comment on whether this type of linear model of a count variable on three environmental variables would be sensible or not, because *d* is not an interval variable – we will return to this point later.

Since the coefficients in (10.2) depend on the units of the variables, we prefer to consider the regression using all variables in comparable units. Usually this is done by standardization of the variables, so that they are all in units of standard deviation. Let us denote these standardized variables (i.e., centred and normalized) with an asterisk, then the regression model becomes:

$$E(d^*) = 0.347 \times \text{depth}^* - 0.446 \times \text{pollution}^* - 0.002 \times \text{temperature}^* \quad (10.3)$$

The constant term now vanishes and the coefficients, called *standardized regression coefficients,* can be compared with one another. Thus it seems that pollution has the strongest influence on the average level of species *d*, reducing it by 0.446 of a standard deviation for every increase of one standard deviation of pollution. The effect of temperature is minimal and, in fact, is nonsignificant statistically ($p = 0.99$), while depth and pollution are both significant ($p = 0.039$ and $p = 0.010$, respectively), so we drop temperature and consider just the regression on the other two variables, which maintains the value of the coefficients, but slightly smaller p-values: $p = 0.035$ and $p = 0.008$, respectively:

$$E(d^*) = 0.347 \times \text{depth}^* - 0.446 \times \text{pollution}^* \quad (10.4)$$

Geometry of multiple linear regression

When referring to the multiple regression model, it is often said that a *hyperplane* is being fitted to the data. For a single explanatory variable this reduces to a straight line in the familiar case of simple linear regression. When there are two explanatory variables, as in (10.4), the model is a two-dimensional plane in three dimensions, the third dimension being the response variable d^* – a view of this plane in three dimensions is given in Exhibit 10.1, with standardized depth* and pollution* forming the two horizontal dimensions and d^* the vertical one. Notice how the plane is going down in the direction of pollution, but going up in the direction of depth, according to the regression coefficients (see the web site of the book which shows a video of this three-dimensional image). Notice too the lack of fit of the points to the plane – the value of R^2 for the regression is 0.442, which means that 44.2% of the variance of *d* is being

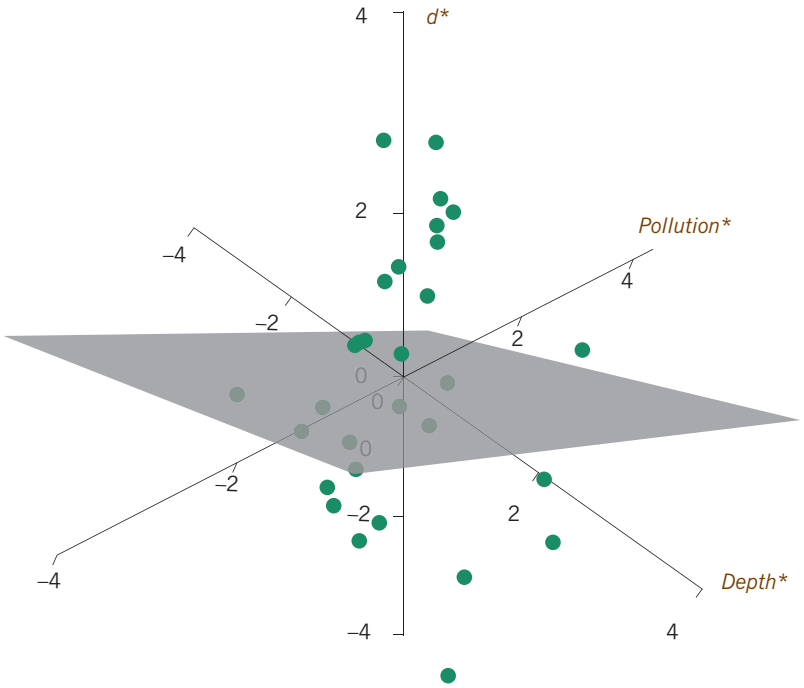


Exhibit 10.1: *Regression plane defined by Equation (10.4) for standardized response d* and standardized explanatory variables pollution* and depth*. The view is from above the plane*

explained, and 55.8% of the variance unexplained and considered residual, or error, variance.

The linearity of the plane means that predictions of the same mean values form parallel straight lines in the plane. From a mountaineer's point of view, if you are standing on the plane and want to stay at the same height, you need to walk in a straight line. Projecting these parallel straight lines onto the depth–pollution plane gives the *contours*, also called *isolines*, as shown in Exhibit 10.2. Finally, the vector in the depth–pollution plane with coordinates equal to the regression coefficients, [0.347 –0.446], called the *gradient*, indicates the direction of steepest ascent in the regression plane, and is perpendicular to the contours. Given the geometry of the regression plane in Exhibit 10.2, it follows that we can do away with the *d** dimension, just like cartographers do, and consider just the depth–pollution plane and the contours of the regression plane, which are perpendicular to the gradient vector. Exhibit 10.3 shows this "ground view" of the model.

The short arrow labelled *d* is the gradient vector. The dashed line through this vector is called the *biplot axis* for the variable *d*. Contour lines are perpendicular to the biplot axis. Exhibit 10.3(a) corresponds to the darker "shadow" in Exhibit 10.2 in the depth–pollution plane, where the contours are in units of standard

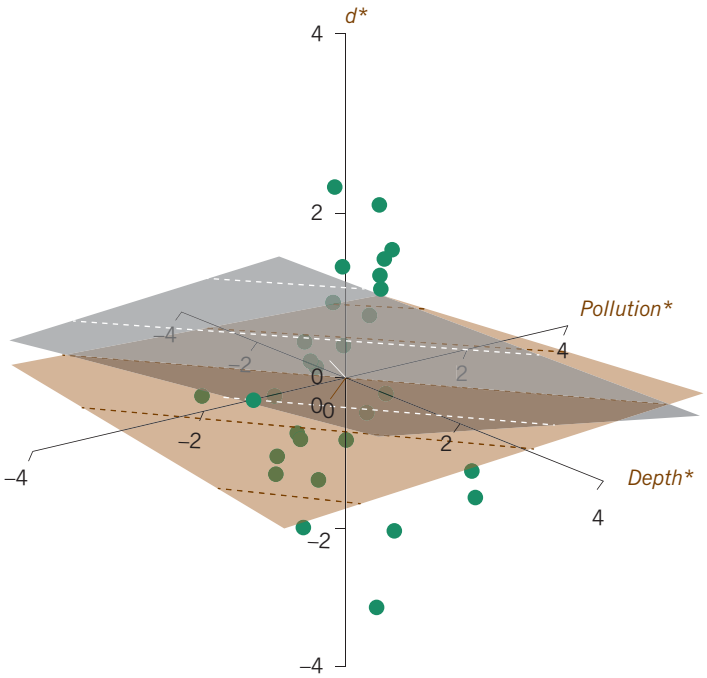


Exhibit 10.2: *Another view of the regression plane, showing lines of equal height (dashed white lines in the plane) and their projection onto the depth—pollution plane (brown dashed lines in the darker "shadow" of the plane). The view is now from below the regression plane but above the depth—pollution plane. The short solid white line in the regression plane shows the direction of steepest ascent, and its projection down onto the depth—pollution plane is the gradient vector*

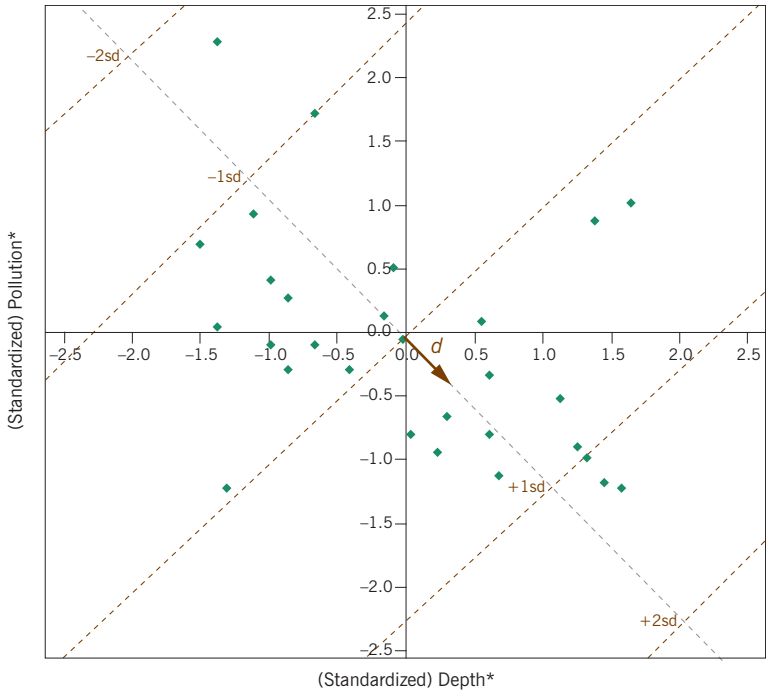
deviation (sd) of species d (sd of $d = 6.7$). The mean of d, equal to 10.9, corresponds to the contour line through the origin. Calibrating the biplot axis in the original abundance units of d, Exhibit 10.3(b) is obtained.

Now the expected abundance, according to the regression model, can be estimated for any sample by seeing on what contour line it lies, which is achieved by projecting the point perpendicularly onto the biplot axis. For example, the sample shown in Exhibit 10.3(b), with standardized coordinates –0.668 and 1.720, is on a contour line with value 4.2. The observed value for this sample is 3, so this means that the regression plane lies above the sample point and thus over-estimates its value. The action of projecting the sample point perpendicularly onto the biplot axis is a scalar product operation – just the regression model (10.4), in fact. The scalar product of the gradient vector [0.347 –0.446] with the sample point vector [–0.668 1.720] is:

$$0.347 \times -0.668 + (-0.446) \times 1.720 = -0.999$$

which means that the prediction is almost exactly one standard deviation below the mean of d (in Exhibit 10.3(a) it is on the contour line –1sd), that is $10.9 - 0.999 \times 6.7 = 4.2$.

(a)



(b)

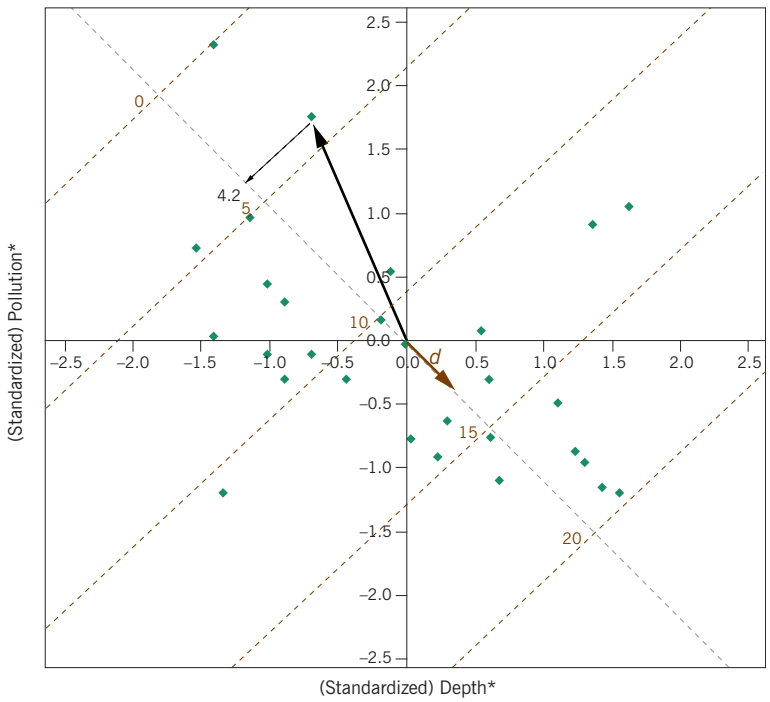
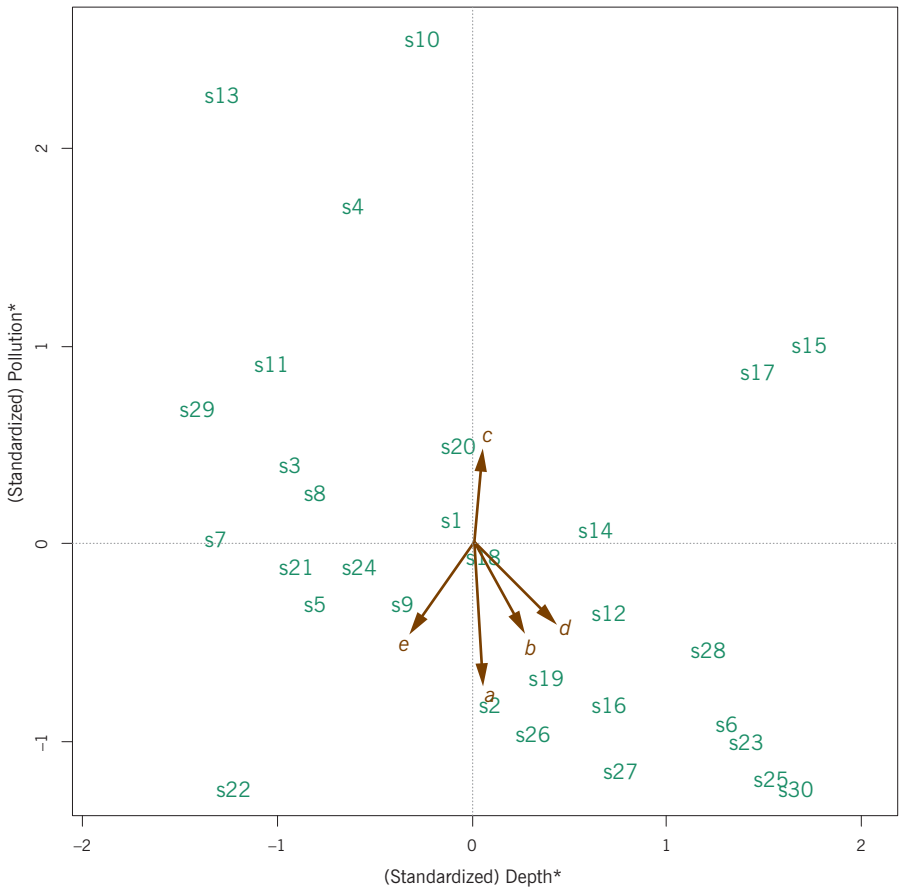


Exhibit 10.3:
Regression plane shown as contour lines in the plane of the two explanatory variables, depth and pollution, both standardized. In (a) the contours are shown of the standardized response variable d, where the units are standard deviations (sd's) and the contour through the origin corresponds to mean 0 on the standardized scale, i.e. the mean on the original abundance scale. In (b) the contours are shown after unstandardizing to the original abundance scale of d. The sample shown in (b) corresponds to a height of 4.2 on the regression plane*

Regression biplot

We have given a different geometric view of multiple regression, for the case of two predictor variables, reducing the regression model to the gradient vector of regression coefficients in the plane of the predictors (we will come to the case of more predictor variables later). The contours of the plane are perpendicular to the gradient vector. We can now perform the regressions of the other four species with the same two predictors. Each species has a different pair of regression coefficients defining its gradient vector and all five of these are plotted together in Exhibit 10.4. The fact that *b* and *d* point in similar directions means that they have similar regression relationships with the two predictors, and the samples will have similar projections onto the two biplot axes through *b* and *d*. Species *a* and *c* point in opposite directions and thus have opposite relationships, *a* negative with pollution and *c* positive. While samples high in the vertical direction such as s10 and s13 have high (modelled) abundances of *c*, they also have the lowest abundances of *a*.

Exhibit 10.4: *Regression biplot of the five species with respect to the predictors depth and pollution*



Each regression has an associated R^2 value: for the five species these are (as percentages) (*a*) 52.9%, (*b*) 39.8%, (c) 21.8%, (*d*) 44.2%, (e) 23.5%. An overall measure of variance explained for all five regressions in the biplot is the ratio of the sum of the explained variances in each and the sum of the total variances, which gives a value of 41.5%. As far as statistical significance is concerned, all species have significant linear relationship with pollution, but only *d* and e are significantly related to depth as well (these have the highest standardized regression coefficients on the horizontal axis in Exhibit 10.4).

Generalized linear model biplots with categorical variables

In Chapter 9 we have already shown how the environmental variable sediment (Exhibit 1.1), which is categorical, can be added to a MDS display. Each category is placed at the average of the samples in which it is contained – we call these *supplementary points*. Similarly, we situated species in an MDS map as supplementary points by positioning them at their weighted averages of the sample points, with weights equal to the relative abundances. This approach can be used here as well, but their positions do not reflect any formal relationship between the species and the predictors. *Logistic regression* can be used in this case to give gradient vectors to represent these categories.

Logistic regression models the logarithm of the odds of being in a given category, in this case a particular sediment category. Modelling the log-odds (i.e., the *logit*) for each sediment category as a function of (standardized) depth and pollution using logistic regression leads to three sets of regression coefficients in the linear part of the model. For example, for gravel (G), the model is:

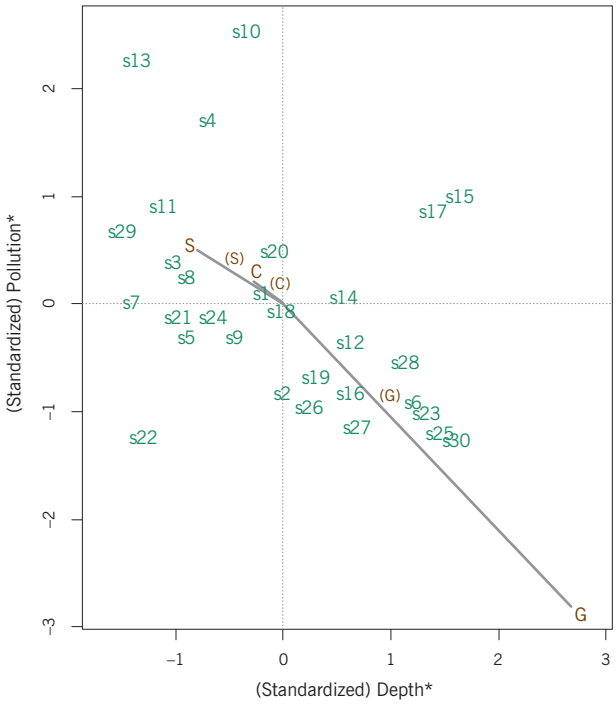
$$\text{logit}(p_G) = \log\left(\frac{p_G}{1-p_G}\right) = -3.322 + 2.672 \times \text{depth}^* - 2.811 \times \text{pollution}^*$$

The three sediment categories are shown according to their logistic regression coefficients in Exhibit 10.5(a), connected to the origin. In fact, the above logistic regression is the only one that is statistically significant, those for clay (C) and sand (S) are not. The categories as supplementary points are also shown in Exhibit 10.5(a) by smaller labels in parentheses.

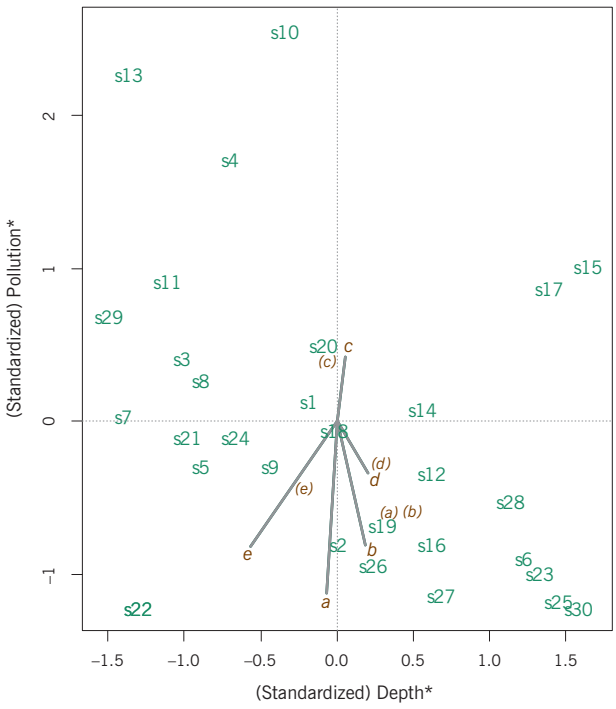
Rather than use linear regression to display the species in a regression biplot, as in Exhibit 10.4, there are two other alternatives: Poisson regression, which may be considered more appropriate because it applies to count response data, or fuzzy coding. The Poisson regressions lead to the coefficients displayed as vectors in Exhibit 10.5(b). Significance with respect to the two predictors is the same as for the linear regression (see above), with in addition species *b* being significantly related to depth. Both Poisson and logistic regression are treated in more detail in Chapter 18.

Exhibit 10.5: *(a) Logistic regression biplot of the three sediment categories and (b) Poisson regression biplot of the five species as predicted by depth and pollution. In each biplot the gradient vectors are shown connected to the origin. In addition, the positions of the sediment categories and the species as supplementary points are given in their respective biplots by their labels in parentheses*

(a)



(b)



Fuzzy-coded species abundances

For the fuzzy coding, because there are several zeros in the abundance data, we can set up fuzzy codes for the species with a "crisp" code just for zeros and then three fuzzy categories for the nonzero values. Thus, for species *a*, for example, the code *a0* refers to zero abundance and *a1*, *a2* and *a3* refer to low, medium and high positive abundances. Exhibit 10.6 shows the two ways of representing the fuzzy categories, first in terms of their (linear) regressions on depth and pollution, and second, in terms of their supplementary point positions as weighted averages of the samples. Overall, 10.6(a) and (b) tell the same story: most of the variation is in a vertical direction, along the pollution direction, with high values of species *e* (i.e., category *e3*) ending up bottom left, while the corresponding categories for species *b* and *d* end up bottom right. The trajectories of each species contain features that are not possible to see in the previous biplots. For example, species *d* has an interesting nonlinear trajectory, with low positive values (*d1*) pulled out towards the shallowest depths. Since the sample size is small, this feature may not be statistically significant – we shall return to this aspect later, the point we are making here is that fuzzy coding can reveal more information in the relationships than linear models.

More than two predictors

The two predictor variables depth and pollution form what is called the *support* of the biplot. With the aid of three-dimensional graphics we could have a third variable, in which case the gradient vectors would be three-dimensional. But if we only have a two-dimensional "palette" on which to explore the relationships, multivariate analysis can provide the solution, at the expense of losing some information. As in the case of MDS, however, we are assured that a minimum amount of information is lost.

Without entering into all the details of a multivariate method called *canonical correlation analysis*, which is a form of linear regression analysis between two sets of variables, we simply show its results in Exhibit 10.7, which visualizes all the (linear) relationships between the five species and the three predictor variables depth, pollution and temperature. The configuration of the samples looks very similar to a 90 degree counter-clockwise rotation of the scatterplots in previous biplots. However, the support dimensions are no longer identified with single predictors but are rather linear combinations of predictors. The two canonical axes are defined as follows:

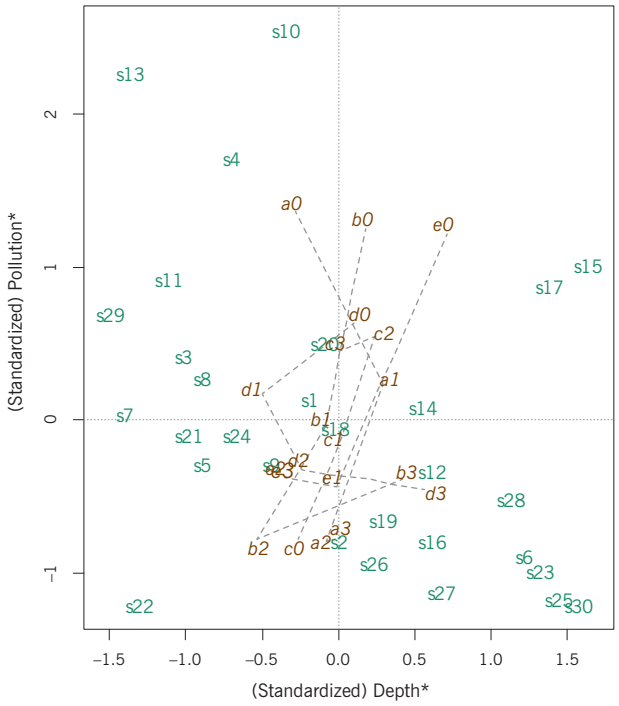
$$\text{canonical dimension } 1 = 0.203 \times \text{depth}^* - 0.906 \times \text{pollution}^* - 0.009 \times \text{temperature}^*$$

$$\text{canonical dimension } 2 = 1.057 \times \text{depth}^* + 0.607 \times \text{pollution}^* - 0.102 \times \text{temperature}^*$$

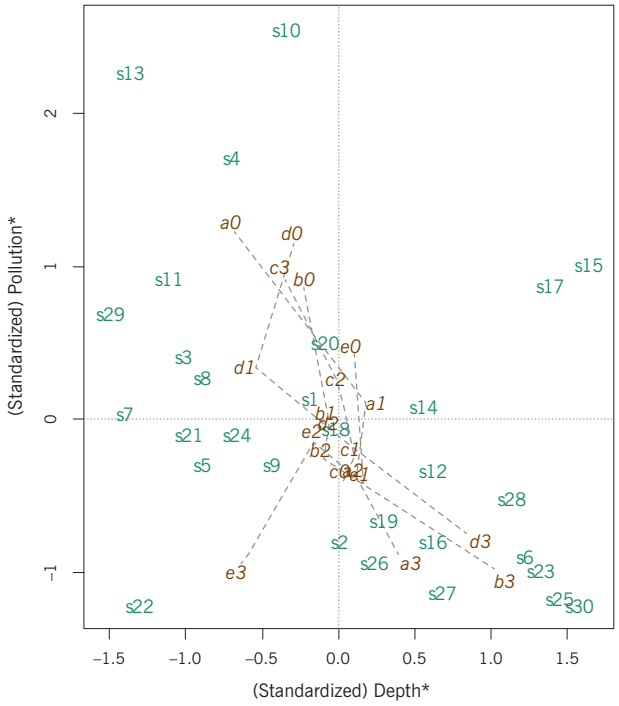
These dimensions are established to maximize the correlation between the species and the environmental variables. The first dimension is principally pollution

Exhibit 10.6: *Fuzzy coding of the species, showing for the fuzzy categories (a) their regressions on (standardized) depth and pollution, and (b) their weighted average positions with respect to the samples (i.e., supplementary points)*

(a)



(b)



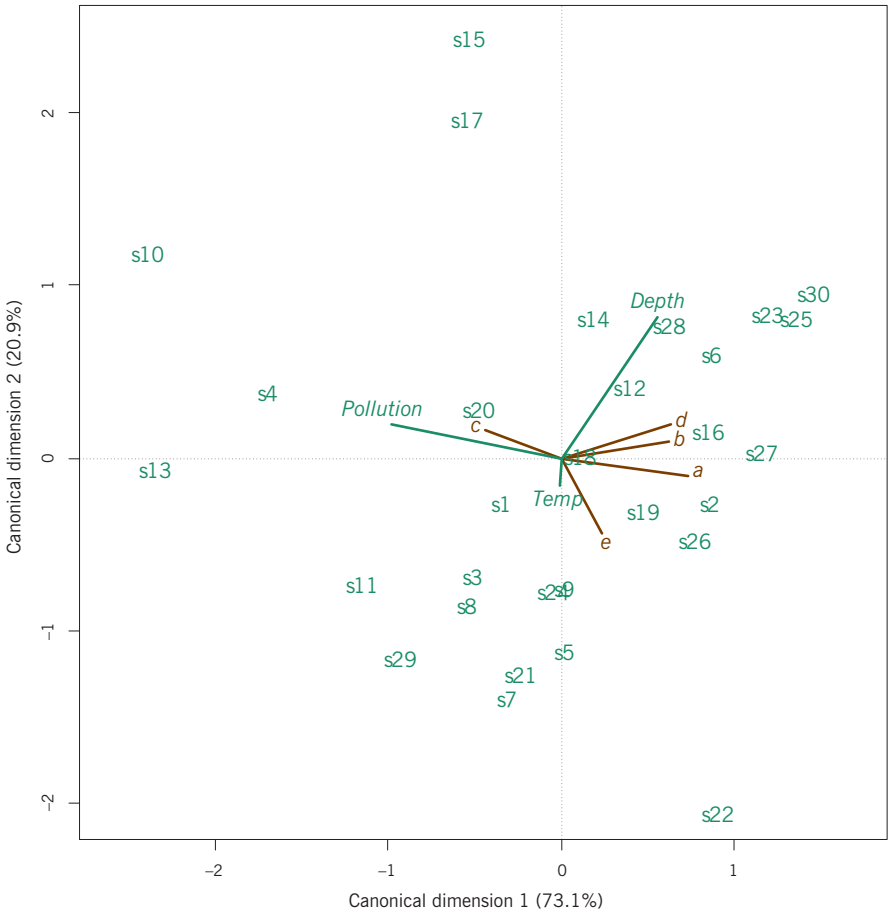


Exhibit 10.7: *Canonical correlation biplot of the five species with respect to the predictors depth, pollution and temperature*

but to a lesser extent depth, while the second is mainly depth but to a lesser extent pollution. The first dimension is the most important, accounting for 73.1% of the correlation between the two sets, while the second accounts for 20.9%, that is 94.0% for the two-dimensional solution. Temperature plays a very minor role in the definition of these dimensions. The canonical dimensions are standardized, hence the support has all the properties of previous biplots except that the dimensions are combinations of the variables, chiefly depth and pollution. In addition, the canonical axes have zero correlation, unlike the depth–pollution support where the two variables had a correlation of –0.396. Now that the support is defined, the species can be regressed on these two dimensions, as before, to show their regressions in the form of gradient vectors. The three environmental variables can be regressed on the dimensions as well, and their relationship shown using their gradient vectors, as in Exhibit 10.7. Notice that the angle between pol-

lution and depth suggests the negative correlation between them – see Chapter 6. In fact, the cosine of the angle between these two vectors is –0.391, very close to the actual sample value of –0.396. Notice as well the absence of relationship of temperature with the canonical dimensions.

Exhibit 10.7 is two biplots in one, often called a *triplot*. We shall return to the subject of triplots in later chapters – they are one of the most powerful tools that we have in multivariate analysis of ecological data because they combine the samples, responses (e.g., the species) and the predictors (e.g., the environmental variables) in a single graphical display, always optimizing some measure of variance explained.

SUMMARY: Regression biplots

1. When there are two predictor variables and a single response variable in a multiple regression, the modelled regression plane can be visualized by its contours in the plane of the predictors (usually standardized). The contours, which are parallel straight lines, show the predicted values on the regression plane.
2. A regression biplot is built on a scatterplot of the samples in terms of the two predictors, called the *support* of the biplot. The gradient of the response variable is the vector of its regression coefficients, indicating the direction of steepest ascent on the regression plane. The gradient is perpendicular to the contour lines.
3. Several (continuous) response variables can be depicted by their gradient vectors in the support space, giving a *biplot axis* for each variable, and sample points can be projected perpendicularly onto the biplot axes to obtain predicted values according to the respective regression models.
4. When response variables are categorical, their gradient vectors can be obtained by performing a logistic regression on the predictors. Alternatively, the categories can be displayed at the averages of the sample points that contain them.
5. When response variables are counts (e.g., abundances), their gradient vectors can be obtained by Poisson regression. Again, there is the alternative of displaying them at the weighted averages of the sample points that contain them, where weights are the relative abundances of each variable across the samples.
6. For more than two predictor variables the support space of low dimensionality can be obtained by a dimension-reducing method such as canonical correlation analysis. Dimension reduction is the main topic of the rest of this book.

CHAPTER 11

Multidimensional Scaling Biplots

The ecological literature abounds with different measures of distance and dissimilarity between samples, computed on species abundance data, biomass data, presence/absence data, compositional data, and so on. We have seen that we can perform cluster analyses on such proximity measures but also can map the samples in a spatial display using MDS. We have also seen how the variables themselves can be added, through regression or through averaging, to a support space that is usually two-dimensional for ease of interpretation. In this chapter we unite the idea of an MDS display with that of the regression biplot, to obtain joint displays of samples and variables. This step is a precursor to the many methods of multivariate analysis that are used in practice, notably principal component analysis, log-ratio analysis and correspondence analysis, which are the subjects of future chapters.

Contents

The "Barents fish" data set

This data set consists of the abundances of 30 fish species at 89 sampling stations from the shrimp survey in the Barents Sea in the period April-May 1997. These data are part of a much larger data set for the Barents Sea over several years.[1] The locations of the samples are shown in Exhibit 11.1. Apart from the abundances, available environmental covariates include bottom depth, temperature

[1] Data were collected during the former annual shrimp surveys in the Barents Sea, by the Norwegian Institute of Fisheries and Aquaculture (NIFA) and the Institute of Marine Research (IMR), Norway.

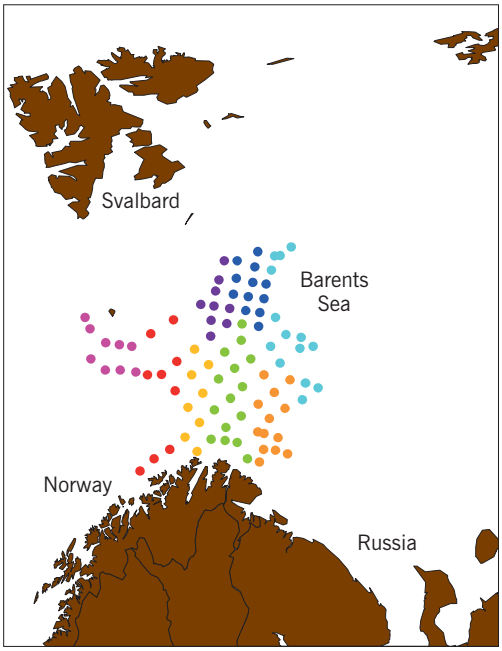


Exhibit 11.1: *Locations of samples in "Barents fish" data set. At each sampling point the data consist of the abundances of 30 fish species, the bottom depth, the temperature and the spatial position (latitude and longitude). The stations have been colour coded into approximately neighbouring groups, using great circle distances, for comparison with the MDS map based on the abundances (coming in Exhibit 11.3)*

and the latitude/longitude coordinates for each sample. A small part of the data set is shown in Exhibit 11.2, showing that the species abundances differ widely both among samples and among species (see the sums of the rows and columns shown). Nevertheless, the samples come from equal volume sampling, a 20-minute bottom trawl in each case.

Exhibit 11.2: *Part of the "Barents fish" data set: 89 samples (rows), 4 environmental variables and 30 fish species (columns)*

Station	Environmental data				Species abundance data										
ID No	*Latitude*	*Longitude*	*Depth*	*Temp.*	*Re_hi*	*An_de*	*An_mi*	*Hi_pl*	*An_lu*	*Me_ae*	*Ra_ra*	···	*Ca_re*	*Tr_spp*	*Sum*
356	71.10	22.43	349	3.95	0	0	0	31	0	108	0	···	0	0	*845*
357	71.32	23.68	382	3.75	0	0	0	4	0	110	0	···	0	0	*1,740*
358	71.60	24.90	294	3.45	0	0	0	27	0	788	0	···	0	0	*1,763*
359	71.27	25.88	304	3.65	0	0	1	13	0	295	0	···	0	0	*767*
363	71.52	28.12	384	3.35	0	0	0	23	0	13	2	···	0	0	*1,347*
364	71.48	29.10	344	3.65	1	0	0	20	0	97	0	···	0	0	*801*
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
462	70.83	21.32	167	4.45	0	0	0	10	2	50	0	···	0	0	*232*
465	73.38	17.37	462	1.95	0	0	0	0	0	0	0	···	0	0	*36*
				Sum	*316*	*135*	*45*	*8,564*	*9*	*6,141*	*305*	···	*62*	*653*	*63,896*

Relating MDS map to geography

In order to perform MDS on these data, either Bray-Curtis or chi-square can be computed, bearing in mind the differences between them when they are applied in their usual forms: Bray-Curtis is computed on the original abundances, whereas chi-square is applied to the relative abundances in each sample. There is also the issue about whether abundances should be transformed before applying Bray-Curtis, for example a square root or even fourth root transformation. We first compute Bray-Curtis on the raw data and then consider transformations later in this chapter. Exhibit 11.3 shows the nonmetric MDS solution in two dimensions of the Bray-Curtis dissimilarities, with the same colour coding as in the geographical map of Exhibit 11.1. The MDS solution is interpreted spatially as the similarity between the samples in terms of their species abundances, while Exhibit 11.1 represents the geographical proximities between the sampling locations. Already we can see in Exhibit 11.3 that points in the same spatial group (colour-coded) are often close together. So we can consider here the interesting question of measuring how similar the biologically determined map is to the geographical map. A simple approach

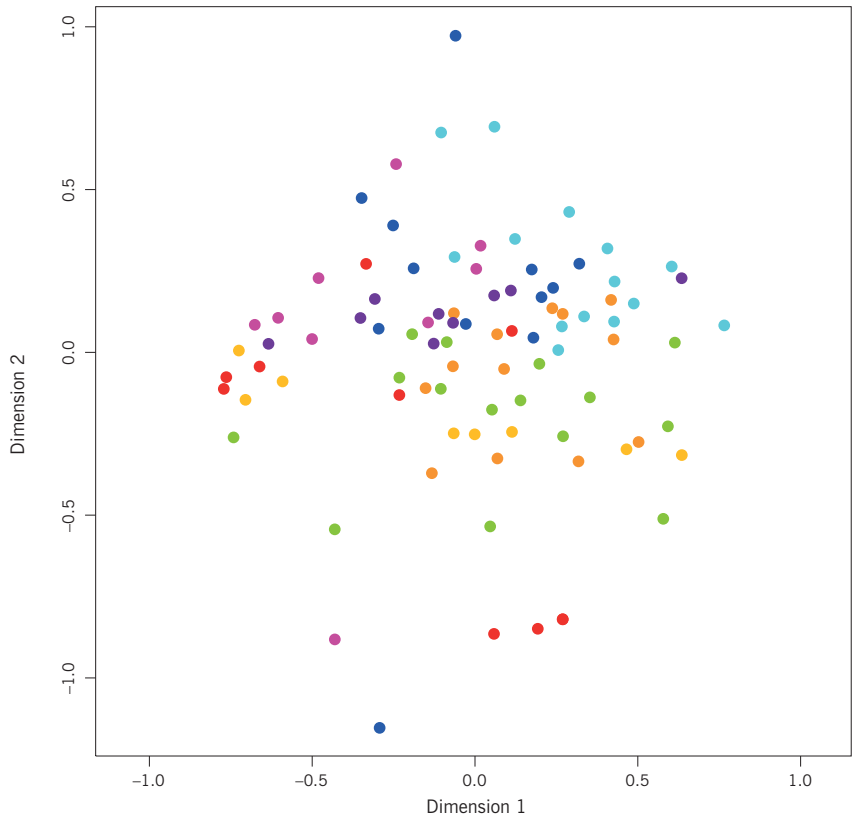
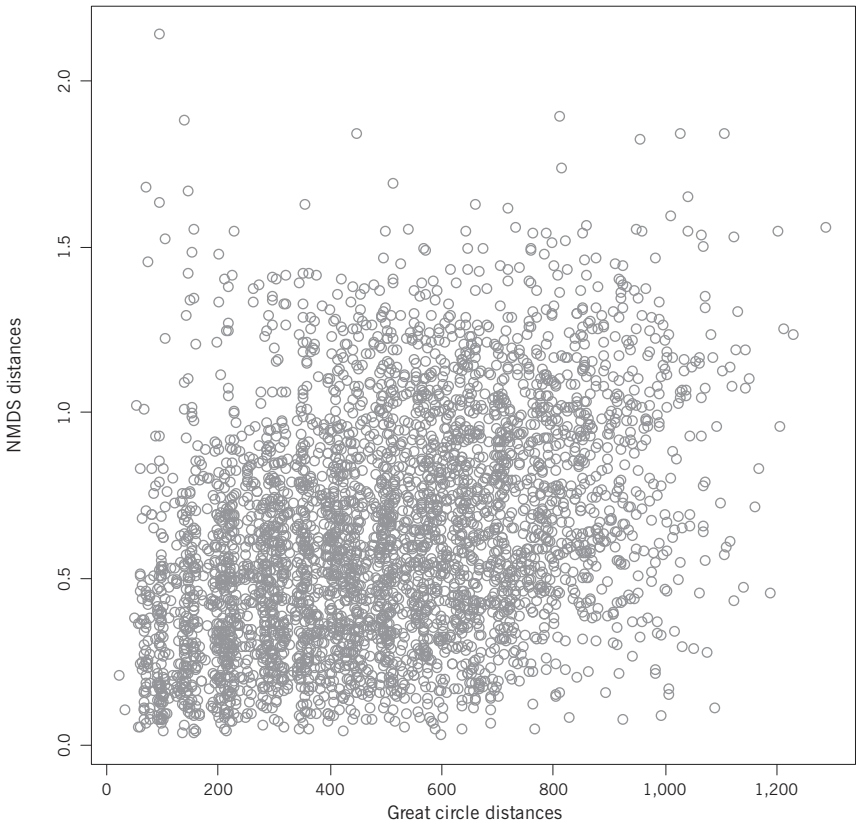


Exhibit 11.3: *Nonmetric MDS of the Bray-Curtis dissimilarities in community structure between the 89 samples, with the same colour coding as in the map of Exhibit 11.1*

Exhibit 11.4:
Scatterplot of inter-sample geographical (great circle) distances and distances in Exhibit 11.3. Spearman rank correlation = 0.378



is to plot the two sets of inter-sample distances against one another (there are ½ × 89 × 88 = 3,916 of them), shown in Exhibit 11.4. The Spearman rank correlation between the two sets of distances, which measures how similar the two sets of distances are, is equal to 0.378.

An alternative approach is to use *Procrustes analysis*, a method that is specifically designed to measure the similarity between two maps. One of the maps, in this case the geographical map, is used as a *target* matrix and the other one, the MDS map, is rotated, translated and rescaled to best fit the target. One possible problem here is that the Euclidean distances between the latitude/longitude coordinates do not give the great circle distances. This can be solved by using a map projection in R (see Appendix C) to get coordinates for which Euclidean distances are approximate great circle distances. The Procrustes correlation between Exhibits 11.1 (using projected coordinates) and 11.3 turns out to be 0.549 and highly significant ($p < 0.0001$), hence there is a relationship between the geography and the fish community structure.

Adding continuous variables to a MDS map

As presented in Chapter 10, we use the regression biplot to add species and environmental variables to an MDS map such as Exhibit 11.3. A Poisson regression of every species can be performed on the two dimensions of the map, and the species depicted by the gradient vector of its regression coefficients. For every environmental variable (here we have two, but we can include latitude and longitude as well for the moment) a linear regression can be performed on the two dimensions of the map, again providing a gradient vector in each case. The result is shown as a separate display in Exhibit 11.5.

Notice several technical aspects of this display. First, each Poisson regression models the logarithm of the mean species abundance, so that biplot axes through the species vectors, if calibrated, would have logarithmic scales. On the other hand, the biplot axes indicated by the environmental variables would be calibrated linearly – see Chapter 10. Second, to equalize the scales of the environmental variables, they were standardized. Third, because of the dispar-

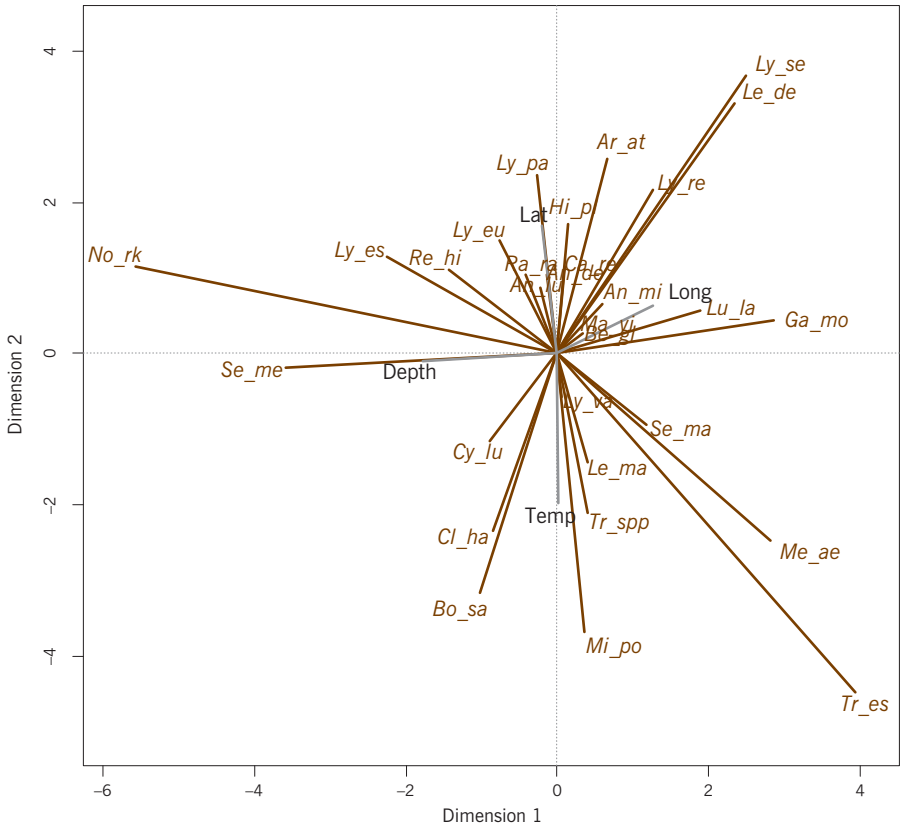


Exhibit 11.5: *Gradient vectors of the species (from Poisson regressions) and of the environmental variables (from linear regressions) when regression is performed on the dimensions of Exhibit 11.3*

ity in scale between the set of species (on a logarithmic scale) and the set of environmental variables (standardized), one can only compare vectors' lengths within a group, and not between groups. For example, the percentage of variance explained in the regression of the species *Mi_po* (*Micromesistius poutassou*, blue whiting) and of the variable temperature (Temp), both negative on the vertical axis, is of the order of 40%, yet *Mi_po* has a vector about twice as long as temperature.

The interpretation of the environmental variables in Exhibit 11.5 is quite clear, and has a strong relationship to the spatial distribution of the samples. Increasing latitude points conveniently "north" in the solution and longitude points "east".[2] Temperature points "south", sea water getting warmer with decreasing latitudes, and depth points "west", so the samples are deeper as longitude decreases. The directions of the species, especially those with longer gradient vectors, give the biological interpretation of the ordination. In cold waters in the "north", species typical of Arctic water masses tend to dominate, whereas more Atlantic species characterize samples in the "south".

Nonlinear contours in a MDS map

The vectors in Exhibit 11.5 are assuming that the relationships between the environmental variables and the dimensions of the MDS map are linear, in which case all we need to know is each variable's gradient vector. To check this, an alternative way to depict the changing nature of the environmental variables in the MDS map is to plot their contours nonlinearly, as shown in Exhibit 11.6. Here we can see the changes in the values of each variable: depth, for example, does look like it is increasing more or less linearly to the west, and longitude linearly to the upper right (see the directions of linear change of these variables in Exhibit 11.5). Latitude and temperature have the same but less linear pattern, and notice that the contours are in inverse directions: as the contours of latitude go up, the contours of temperature go down. For these two variables, the assumption of linearity may be rather too simple, but nevertheless Exhibit 11.5 did show that their correlation was negative. We could also add contours like these for selected species of interest.

Fuzzy coding of environmental variables

To capture possible nonlinear relationships between continuous variables and the MDS ordination, fuzzy coding offers an interesting alternative display. Let us code each of the environmental variables into four fuzzy categories, as described in Chapter 3. Then each category is placed on the ordination at its weighted average position, shown in Exhibit 11.7. Latitude and temperature can be seen to

[2] Notice that the signs of the MDS dimensions are random, so the axes can be reversed at will to facilitate the interpretation.

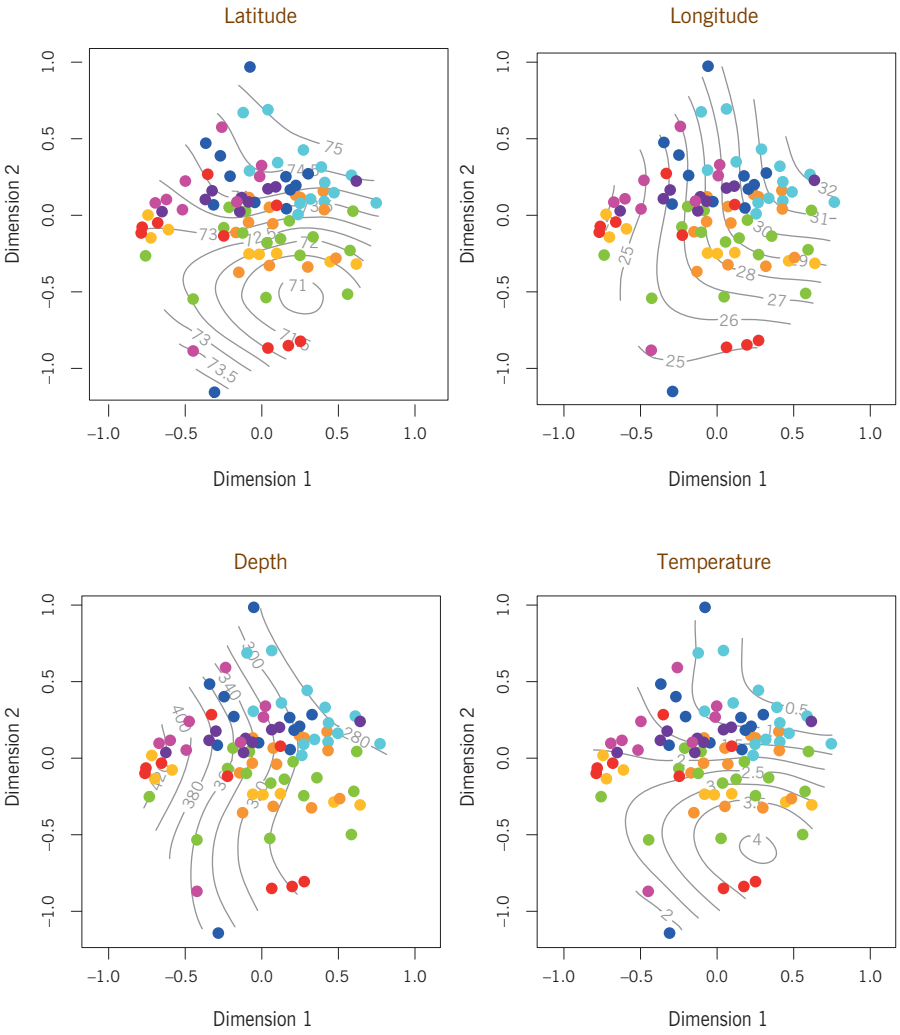


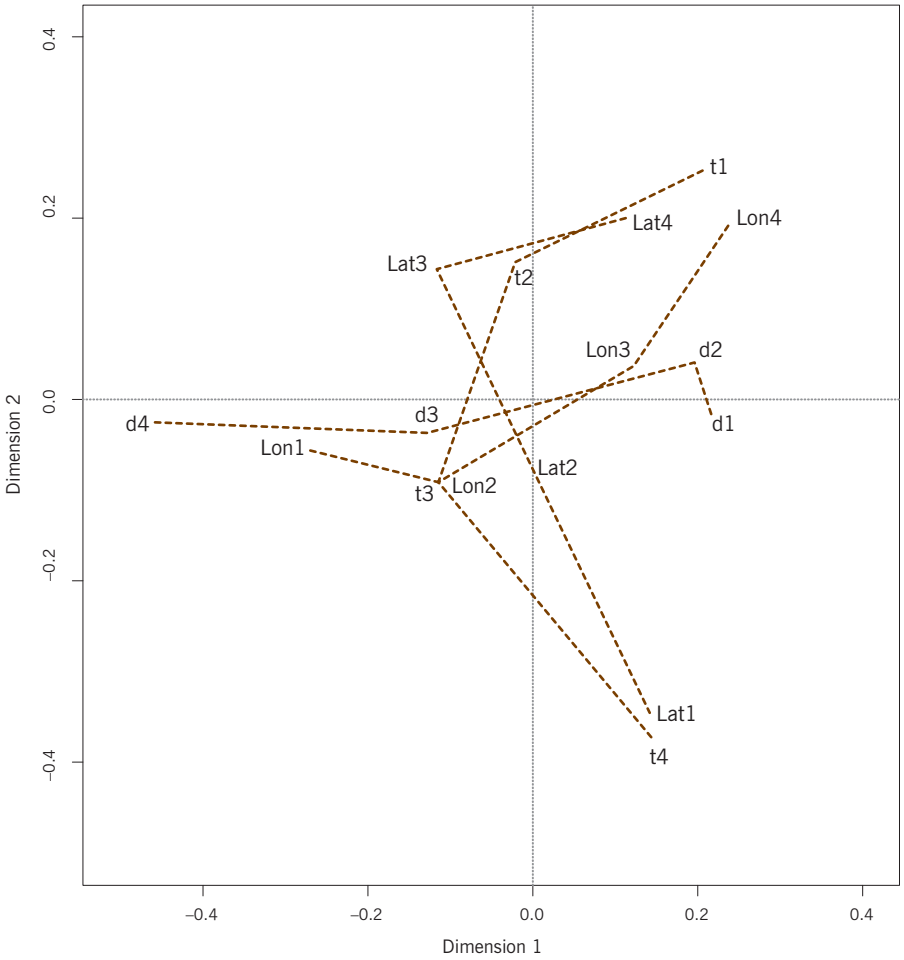
Exhibit 11.6: *Nonlinear contours of the four environmental variables showing their relationship with the two MDS dimensions*

have curved trajectories and going in the opposite sense, whereas longitude and depth have more straight-line trajectories, in directions similar to their gradient vectors in Exhibit 11.5.

Fuzzy coding of interactions and spatial coordinates

Each variable was regressed separately on the ordination axes, but there are situations when interactions are important (but they are not included here). For example, in Exhibit 11.7 it is clear that there is a nonlinear relationship between temperature and depth since "shallower" samples (categories `d1` and `d2`) are on the side of both high and low temperatures. However, this is not necessarily an interaction effect, which would be that the trajectory of depth

Exhibit 11.7: *Fuzzy categories of the four environmental variables, positioned at their respective weighted averages of the samples. The sample ordination is given in Exhibit 11.3, and linear relationships of species and variables in Exhibit 11.5*



categories differs depending on the temperature category. Another example is the pair of variables latitude and longitude: in both Exhibits 11.5 and 11.7, it is not possible to determine if the effect of longitude is contingent on latitude. In regular statistical modelling involving continuous variables, an interaction is coded by the product of the two variables and this product is included in the model as well as the linear terms. But for our purpose, showing gradient vectors in an MDS ordination for the linear terms of latitude and longitude and for their product is very difficult to interpret (the same problem occurs if one codes polynomial terms of latitude and longitude, which is often recommended to account for a nonlinear spatial component – the biplot representation of powers and cross-products of latitude and longitude coordinates is hard to interpret).

Fuzzy coding offers a better alternative, shown in Exhibit 11.8. The fuzzy coding of latitude and longitude (or any other pair of variables whose interaction needs to be explored) is computed, giving two sets of three numbers as shown, and then all pairs of these numbers give nine fuzzy categories coding the eight compass points and a central category. For two continuous variables such as depth and temperature, the eight outer categories would code, for example, depth and temperature high ("north-eastern" category), temperature near the centre, depth high ("eastern" category, if depth is considered on the horizontal axis of the scheme in Exhibit 11.8), temperature low and depth high ("south-eastern" category), and so on. Finally, for the spatial variables, Exhibit 11.9 shows the positions of the nine categories. The points are connected by lines which would form a square grid if no interaction were present. There is a clear interaction, with the eastern regions having less difference than the western regions. This is a richer result than showing latitude and longitude by two simple vectors, and reflects the more complex nature of the relationship of fish abundances to the geography.

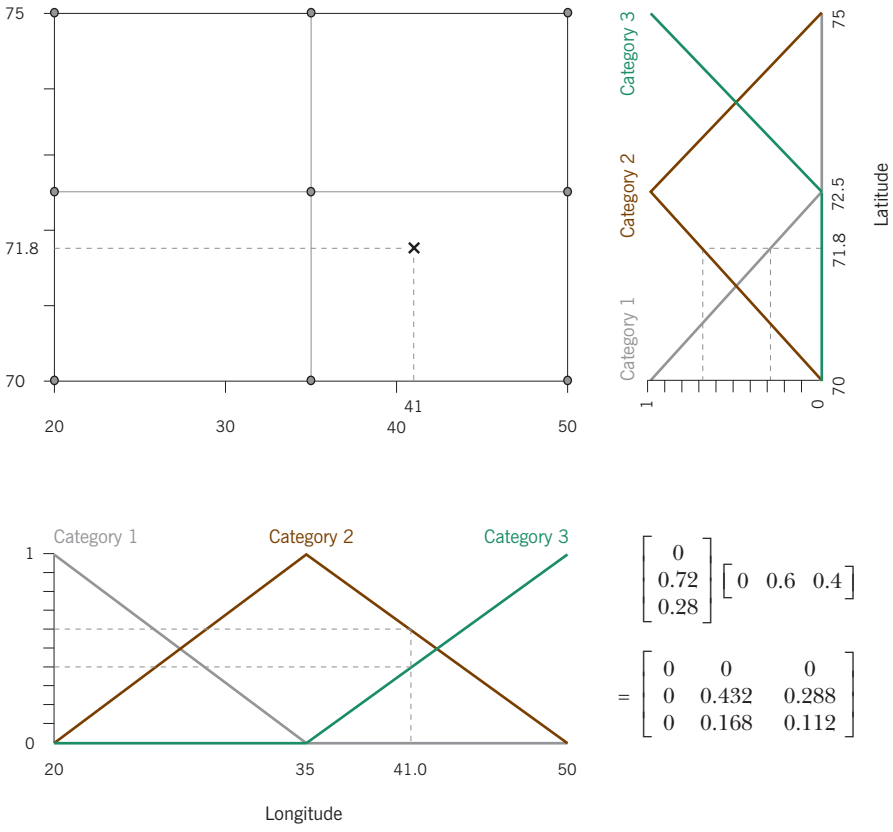
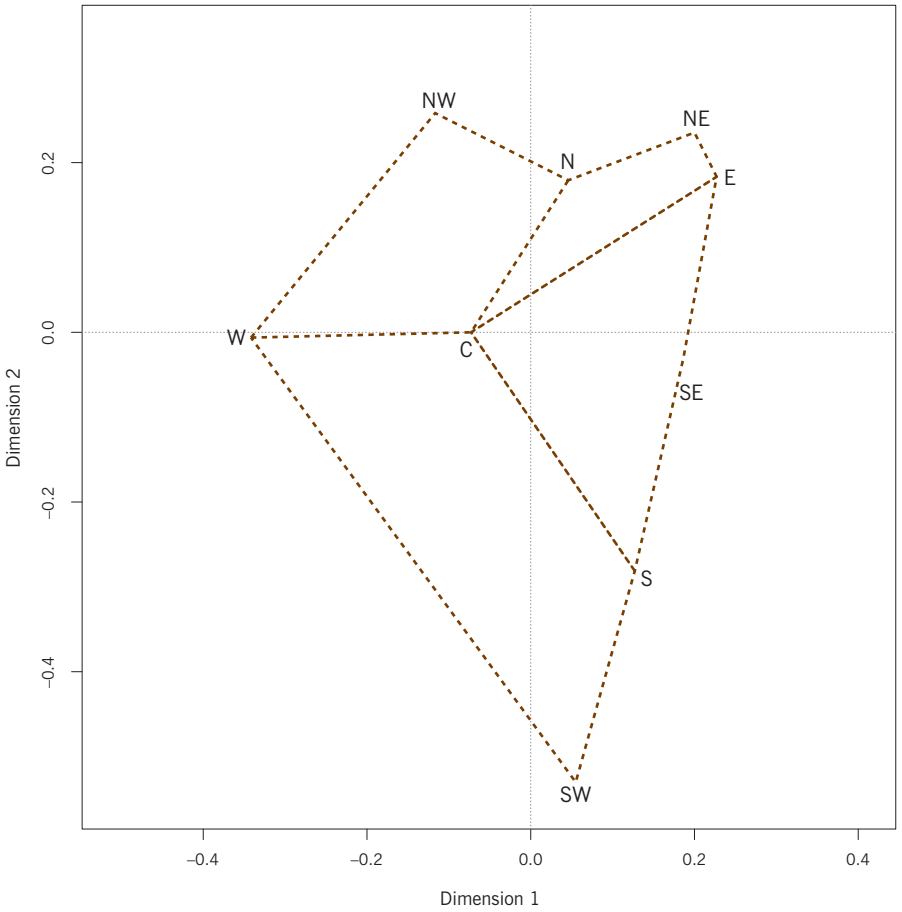


Exhibit 11.8: *Coding the latitude–longitude interaction into fuzzy categories: for example, each is coded into three fuzzy categories and then all pairwise products of the categories are computed to give nine categories coding the interaction. For example, the point with latitude 71.8°N and longitude 41°E has fuzzy coding [0.28 0.72 0] and [0 0.6 0.4] respectively. The first set is reversed to give values from north to south, and all combinations of the fuzzy values give nine categories coding the eight compass points and a central location*

Exhibit 11.9:
The positions of the nine fuzzy categories coding the interaction between latitude and longitude. Labels are the eight compass points, and C *for central position*



SUMMARY: Multidimensional scaling biplots

1. Multidimensional scaling (MDS) makes an ordination map of a matrix of proximities within a set of samples, based on variables observed in each sample, for example species abundances, environmental or geographical variables, and so on.

2. Once this ordination map is achieved, the variables on which it was originally based can be related to the dimensions of the map using the regression biplot approach, whereby regression models are fitted to each variable as a response and the ordination dimensions as predictors.

3. When the spatial coordinates of the samples are known, it is relevant to relate the ordination map to the spatial map. This can be done either by comparing inter-sample spatial distance with inter-sample distance in the ordination map, or by performing Procrustes analysis on the two configurations.

4. The nonlinear contours of a concomitant continuous variable observed on all the samples can also be visualized, one at a time, and compared to the straight-line contours implied by the regression biplot.
5. Fuzzy coding is useful for visualizing these nonlinear contours in a simple way that allows several variables to be visualized simultaneously.
6. Fuzzy coding can also be used to code interactions between continuous variables, such as latitude and longitude, and thus enrich the interpretation of the MDS result.

CHAPTER 12

Principal Component Analysis

In the biplots shown so far, there has been a two-step process in their construction. First, a scatterplot of the samples is made, using as support axes either two observed variables or two dimensions from an MDS. In the latter case the display has been optimized according to the objective function in the MDS, either to come as close as possible to reproducing the proximities (metric MDS) or to come as close to reproducing their ordering (nonmetric MDS). Given the MDS ordination, we have shown how different variables can be added using regression coefficients from various types of linear models or as weighted averages of sample points. The visualization of these variables is optimized conditional on the ordination – that is, the ordination is not necessarily the best ordination for explaining the variance of the added variables. In this chapter we present the first method that simultaneously optimizes both the ordination of the samples and the explained variance of the variables. The method, principal component analysis, applies to matrices of interval-scale continuous measurements.

Contents

The "climate" data set

Climate data are important in many ecological research projects, since they form part of the body of environmental indicators that can help to explain biological patterns. In a marine research project in Kotzbehue, Alaska, a set of annual climate variables were gathered together in a table, part of which is shown in Exhibit 12.1. These are annual data over 23 years, from 1981 to 2003. The variables

Exhibit 12.1: *Annual climate data for years 1981-2003, consisting of 17 climate indices and meteorological variables. Part of the 23 × 17 data matrix is shown*

Year	*AO*	*AO_winter*	*AO_summer*	*NPI*	*NPI_spring*	*NPI_winter*	*Temp*	⋯	*IceCover*	*IceFreeDays*
1981	–0.4346	–0.1683	–0.2410	–2.09	–0.15	–4.46	–3.9	⋯	–0.64	140
1982	0.2977	–0.3750	0.3083	0.75	0.13	1.70	–4.7	⋯	–1.65	144
1983	0.0319	0.1733	0.4653	–2.54	0.30	–5.44	–4.4	⋯	–0.34	116
1984	–0.1917	0.2627	0.0240	–1.20	–0.23	–2.62	–7.0	⋯	0.15	134
1985	–0.5192	–1.2667	0.2678	0.52	–0.43	1.11	–5.9	⋯	–0.21	120
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2002	0.0717	0.4543	0.0187	0.13	–0.18	0.30	–3.3	⋯	0.78	203
2003	0.1521	–0.6453	0.0399	–1.67	–0.40	–3.84	–3.8	⋯	–1.60	179
mean	0.0466	0.0587	0.1652	–0.440	0.023	–0.950	–5.15		–0.317	151.8
variance	0.1699	1.1687	1.0505	1.166	0.491	5.603	1.08		0.888	398.5

form three sets: climate indices such as the Arctic Oscillation (AO) and North Pacific Index (NPI),[1] meteorological variables such as temperature and rainfall, and various measures relating to Arctic ice such as ice coverage and number of ice-free days in the year.

MDS of the sample points

Each year is described by a set of 17 variables – each variable is assumed to be on an interval scale: this means that to compare two values of any particular variable it is the difference between two values that is important rather than the ratio or percentage difference. There are many different scales amongst the variables, some are pure indices without any scale, others are in units of degrees centigrade or centimetres of precipitation, and another is a count of days. To measure overall differences between the years based on this disparate set of variables, we need to equalize their scales in some way so that large values do not count more just because they are on a different scale – notice, for example, in the last two columns of Exhibit 12.1 the large differences across years in the *IceFreeDays* column (i.e., high variance) compared to the small differences in the *IceCover* column (i.e., low variance). As explained in Chapters 3 and 4, the most common way of equalizing the scales is to express each variable relative to its standard deviation so that all variables have equal variance. But, depending on the nature of the data, some other way may be preferred – see the discussion in Chapter 3 for alternatives. In some other situations, when all the variables are measured on the same scale, standardization might not be necessary, so that

[1] The Atlantic Oscillation Index, based on sea-level pressure differences, is positive when there is low pressure over the North Pole, keeping the cold air there, while it is negative when cold air is released southwards. The North Pacific Index measures interannual to decadal variations in the atmospheric circulation, which anticipate changes in sea surface temperatures.

the natural variances of the variables come into play and are not equalized in any way.

Once the variables are standardized, there are several possibilities for computing an overall measure of difference, i.e. distance or dissimilarity, between any two years. The most obvious choices are the sum (or average) of the absolute differences between the 18 variables (city-block distance) or the Euclidean distance. In principal component analysis the Euclidean distance is used, followed by classical MDS. We will use a very slight adaptation of the Euclidean distance, averaging the squared distances between the variables rather than summing them, so that the measure of distance is unaffected by the number of variables included. For the first two years in Exhibit 12.1, the distance between them is computed as follows:

$$d_{1981,\,1982} = \sqrt{\frac{1}{18}\left[\frac{(-0.4346-0.2977)^2}{0.1699} + \frac{(-0.1683-(-0.3750)^2}{1.1687} + \cdots + \frac{(140-144)^2}{398.5}\right]} = 1.277$$

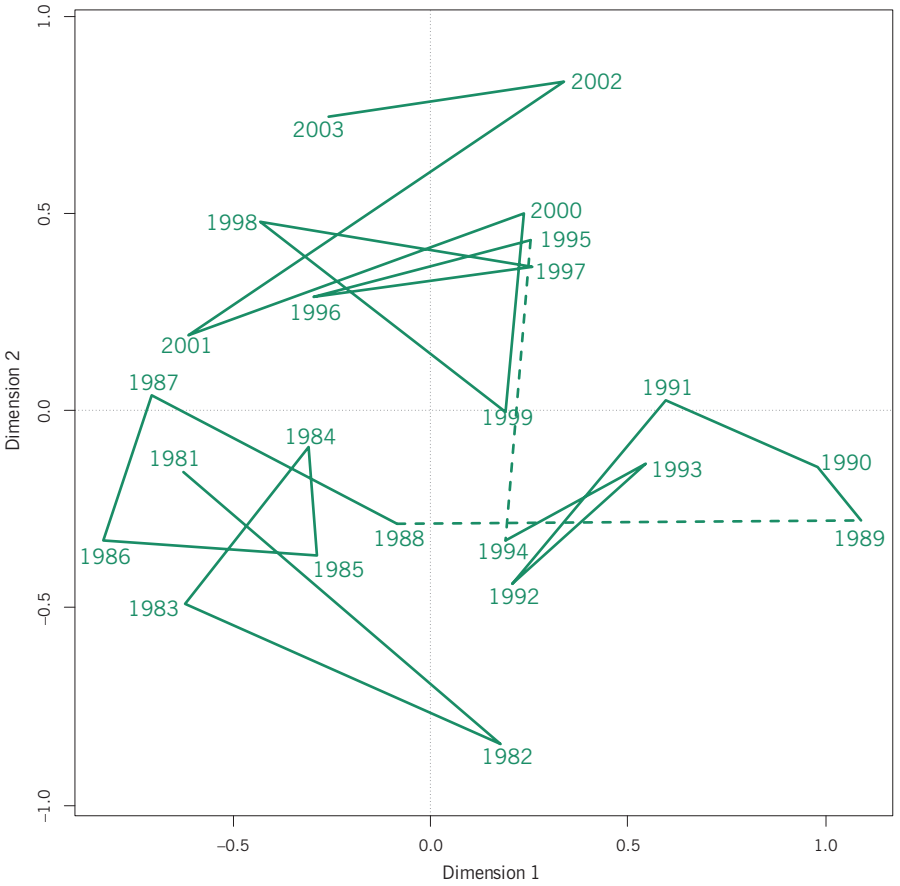
The denominators 0.1699, 1.1687, …, 398.5 are the variances of the variables *AO*, *AO_Winter*, …, *IceFreeDays*, so that the values inside the square of each numerator are divided by the corresponding variable's standard deviation. As described in Chapter 4, this distance function is called the *standardized Euclidean distance*.

Applying classical MDS to these distances, Exhibit 12.2 is obtained, explaining 45.6% of the variance. The years have been connected in sequence and there seem to be big changes in the climate variables from 1988 to 1989 and from 1994 to 1995, shown by dashed lines. Thus three groups of climate "regimes" are apparent, from 1982 to 1988 bottom left, then 1989 to 1994 on the right and finally from 1995 to 2003 in the upper section of the map. Next, adding the variables to the map will give the interpretation for these groupings.

Adding the variables to make a biplot

The 17 standardized variables are now regressed on the MDS dimensions, and their gradient vectors of regression coefficients are shown in Exhibit 12.3. The reason why the three groups of years separate is now apparent. The first period from 1982 to 1988 is characterized by high ice at all times of the year and to a lesser extent low winter temperatures (remember that the longer vectors here, corresponding to higher regression coefficients, will be the more important variables to interpret). In this period the climate indices, which are pointing to the right, will generally be below average. The period from 1989 to 1994, especially 1989, show a relatively sharp increase in all the climate indices. Then for the period 1995 to 2003, the climate indices move generally to average values

Exhibit 12.2: *MDS map of the 23 years according to the standardized Euclidean distances between them, across 17 climate variables. Variance explained by the two dimensions is 27.8% and 17.8%, totalling 45.6%*



again but the annual temperature is generally higher, total ice is lower and the number of ice-free days higher.

Principal component analysis

The MDS solution for the years and the addition of the variables by regression once again looks like a two-step process, but what we have done is in fact the principal component analysis (PCA) solution of the climate data set, which is a one-step process. The difference between this analysis and all the other two-step analyses described before in this book is that here both the display of the cases and the display of the variables are simultaneously optimized. If one computes the overall variance explained of the 17 variables by the two MDS dimensions in this case, one gets exactly the same percentage of variance, 45.6%: 27.8% by the first dimension and 17.8% by the second. To summarize, PCA of a cases-by-variables data matrix can be thought of as an MDS of the Euclidean distances between the cases plus the regressions of the variables on the dimensions of the MDS solution.

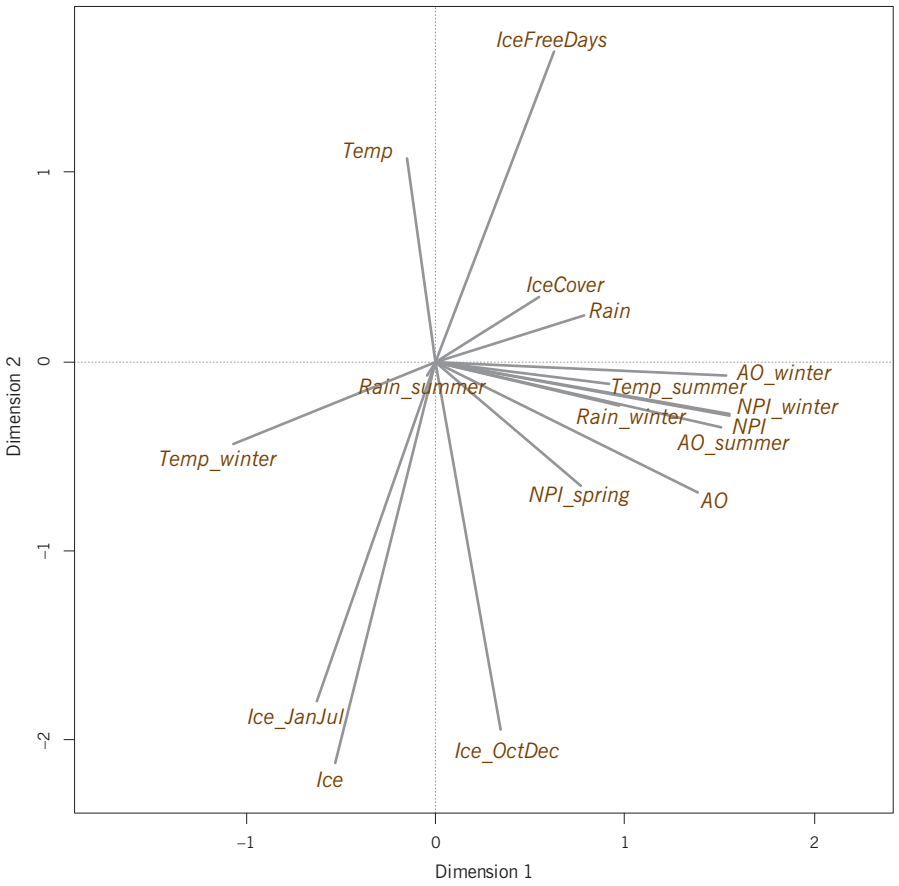


Exhibit 12.3: *Regression relationships of the variables with the two dimensions of the MDS map in Exhibit 12.2. Superimposing this configuration on Exhibit 12.2 would give a biplot of the years and the variables. This would be the so-called* row-principal *biplot, explained on the following page*

To compute a PCA it is not necessary to do these two consecutive steps: they can be done in a single step using a famous theorem in mathematics called the *singular value decomposition,* or SVD. This result is similar to the eigenvalue-eigenvector theorem for square matrices but applies to rectangular matrices. Applying the SVD to a matrix results in a least-squares approximation of the matrix of a lower rank, where rank is the algebraic equivalent of dimensionality. In the application to the climate data, the SVD provides a rank 2 approximation to the 17-dimensional standardized data matrix, and it is this two-dimensional approximation that is represented in Exhibits 12.2 and 12.3. The approximation explains 45.6% of the variance in the original matrix, and this is the same if one thinks of the explanation of the row points (i.e., the years as displayed in Exhibit 12.2) or the variables (i.e., the climate variables as displayed in Exhibit 12.3). Thus, Exhibit 12.2 explains 45.6% of the (squared) Euclidean distances between the rows, and at the same time the two dimen-

sions of the solution are predictors of the 17 variables, also explaining 45.6% of their total variance.

Scaling of the solution

There is one subtle but important difference, not discussed before, between the regression biplots of Chapter 10 and the MDS and PCA biplots in Chapter 11 and in Exhibits 12.2 and 12.3. In the regression biplots of Chapter 10 the support space of the sample points was constructed using standardized variables, with variance one on both axes, whereas in Chapter 11 and this chapter so far, the samples had different variances on the axes. In Exhibit 12.2, for example, the variances along dimensions 1 and 2 were 6.12 and 3.90 respectively, explaining 27.8% and 17.8% of the total variance of the sampled years. This means that the year points are more spread out horizontally than vertically, although this may not appear obvious in Exhibit 12.2. In other examples there may be a much greater disparity between the horizontal and vertical spread of points in the distance map, in which case the discussion in this section is more of an issue.

To mimic the regression biplots of Chapter 10, we could rescale the coordinates of the sample points (i.e., year points here) in the MDS map to have unit variance on both dimensions, and then add the variables by regression. This will not affect the variance explained in the regression analyses but has some advantage in that the gradient vectors are then standardized regression coefficients and more easily compared.

Let us introduce some terminology that will be essential in future descriptions of different types of biplot. If a set of points, row or column points, has equal sum of squares on each dimension (usually equal to 1, but not necessarily), we call their coordinates on the dimensions *standard coordinates*. If they have sum of squares equal to (or proportional to) the variance explained by the dimensions, then their coordinates are called *principal coordinates*. In the case of PCA, the Exhibit 12.2 displays the year points (rows) in principal coordinates, and Exhibit 12.3 displays the variables (columns) in standard coordinates. The superimposition of these two displays is a true biplot, called the *row-principal biplot*. The other possibility, given in Exhibit 12.4, is the *column-principal biplot*, where the years are in standard coordinates, and the variables in principal coordinates. In this case it may seem hardly different to the combination in Exhibits 12.2 and 12.3, apart from the scale on the axes, so we should clarify the difference in interpretations of the two alternatives. In the row-principal biplot obtained by superimposing Exhibits 12.2 and 12.3, the year points are a spatial approximation of the inter-year Euclidean distances. In the column principal biplot of Exhibit 12.4 where the year coordinates have been standardized, this distance approximation property is not true any more. In Exhibit 12.4 the focus is on the climate variables and their spatial properties, in particular the angles

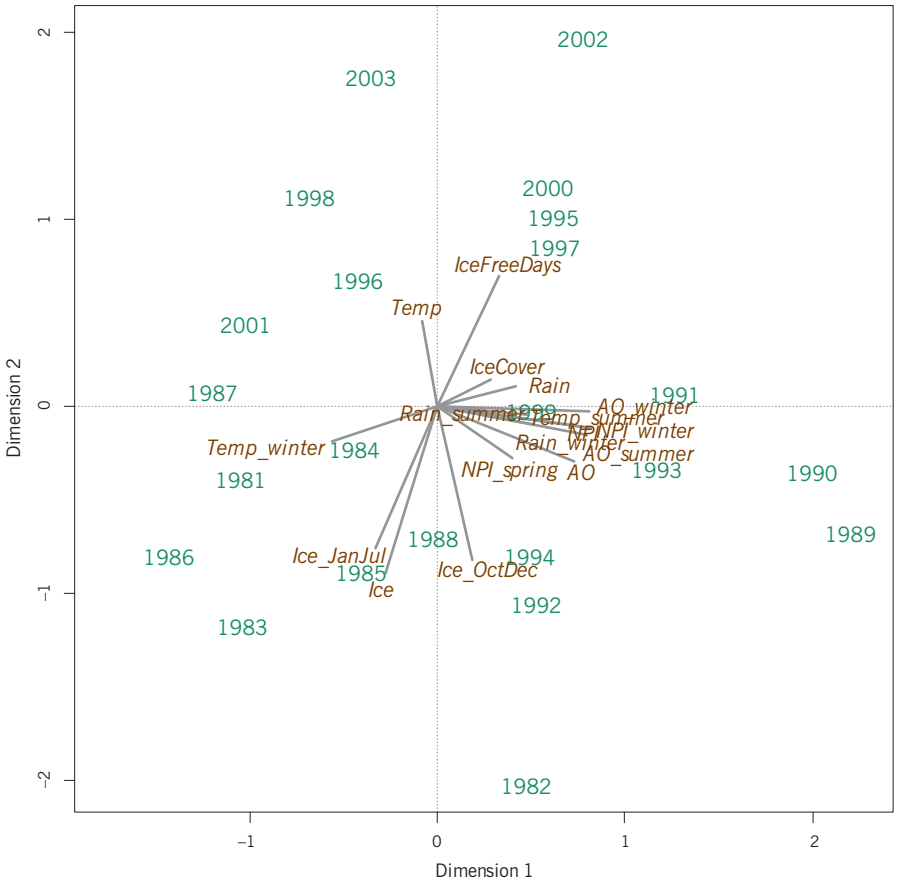


Exhibit 12.4:
Column-principal biplot of the climate data. Here the year points have coordinates that are standardized, while the sum of squares of the variable points on each dimension is proportional to the variance explained

between them, which have cosines that are approximately equal to the pairwise inter-correlations (see Chapter 6).

However, in Exhibit 12.4 the coordinates of the columns are standardized regression coefficients. In addition, because the two support dimensions are uncorrelated, the standardized regression coefficients are identical to the correlation coefficients of the column variables with the dimensions. For example, in Exhibit 12.4, the variable *IceFreeDays* can be seen to have a correlation with the first and second dimensions of approximately 0.3 and 0.7, respectively. In the same display sample points lie more or less within plus/minus two units, that is two standard deviations (because they are standardized), whereas the column points all have absolute values less than one (because their coordinates are correlations). One final remark: the sample points have means equal to zero, but the variable points are not centred. This is why, for

the biplot in Exhibit 12.4, the legend says that the sample coordinates are standardized (mean zero, variance one), whereas the sum of squares of the variable points on each dimension is proportional to the respective variance explained.

The circle of correlations

In the classical MDS of the Euclidean distances of this 17 variable problem, there are a maximum of 17 dimensions in the MDS solution, of which Exhibit 12.2 shows the best two corresponding to the two largest eigenvalues. (If there were 17 or fewer samples the maximum dimensionality of the solution would be one less than the number of samples.) Because the 17 "response" variables are standardized, a property of the standardized regression coefficients computed on standardized dimensions (where the coefficients are, we repeat, the correlations of the variables with the ordination axes) is that their sum of squares over all the dimensions for a particular variable is equal to 1, in other words the variable is fully explained by the complete set of MDS dimensions. In Exhibit 12.4, which shows the standardized regression coefficients with respect to the first two dimensions only, the sum of squares of the two coordinates for each variable, in other words the squared length of each vector shown, is equal to the proportion of variance explained for the respective variable. So we can draw a unit circle around the variable vectors and the variables that are better explained by the dimensions will be longer and closer to the unit circle. Exhibit 12.5 shows this circle. For example, the variable *Ice*, which lies close to the unit circle, has a very large part of its variance explained by the dimensions – the percentage is actually 88% – whereas *Temp* has only 21.2% (the length of the *Temp* vector is just under 0.5, and the length squared is the part of variance explained).

Having explained this relationship between the squared coordinates and the parts of variance explained for each variable, it follows that the average of the squared coordinates on each axis is equal to the part of variance (for all variables) explained by the axis: these averages are computed to be 0.278 and 0.178 respectively – see the caption of Exhibit 12.2.

In addition, as mentioned before, the cosines of the angles between the variables in Exhibit 12.5 are approximations of the correlations between the variables, and the approximation is improved when the variables are well explained, that is close to the unit circle. Thus we can be pretty sure that *IceFreeDays* and *Ice* are negatively correlated – the actual correlation is –0.57, the second most negative correlation amongst the variables. However, look at Exhibit 6.2 again – in order to see the correlation exactly as the angle cosine, we would need to see the two vectors *IceFreeDays* and *Ice* in their actual positions, not projected down onto this approximate MDS map.

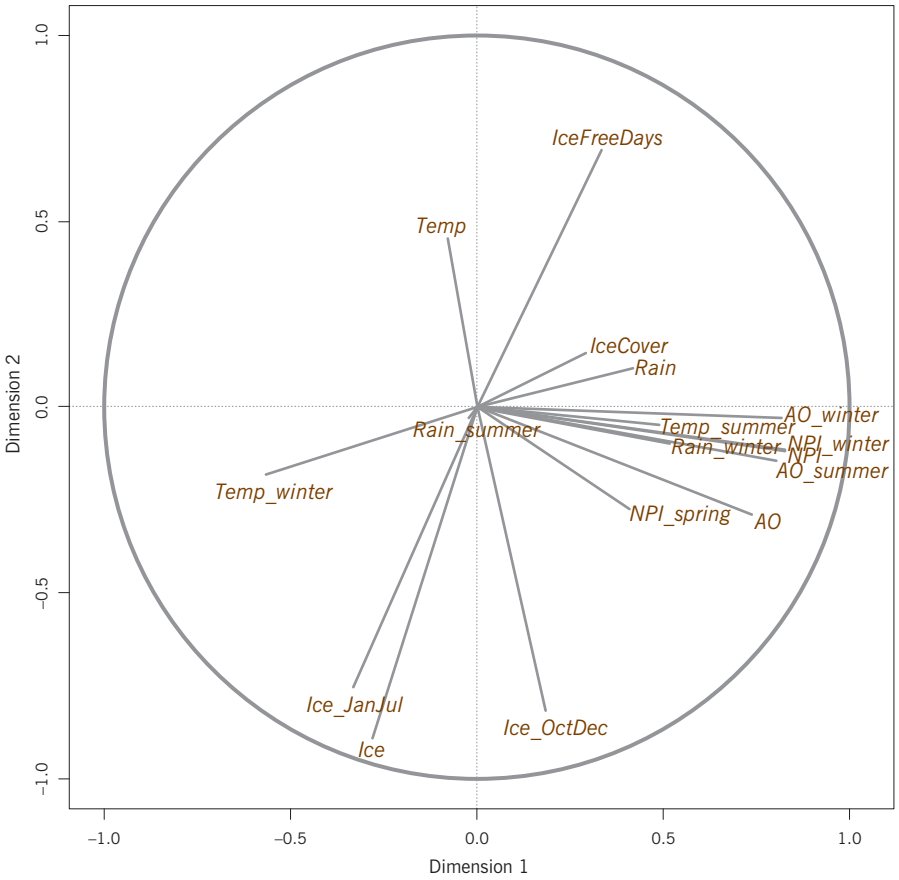


Exhibit 12.5:
Plot of the variables as in Exhibit 12.4, that is as standardized regression coefficients (i.e., principal coordinates in this PCA, which are the correlations between the variables and the dimensions), all lying within the unit circle. The closer the variable vector is to the unit circle, the better it is explained by the dimensions. The angle cosines between the vectors also approximate the correlations between the variables

Deciding on the dimensionality of the solution

Mostly for convenience of plotting, the biplots from PCA and MDS are shown with respect to the best two *principal axes*. But are these axes "significant" in the statistical sense? And what about other dimensions? If the third dimension of the solution were also "important", we could try to use three-dimensional graphics to visualize the biplots. But before going to these lengths we need a way of deciding how many dimensions are worth interpreting. One of the simplest ways of judging this, albeit quite informal, is to make a so-called *scree plot*, that is a bar graph of all the eigenvalues, or parts of variance, of the solution and then look for the so-called *elbow* in the plot. In Exhibit 12.6 on the left we show the scree plot of the eigenvalues of the present 23 × 17 example of climate data, whereas the scree plot on the right is of a 23 × 17 matrix of normally distributed random variables. Both PCAs have the same total variance of 17, equal to the number of variables, since each standardized variable has variance 1. Notice how for the random data the eigenvalues fall off gradually in value, whereas for

the climate data the first two stand out from the rest. From eigenvalue 3 onwards the values in Exhibit 12.6(b) are similar or greater than those in Exhibit 12.6(a), so it does seem that the first two dimensions of the climate data are nonrandom. Later in Chapter 17 we will make formal tests for dimensionality, based on the same method of comparing eigenvalues obtained in a PCA with those from PCAs of randomly generated data.

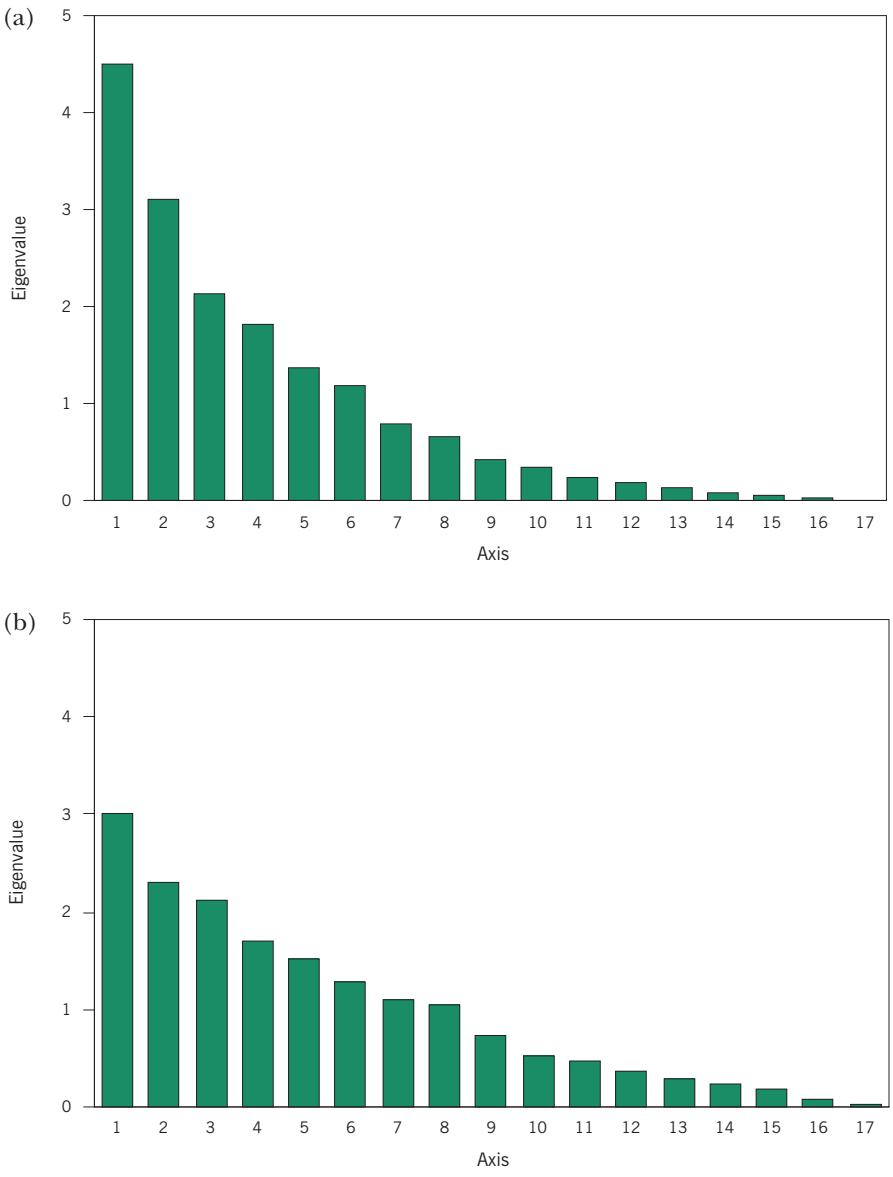


Exhibit 12.6: *Scree plots of the eigenvalues for (a) the climate data matrix; (b) a random data matrix*

SUMMARY: Principal component analysis

1. Principal component analysis (PCA) can be thought of as a multidimensional scaling (MDS) of a sample of multivariate observations, followed by the addition of the variables to the MDS map by linear regression on the dimensions, to give a biplot. Proximities between the multivariate sample points are defined using Euclidean or weighted Euclidean distance.

2. The special feature of PCA is that the visualization of both the sample points and the variables is optimized simultaneously: that is, the MDS optimally displays the sample points by least-squares and the dimensions are at the same time the best ones for predicting the variables by least-squares regression. In fact, the PCA solution is obtained in a single computational step, rather than a two-step MDS and regression approach.

3. There are two ways of displaying the results of PCA in the form of a biplot, differing only by scale factors of the sample (row) and variable (column) coordinates: the row-principal biplot and the column-principal biplot.

4. In the row-principal biplot the sample points are scaled in principal coordinates, as they would be from the MDS solution: they have mean 0 on each principal axis and their variances on each axis are the parts of variance explained. The variables added by regression will have equal variance on the axes (usually equal to 1) and their coordinates are thus called *standard coordinates*. Visually, the sample points will be spread out more on the first (horizontal) axis than on the second (vertical) axis, whereas the variables will be equally spread out on the two axes.

5. In the column-principal biplot the sample points are standardized on each principal axis, usually to have variance 1, in which case the regressions of the (standardized) variables on the principal axes are standardized regression coefficients, identical to the correlations between the variables and the axes. Now the variables, which are in principal coordinates, will be more spread out on the first axis than the second, whereas the sample points are equally spread out on the axes.

6. In the column-principal biplot the variables can be depicted as vectors inside a unit circle: the closer they lie to the circle the better they are explained by the principal dimensions. Also the angle cosines between the vectors are approximations of the correlations between them.

7. PCA is a dimension-reduction technique that attempts to separate "signal" (i.e., true structure) in the data, from "noise" (i.e., random variation), concentrating the signal in the first principal axes. Choosing how many axes are nonrandom can be performed informally by inspection of the scree plot of the eigenvalues in descending order, observing which eigenvalues stand out from the rest.

CORRESPONDENCE ANALYSIS

CHAPTER 13

Correspondence Analysis

Correspondence analysis is one of the methods of choice for constructing ordinations of multivariate ecological data. Ecological data are often collected as counts, for example abundances, or other positive amounts such as biomasses, on a set of species at different sampling sites. Correspondence analysis is similar to PCA, but applies to data such as these rather than interval-scale data. It analyses differences between *relative* values: for example, if at a sampling site there is an overall abundance count of 320 individuals across all the species, and if a particular species is counted to be 55, then what is relevant for the analysis is the relative value of the abundance, 55/320 (17%). Furthermore, to measure inter-sample difference in such relative abundances across the species, the chi-square distance described in Chapter 4 is used to normalize species with different overall abundances. Classical MDS is applied to the chi-square distances to obtain an ordination of the sample points, with one important difference compared to PCA: each sample point is weighted proportionally to its total abundance (e.g., the value 320 mentioned above as an example), so that samples with higher overall abundance are weighted proportionally higher. These sample weights are also used in regressing the species on the dimensions to obtain a biplot. Finally, correspondence analysis has the special property that the analysis can be equivalently defined, and thought of, as the analysis of the rows (e.g., samples) or the analysis of the columns (e.g., species).

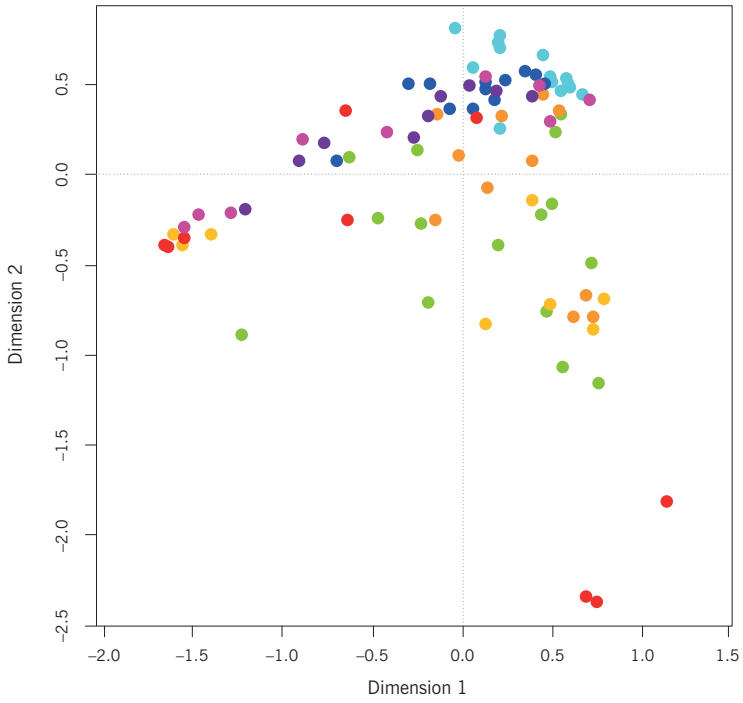
Contents

Weighted MDS of chi-square distances

In Chapter 11 we made a two-step analysis of the "Barents fish" abundance data: first, the Bray-Curtis dissimilarities between sampling sites were computed and a nonmetric MDS performed, and second, the species counts were regressed on the ordination dimensions using Poisson regression to obtain a biplot of the samples and species. These regressions were optimal conditional on the ordination obtained in the first step, so the question to consider now is what the ordination should be in two dimensions, say, in order that the regressions are the best that one can get using the two ordination axes as predictors. Like PCA, correspondence analysis, abbreviated as CA from now on, is going to be doubly optimal: the display of the sample points will be optimal and the biplot of the samples and species will be optimal in that the species regressions will explain maximum variance. One major difference in the CA approach is that it measures distance between the profiles of the abundances (i.e., vectors of relative abundances), described in Chapter 4 – see Exhibit 4.6 and the surrounding description. Then it uses the chi-square distance function between the profiles – see Exhibit 4.7 and its surrounding description. Furthermore, the sense of the optimality is by *weighted* least-squares in both the MDS of the sample profiles and in the regressions of species on the ordination axes – the sample weights are proportional to the abundance totals at the different sampling points. For example, the abundance totals at the 89 sites (see Exhibit 11.2) are 845, 1,740, 1,763, 767,..., 232, 36, with a grand total of 63,896. The weights, which are positive and add up to 1, will be 845/ 63,896 = 0.0132, 1,740/63,896 = 0.0272, and so on, until 232/63,896 = 0.0036 and 36/63,896 = 0.0006. Thus sites with higher abundances will be weighted more than those with lower abundances: for example, the profile of the second site will get a weight of 0.0272 (2.72%) whereas the last site, where overall abundance was low, will get a weight of 0.0006 (0.06%).

In the previous description of MDS methods there was no question of weighting the points, in other words all were weighted equally. It is a fairly simple adaptation of the methodology to accommodate different weights, which means that points with higher weight will tend to be better displayed than points with lower weight. This reweighting can make a big difference to the final MDS solution, as illustrated for this particular data set in Exhibit 13.1. In the unweighted MDS there is a curve of points from the left across to the top and then down to the three red points. This curve is essentially reproduced on the right hand side of the weighted MDS, following the vertical axis from top to bottom, but two samples have separated out on the left. These latter samples have high abundances, and so have high weights and become much more prominent in the weighted analysis. The ordination map in Exhibit 13.1(b) is based on the CA solution.

(a)



(b)

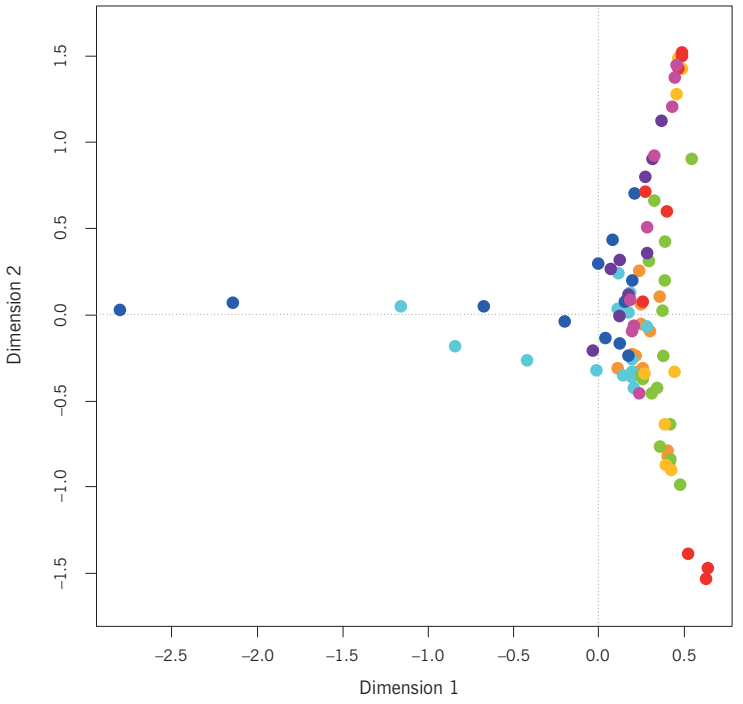
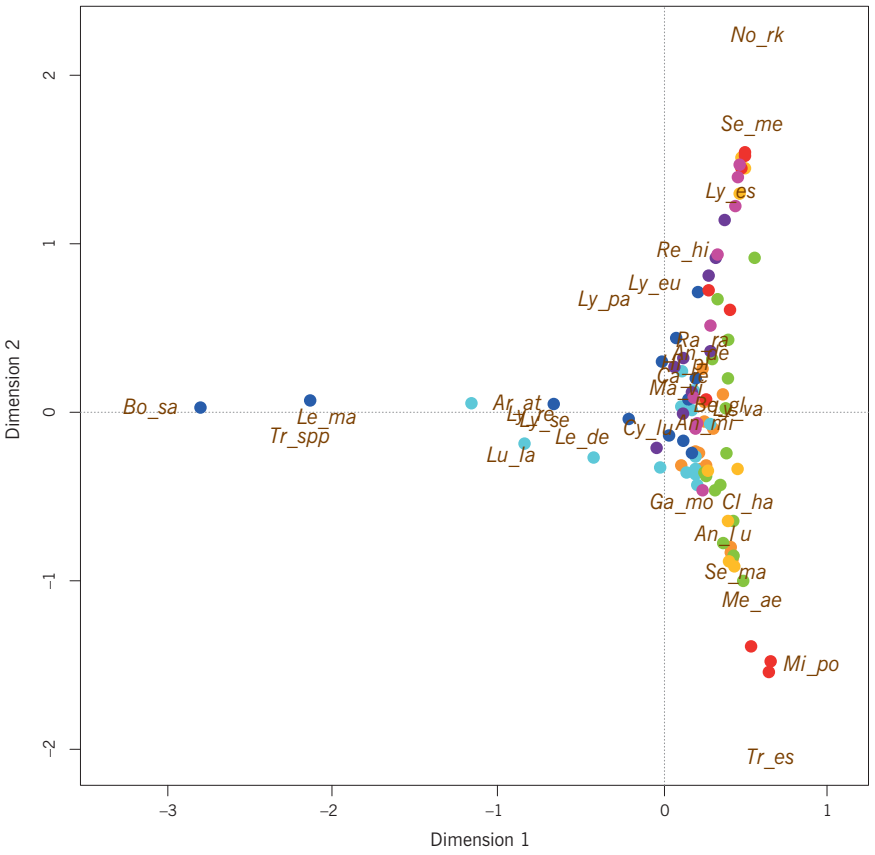


Exhibit 13.1:
Unweighted MDS (a) and weighted MDS (b) of the chi-square distances between sampling sites, for the "Barents fish" data. Colour coding as in Chapter 11

Display of unit profiles

CA has some special features that are not present in PCA, chiefly because the displayed profiles consist of nonnegative values that add up to 1. These are also called *compositional data*, the proportions of the species in each sample. One of the classic ways of displaying the species in CA is to show where the so-called *unit profiles* are in the ordination space: these are vectors of zeros except for a 1 in the position corresponding to a species, as if there were a sample with only that species observed in it. The unit profiles for the species are shown as supplementary points in Exhibit 13.2. Thus the unit point of the species *Bo_sa* (*Boreogadus saida,* polar cod) is on the extreme left, and species points *Tr_spp* (*Triglops* species) and *Le_ma* (*Leptoclinus maculatus,* spotted snake blenny) are also separate from the others on the left hand side. It turns out that the sample on the extreme left has very high relative abundances of these three species, and this explains its outlying position in the direction of these species. These high distances are not apparent in Exhibit 13.1(a) because all the points are weighted equally, whereas CA gives prominence to the samples with higher weight, and

Exhibit 13.2: *Row-principal CA biplot (asymmetric map) of "Barents fish" data. The sample profiles are shown as well as unit profiles for the species. There is a barycentric (weighted average) relationship between the samples and species points. Explained variance is 47.4%*



Bo_sa is a species with high overall abundance. The CA solution in Exhibit 13.2 is sometimes called an *asymmetric map:* the sample (row) points are in principal coordinates and it turns out that the species (column) points are in standard coordinates in the CA sense. To make this more precise, each species also has a weight in CA, its total abundance relative to the grand total. In Exhibit 11.2 the last species *Tr_spp*, for example, has an abundance of 653, which gives it a weight of 653/63,896 (1.02%). With these weights the species points in Exhibit 13.2, representing unit profiles, have weighted sum of squares equal to 1 on each dimension, and thus have coordinates referred to as standard coordinates.

Barycentric (weighted average) relationship

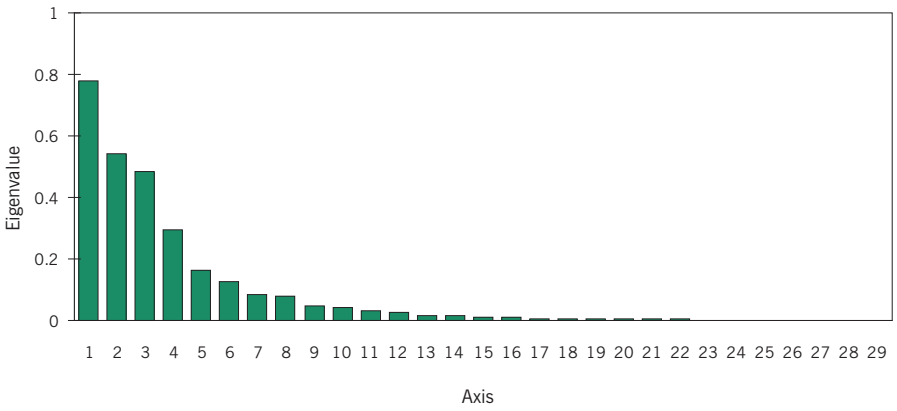
Before continuing let us start to call the sample and species weights *masses*, which is the preferred term in CA. This also distinguishes these masses from other sets of weights which we discuss now. For example, the masses of the 30 species in the "Barents fish" data are the relative abundances of the species in the whole data set. So the masses reflect the expected, or average, relative abundances in a sample if there were no differences in species distribution across the study region.

The joint display of samples and species in Exhibit 13.2 has an additional property that is particular to CA and is, in fact, one of the reasons for its relevance in ecology. Each sample point, originally the profile of the sample across the species, is at the weighted average of the species points, where weights are defined here as the elements of the profile. Let us take the sample on the extreme left of Exhibit 13.2 as an example. This sample has a total abundance of 4,399 and its profile across the 30 species consists of 19 zeros and 11 positive values, of which a few are extremely high compared to the species masses. For example, 82.9% is in the species *Bo_sa* (3,647 out of 4,399), whereas the relative abundance (i.e., mass) of *Bo_sa* in the whole data set is only 8.3% (5,297 out of 63,896). The sample is situated at the weighted average of the species points, and 82.9% of its weight is on *Bo_sa*, hence its position close to it. It has also much higher than average relative abundances of *Tr_spp* and *Le_ma*. For the same reason, the three sample points at bottom right must have high values in their profiles on the species *Tr_es* and *Mi_po* in order to be situated so close to them in that direction. Weighted averages are also called *barycentres* and this relationship between sample and species points in this version of the CA solution is called the *barycentric relationship*.

Dimensionality of CA solution

CA also leads to an eigenvalue measure of the part of variance on each dimension, as in PCA, and these eigenvalues can be viewed in a scree plot, shown in Exhibit 13.3. Here we introduce some terminology particular to CA: the total variance is called the total *inertia* of the data set, and is equal to 2.781 in this case. The eigenvalues, or *principal inertias*, decompose this total along the prin-

Exhibit 13.3:
Scree plot of eigenvalues in the CA of the "Barents fish" data



cipal axes. Notice that there are only 29 eigenvalues – the dimensionality of the full space is not 30, the number of species, but one less because the profile matrix analysed has constant row sums of 1. To decide on how many dimensions are worth interpreting we proceed as in PCA: it looks like there may be at most four dimensions distinguishing themselves from the others that tend to fall off in a pattern typical of random data. Later in Chapter 17 we will show more formally by a permutation test that in fact there are only three highly significant dimensions. So we should be looking at the third dimension as well. This poses a technological challenge, but it is now fairly easy to observe three-dimensional displays. In Exhibit 13.4 is a snapshot of the three-dimensional view of the points, and if you click on the image in the electronic version of this book it will revolve around the vertical axis.

Contribution biplots

The caption of Exhibit 13.2 refers to the display as a row-principal biplot, but this is not exactly the same as the regression biplots discussed before. In Chapters 11 and 12 the standardized variables were regressed onto the axes using ordinary least squares. Here there are two differences: firstly, the fact that the chi-square distances between profiles are being displayed, and secondly, the fact that each sample is weighted differently according to its corresponding mass. Thus it should be the columns of the standardized profile matrix that are regressed on the axes, standardized by centring with respect to the average profile (in this case, the set of species masses) and dividing columns by the square root of the corresponding masses, i.e. the standardization inherent in the chi-square distance. This gives another version of the biplot which we call the *contribution biplot*, shown in Exhibit 13.5 – just the species vectors are shown, the sample points are identical to those of Exhibit 13.2. With this scaling the species that are the most outlying on the axes are the ones contributing mostly to the CA solution, and thus the important ones for interpretation.

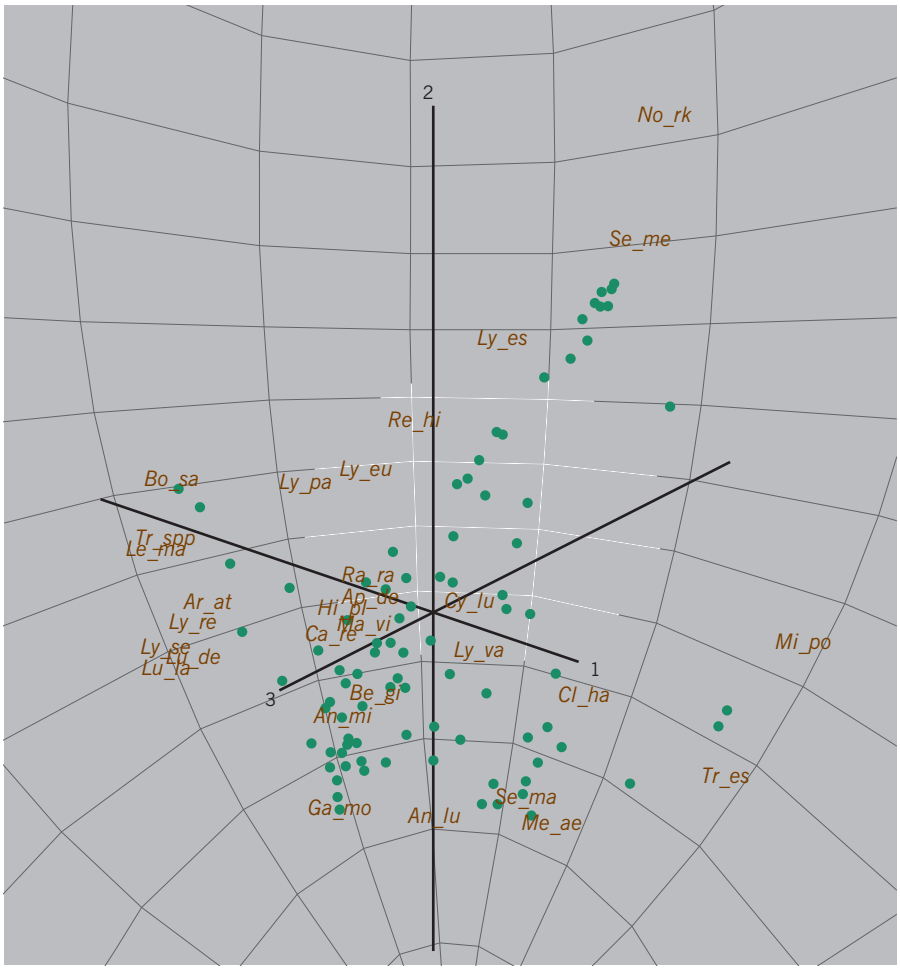


Exhibit 13.4:
Three-dimensional view of the samples and species, row principal biplot scaling. For readers of the electronic version: To see the rotation of these points around the vertical (second) axis, click on the display

Notice first the technical difference between the scalings of the species in Exhibits 13.2 and 13.5. In Exhibit 13.2 the standard coordinates have weighted average sum of squares equal to 1 on each ordination axis, using the species masses. In Exhibit 13.5 the contribution coordinates have unweighted sum of squared coordinates equal to 1 on each axis. These squared coordinates are the part contributions to the respective axes and are thus called *contribution coordinates*. In Exhibit 13.5 the species are shown as gradient vectors, and are oriented in the exact same directions as the unit profiles in Exhibit 13.2, but each species point has been pulled in by different amounts, with the rarer species being pulled in more than the more abundant ones. The exact relationship between the two types of species coordinates is that the contribution coordinates in Exhibit 13.5 are the standard coordinates in Exhibit 13.2 multiplied by the square roots of the respective species masses.

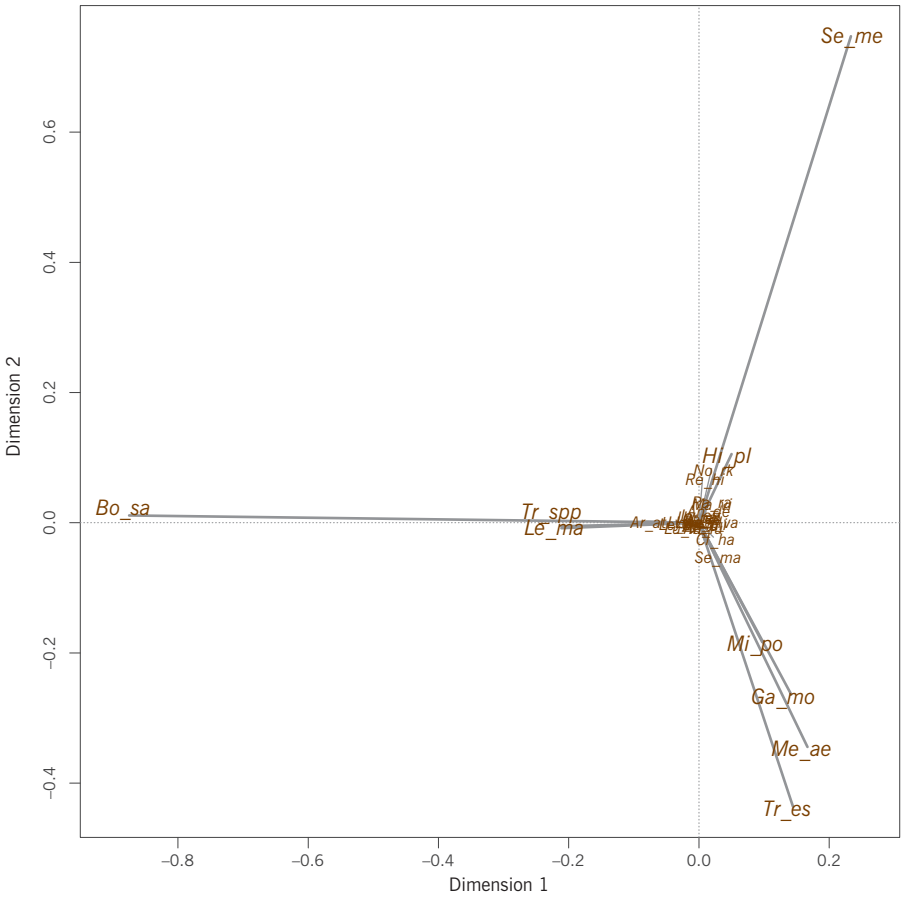


Exhibit 13.5: *Species in contribution coordinates. Combining this configuration with the sample points in Exhibit 13.2 would give the two-dimensional contribution biplot. The species that contribute more than average to an axis are shown in larger font (contributions to all three significant dimensions are taken into account here – the species Hi_pl contributes highly to the third dimension). Those near the origin in tiny font are very low contributors to the CA solution*

When it comes to the interpretation, the species *No_rk* and *Se_me* are ones that exemplify the difference between Exhibits 13.2 and 13.5. *No_rk* is a quite rare species in the data set, only 83 counted out of the total of 63,896, whereas the overall count of *Se_me* is 12,103. Thus *No_rk* is pulled in very strongly from its unit profile position in Exhibit 13.2 to its inlying position in Exhibit 13.5 – whereas it looked like it was the most important point before, it is now one of the species near the origin that are shown with tiny labels. By contrast, *Se_me* is not pulled in so strongly because of its high mass and in Exhibit 13.5 is confirmed to be the most important contributor to that spread of the samples upwards on the second axis. Both versions of the CA ordination are useful: from Exhibit 13.5 we know that *Se_me* is a strong contributor while Exhibit 13.2 tells us that the much sparser data for *No_rk* still correlates with that of *Se_me*. Another way of thinking about the high contributors, nine species in all in Exhibit 13.5, is that we could remove the other 21 species from the data set and get more or less the same result. To illustrate this, Exhibit 13.6

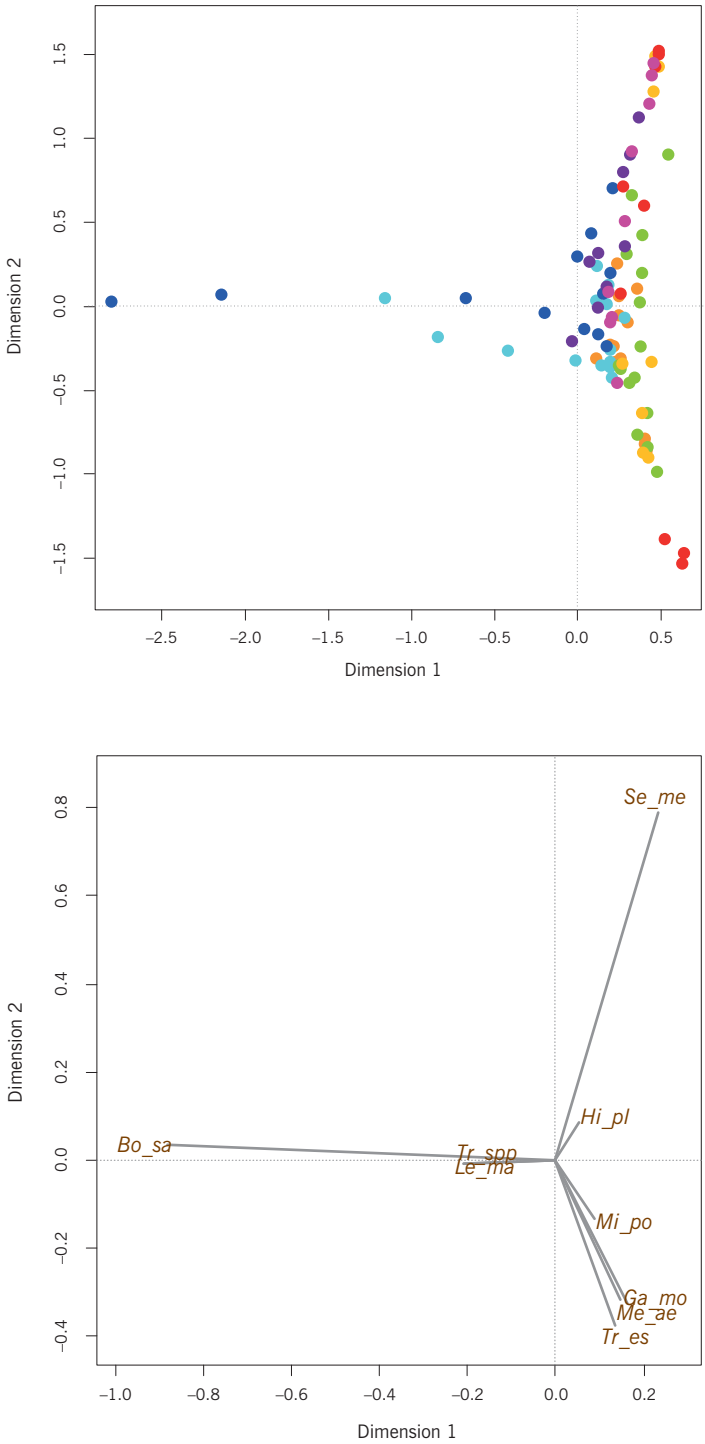


Exhibit 13.6:
Contribution biplot of the "Barents fish" data, retaining only the nine species with high contributions to the three-dimensional solution. The sample and species points are shown separately. The Procrustes correlations with the configurations obtained in Exhibits 13.2 (sample points) and 13.5 (species points), using all 30 species, are 0.993 and 0.997 respectively

shows the contribution biplot of this reduced data set of nine species. The result is almost identical – the Procrustes correlations with the previous results are almost 1.

Symmetric analysis of rows and columns

In all the above we have considered the case of the row profiles, the relative abundances of the species in each sample, with chi-square distances between them, mapped into a space using (weighted) classical MDS, with columns (i.e., species) displayed either as unit profiles or in contribution coordinates. We could turn this problem around by interchanging rows and columns and repeating everything as before. The matrix of column profiles is thus considered – these are the relative abundances across the samples of each species (i.e., the columns of Exhibit 11.1 divided by the column totals). Chi-square distances between these species profiles would be visualized (i.e., columns in principal coordinates), and the sample points added either as unit points (i.e., rows in standard coordinates) or as standardized regression coefficients (i.e., rows in contribution coordinates). In CA the row and column profile matrices, analysed in this similar and symmetric way, lead to exactly the same final solution, and all the sets of coordinates are related by simple scalar multipliers. The following are the basic results to remember, for both row and column points:

$$\text{principal coordinates} = \text{standard coordinates} \times (\text{principal inertias})^{1/2} \qquad (13.1)$$

$$\text{contribution coordinates} = (\text{masses})^{1/2} \times \text{standard coordinates} \qquad (13.2)$$

For example, suppose we had all the results from the analysis of the sample profiles, as discussed up to now and as shown in Exhibits 13.2, 13.4 and 13.5, and we wanted the equivalent results for the analysis of the species profiles. The species principal coordinates would be the species standard coordinates (shown in Exhibit 13.2) multiplied by the square roots of the principal inertias (eigenvalues) on respective axes: $(0.777)^{1/2}$ on first axis, $(0.541)^{1/2}$ on the second, $(0.485)^{1/2}$ on the third – notice that the principal inertias in CA are always less than one, so the principal coordinates are always contracted towards the centre compared to the standard coordinates. To obtain the sample standard coordinates we have to do the reverse operation by taking the sample principal coordinates (also in Exhibit 13.2) and divide by the corresponding square roots of the principal inertias, given above. Finally, to obtain sample contribution coordinates in order to see which are the highly contributing samples to the solution, the standard coordinates for each sample are multiplied by the corresponding square root of the sample mass.

A popular way of showing the results of a CA is to show the simultaneous display of the row and column profiles, that is both in principal coordinates. For the

"Barents fish" data set, this so-called *symmetric map* of the points, where both rows and columns are visualizing their inter-point chi-square distances, is shown in Exhibit 13.7. Because contributions are not directly visualized in the points' coordinates, we can introduce larger and smaller symbols or labels to give an indication of the important points to concentrate on in the interpretation. An advantage of this display is that the row and column points have the same inertias (parts of variance) along the dimensions, so they are spread out the same amount horizontally and vertically, which uses the plotting space better. Strictly speaking, however, the symmetric map is not a biplot as described before. However, when the square roots of the principal inertias along axes are not too different, so that principal and standard coordinates are approximately proportional to one another in the two-dimensional solution (see formula (13.1)), then the map comes close to a true biplot. In this example, $(0.777)^{½} = 0.881$ and $(0.541)^{½} = 0.736$, which are indeed quite close, so Exhibit 13.7 can be interpreted as an approximate biplot.

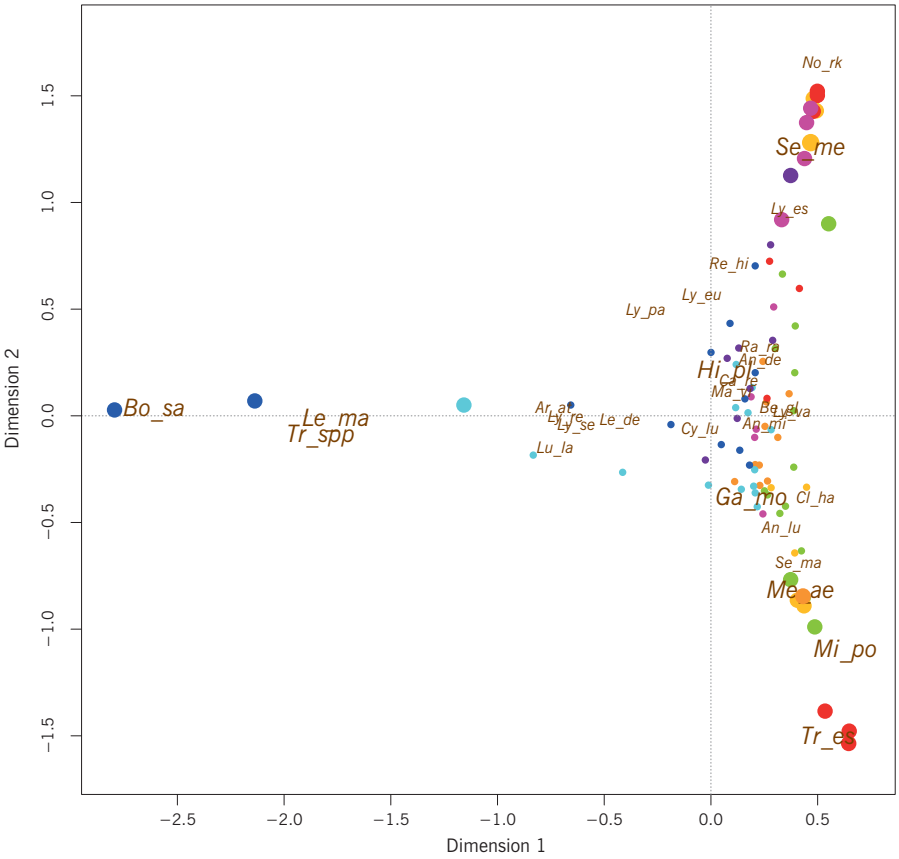


Exhibit 13.7:
Symmetric map of "Barents fish" data set, both samples and species in principal coordinates, with higher than average contributing samples and species in larger symbols and font sizes

SUMMARY: Correspondence analysis

1. Correspondence analysis (CA) is the analogue of principal component analysis (PCA) for data that are nonnegative such as abundance counts, biomasses and percentages. All the data must be measured on the same scale, so that it makes sense to compute row sums and column sums.
2. CA analyses the row profiles and/or the column profiles of the data matrix: these are the rows of data divided by their respective row sums or the columns divided by their respective column sums.
3. Each row and each column is weighted by its respective mass: the masses are the row and columns sums relative to the grand total of the data.
4. Distances between row profiles or between column profiles are defined by the chi-square distance.
5. For a samples-by-species data matrix, CA is generally thought of asymmetrically as an analysis of the sample (row) profiles, visualizing the inter-profile chi-square distances in a low-dimensional map (i.e., samples displayed in principal coordinates).
6. The species can then be visualized in two alternative ways as a biplot: as unit profiles, showing fictitious samples consisting of just one species (i.e., species in standard coordinates), or as gradient vectors showing the regression relationships between the species and the principal axes (i.e., species in contribution coordinates). These alternatives indicate identical orientations of biplot axes, but the latter alternative has the advantage that the more outlying species are the higher contributors to the solution.
7. Thinking of the analysis from the column profile point of view gives another way of interpreting the CA solution, as distances between species. The solution of this problem is identical to the row profile problem, with simple scaling factors linking the two solutions.
8. A popular way of showing the CA result is the symmetric map, where both row and column profiles are visualized simultaneously, that is both in principal coordinates. The row and column points have the same spread along the dimensions, and they can each be interpreted in terms of approximate chi-square distances.

CHAPTER 14

Compositional Data and Log-ratio Analysis

We have already met compositional data in the form of row or column profiles in CA: these are sets of nonnegative values that add up to a constant, usually 1 or 100%. In CA the profiles are computed on data matrices of abundances or biomasses, for example by dividing by their respective row and column totals. In other contexts the original data are compositional, for example chemical or geological data where the total size of the sample, measured in units of weight or volume, is not relevant, just its decomposition into a set of components. Another example of compositional data in biology is that of fatty acid compositions in studies of marine food webs. Compositional data are special because in their original form they have the property of closure, that is the compositional values of each sample have a constant sum. There are particular methodological issues when analysing compositional data, such as subcompositional coherence and the log-ratio transformation, which we shall consider in this chapter. Although this chapter is specific to compositional data, the wider issue of rare observations is discussed and the value of the contribution biplot is again demonstrated.

Contents

Compositional data and subcompositions

To illustrate the main reason why compositional data are a special case, consider the data in Exhibit 14.1. First, there is composition consisting of four fatty acids measured in six samples, with their components adding up to 1. Second, the last component is eliminated and the composition is *closed* again, that is re-expressed as proportions that sum to 1: this is called a s*ubcomposition* of the original com-

Exhibit 14.1: *Compositional data matrix (a) and a subcomposition (b), after eliminating the last component*

(a)

	16:1(n-7)	*20:5(n-3)*	*18:4(n-3)*	*18:00*	*Sum*
B6	0.343	0.217	0.054	0.387	*1*
B7	0.240	0.196	0.050	0.515	*1*
D4	0.642	0.294	0.039	0.025	*1*
D5	0.713	0.228	0.020	0.040	*1*
H5	0.177	0.351	0.423	0.050	*1*
H6	0.209	0.221	0.511	0.059	*1*

(b)

	16:1(n-7)	*20:5(n-3)*	*18:4(n-3)*	*Sum*
B6	0.559	0.353	0.088	*1*
B7	0.494	0.403	0.103	*1*
D4	0.658	0.302	0.040	*1*
D5	0.742	0.237	0.021	*1*
H5	0.186	0.369	0.445	*1*
H6	0.222	0.235	0.543	*1*

position. If researcher A works with the data in Exhibit 14.1(a) and researcher B with the data in Exhibit 14.1(b) and they consider it interesting to compute correlations as a way of measuring association between the components, they will obtain the results in Exhibit 14.2(a) and 14.2(b) respectively. While researcher A finds that the correlations between fatty acid *18:4(n-3)* and the pair *16:1(n-7)* and *20:5(n-3)* are –0.671 and 0.357 respectively, researcher B finds that they are –0.952 and –0.139. There is clearly a paradox here – the relationship between two

Exhibit 14.2: *Correlations between the columns of the compositional data matrices in Exhibit 14.1*

(a)

	16:1(n-7)	*20:5(n-3)*	*18:4(n-3)*	*18:00*
16:1(n-7)	1	–0.038	–0.671	–0.379
20:5(n-3)	–0.038	1	0.357	–0.604
18:4(n-3)	–0.671	0.357	1	–0.407
18:00	–0.379	–0.604	–0.407	1

(b)

	16:1(n-7)	*20:5(n-3)*	*18:4(n-3)*
16:1(n-7)	1	–0.171	–0.952
20:5(n-3)	–0.171	1	–0.139
18:4(n-3)	–0.952	–0.139	1

components should be the same and not depend on whether another component (*18:00*) is present or not. We say that the correlation does not have the property of *subcompositional coherence* – it is incoherent.

The log-ratio transformation

Values that are constant in a composition and any of its subcompositions are the ratios between components. For example, consider the four-part composition $[a,b,c,d]$ with $a+b+c+d=1$, and a three-part closed subcomposition $[a,b,c]/\ (a+b+c)$. Then the ratio a/b in the composition is identical to the ratio $[a/(a+b+c)]\ /\ [b/(a+b+c)]$ in the subcomposition. Since ratios are generally compared multiplicatively rather than additively, the logarithms of the ratios provide a justifiable transformation of the compositional data and have subcompositional coherence. Exhibit 14.3(a) shows the log-ratios $\log(a/b)$ for all six pairs of components a and b in Exhibit 14.1(a), as well as their means and standard deviations. In addition, a distance d_{ab} between the two components a and b is calculated as the square root of the average sum of squares of log-ratios across the samples:

$$d_{ab}=\sqrt{\sum_{i=1}^{n}(1/n)\,[\log(a_i/b_i)]^2}=\sqrt{\sum_{i=1}^{n}(1/n)\,[\log(a_i)-\log(b_i)]^2} \qquad (14.1)$$

(a)

	Log-ratios					
	16:1(n-7) / 20:5(n-3)	*16:1(n-7) / 18:4(n-3)*	*16:1(n-7) / 18:00*	*20:5(n-3) / 18:4(n-3)*	*20:5(n-3) / 18:00*	*18:4(n-3) / 18:00*
B6	0.458	1.849	–0.121	1.391	–0.579	–1.969
B7	0.203	1.569	–0.764	1.366	–0.966	–2.332
D4	0.781	2.801	3.246	2.020	2.465	0.445
D5	1.140	3.574	2.881	2.434	1.740	–0.693
H5	–0.685	–0.871	1.264	–0.187	1.949	2.135
H6	–0.056	–0.894	1.265	–0.838	1.321	2.159
mean	0.307	1.338	1.295	1.031	0.988	–0.043
sd	0.643	1.861	1.585	1.278	1.418	1.960
distance	0.662	2.162	1.942	1.557	1.629	1.790

(b)

	Distance(log-ratios)			
	16:1(n-7)	*20:5(n-3)*	*18:4(n-3)*	*18:00*
16:1(n-7)	0	0.662	2.162	1.942
20:5(n-3)	0.662	0	1.557	1.629
18:4(n-3)	2.162	1.557	0	1.790
18:00	1.942	1.629	1.790	0

Exhibit 14.3: *Logarithms of ratios between all pairs of components and the root mean sum of squares of the log-ratios as a measure of proximity*

Definition (14.1) shows that this distance function is simply a Euclidean distance between the log-transformed components. In Exhibit 14.3(b) the distances have been gathered into a square matrix, which can be used in a cluster analysis or an MDS. If fatty acid *18:00* is removed and the distance function is applied to Exhibit 14.1(b), the distances between the three components of the subcomposition remain identical, hence this measure of distance between the components is subcompositionally coherent.

The "fatty acid" data set

Exhibit 14.4 shows a part of the data set "fatty acid", compositional data on 25 fatty acids from 42 copepods of the species *Calanus glacialis*. The copepods were sampled in three different seasons and the objective is to see how the fatty acid compositions relate to these different seasons. Notice that the components with higher means also have higher standard deviations, which is typical of such data, as it is for count data. In the case of CA, the chi-square distance compensates for this disparity in variances. There is a similar issue in log-ratio analysis, which we describe now.

Log-ratio analysis

Log-ratio analysis (LRA) is the analogue of PCA that visualizes the compositional variables (also called components) transformed to log-ratios – hence it has the property of subcompositional coherence, which neither PCA nor CA have. It is a simple adaptation of PCA and has two forms: an unweighted form and a weighted form. We restrict our discussion to weighted LRA since the weighting has a number of benefits.

Notice in the last line of Exhibit 14.4 "mean(LR)^2", the mean of the squares of the log-ratios in each column. Just as we did for each row of the mini-example

Exhibit 14.4: *Part of 42 × 25 data matrix of fatty acid compositions, expressed as percentages: each set of 25 values in the rows sums to 100%. The mean and standard deviation of each column is given, as well as the mean of the squares of log-ratios for pairs of samples in each column*

	14:00	*i-15:0*	*15:00*	*16:00*	*16:1(n-7)*	⋯	*22:5(n-3)*	*22:6(n-3)*	*Total*
B5	14.229	1.223	0.870	12.204	6.567	⋯	0.543	0.446	*100*
B6	12.153	1.270	1.085	12.318	7.406	⋯	0.353	0.469	*100*
B7	6.640	0.790	0.529	12.272	6.804	⋯	0.656	0.231	*100*
B8	12.410	1.167	0.822	11.543	7.668	⋯	0.425	0.436	*100*
H5	6.764	0.338	0.272	8.056	6.207	⋯	0.298	0.464	*100*
H6	6.896	0.324	0.262	8.046	6.494	⋯	0.313	0.520	*100*
⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋮
E5	5.410	0.407	0.273	12.321	6.622	⋯	0.273	0.257	*100*
E6	9.200	0.813	0.606	9.741	19.193	⋯	0.542	0.601	*100*
mean	8.366	0.678	0.546	9.196	12.818	⋯	0.640	9.100	*100*
sd	2.131	0.277	0.181	1.816	8.263	⋯	0.251	2.715	
mean(LR)^2	0.114	0.321	0.252	0.080	0.664		0.811	0.142	

in Exhibit 14.3, so we can compute log-ratios between all the 42 values in each column (there will be ½ × 42 × 41 = 861 ratios in total), which would be the basis for a distance calculation between pairs of samples. The mean square of these log-ratios has the property that it will be higher for rarer components, which can have bigger ratios than those between components at a higher level. For example, a rare component with mean 0.03% could easily have two values of 0.05% and 0.01%, which gives a ratio of 5, whereas such a large ratio would hardly ever occur for a component with values of the order of 10%, varying between 6% and 14%, say. In weighted log-ratio analysis this effect is compensated for by assigning weights to each component proportional to its mean, so that rarer components get smaller weights. This is exactly the same idea as in CA.

Technically, weighted LRA can also be thought of as a two-step procedure, performing MDS on inter-sample distances based on the log-ratios, where the components have been weighted as just described, and then adding the component variables by regression on the MDS dimensions. But more simply, it reduces to a PCA of the log-transformed data matrix which is centred row-wise, that is each row of the logged data is centred to have mean zero. This centred matrix is then subject to PCA, incorporating the column weights. Because PCA will then automatically centre the data column-wise, it follows that the log-transformed compositional data matrix is actually double-centered, row-wise and column-wise. Notice that the actual log-ratios for all pairs of components do not have to be calculated, thanks to the double-centering. LRA is thus a weighted PCA of the previously log-transformed and row-centered data, with some special features of the interpretation.

Interpretation of log-ratio analysis

Exhibit 14.5 shows the weighted LRA of the "fatty acid" data set, with samples in principal coordinates (thus approximating the log-ratio distances between them) and fatty acids in standard coordinates. Separately, we have verified that only the first two dimensions are significant. There are three clearly separated groups of samples, which we have labelled A, B and C, coinciding exactly with the three seasons in which they were sampled. As in Exhibit 13.5 we have separated the higher than average contributors to the first two dimensions from the others: these seven fatty acids are thus indicated with larger labels, and account for 90% of the variance in this biplot. A novelty of the log-ratio biplot is that it is not the vectors from the origin that define the biplot axes, but the vectors linking the component variables – these vectors are called *links*. For example, the link from *16:1(n-7)* at bottom left to *18:00* top left represents the log-ratio log(*18:00*/*16:1(n-7)*), and the direction of this link is exactly lining up with group A at the top and group C at the bottom. Similarly, the link from *18:00* to *18:4(n-3)*, as well as several others made by the group of three high-contributing fatty acids in-between, separates group A from group B. And the link from *16:1(n-7)* and *18:4(n-3)* is one that

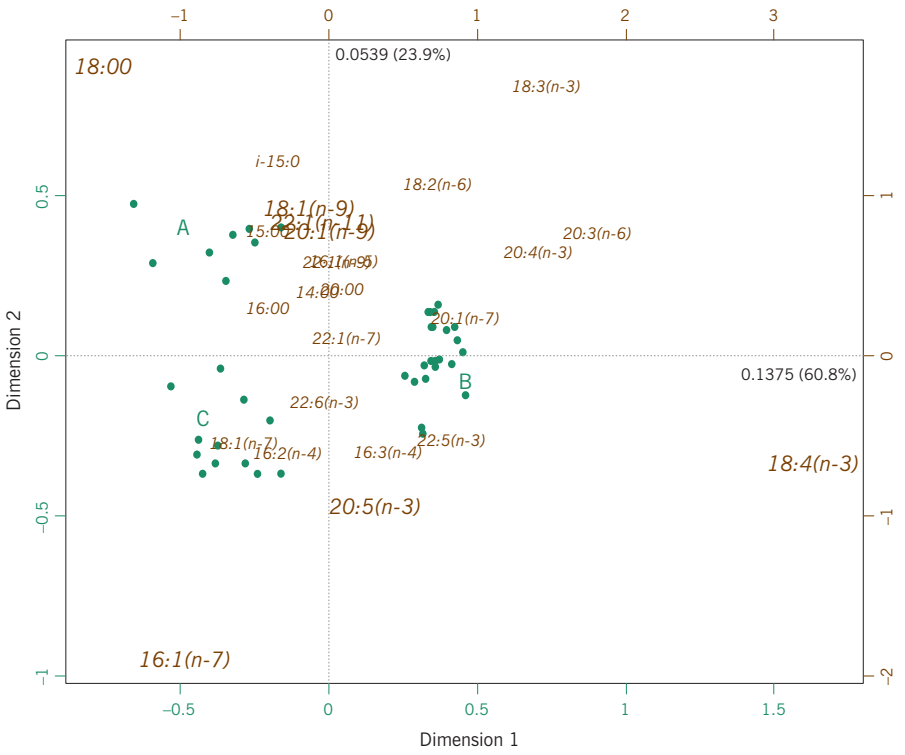


Exhibit 14.5: *Row-principal LRA biplot of "fatty acid" data set. 84.7% of the log-ratio variance is explained. The seven higher-than-average contributing fatty acids are shown in larger font. Notice the different scales for sample points and fatty acid points*

separates group C from group B. Exhibit 14.6 illustrates the group separation in a simpler scatterplot of two of these log-ratios that are suggested by these results. Since the third log-ratio log[*18:4(n-3)/18:00*] separating groups A and B is the horizontal axis of the scatterplot minus the vertical one, it can be depicted by a 45 degrees descending line, shown by the dashed arrow, perfectly coinciding with the separation of the A and B samples.

An interesting feature of the log-ratio biplot is that if components fall on straight lines (as, for example, *18:00*, *18:4(n-3)* and the group of three fatty acids inbetween, *18:1(n-9)*, *22:1(n-11)* and *20:1(n-9)* in Exhibit 14.5) then a model can be deduced between them. The Bibliographical Appendix gives a reference to this way of diagnosing models in biplots.

Relationship between CA and LRA

CA also analyses compositions, albeit compositions (i.e., profiles) computed on a matrix of counts or abundances. In fact, CA can be used to analyse purely compositional data, and likewise, LRA can be used to analyse count data or other strictly positive ratio-scale data. There is an interesting relationship between the two methods: leaving out some technical details, the main result is that if one ap-

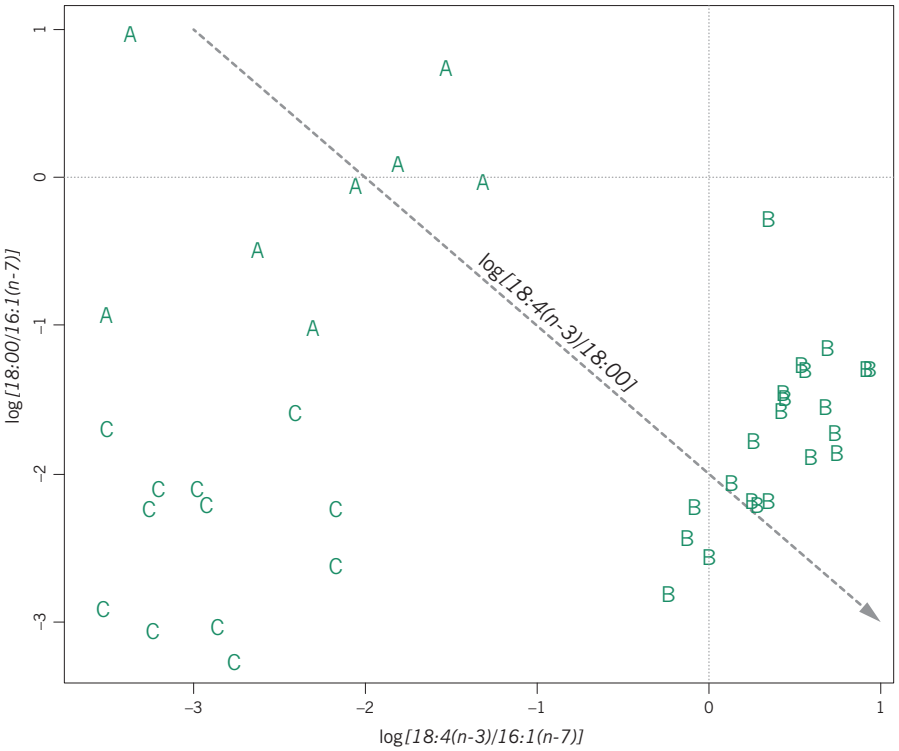
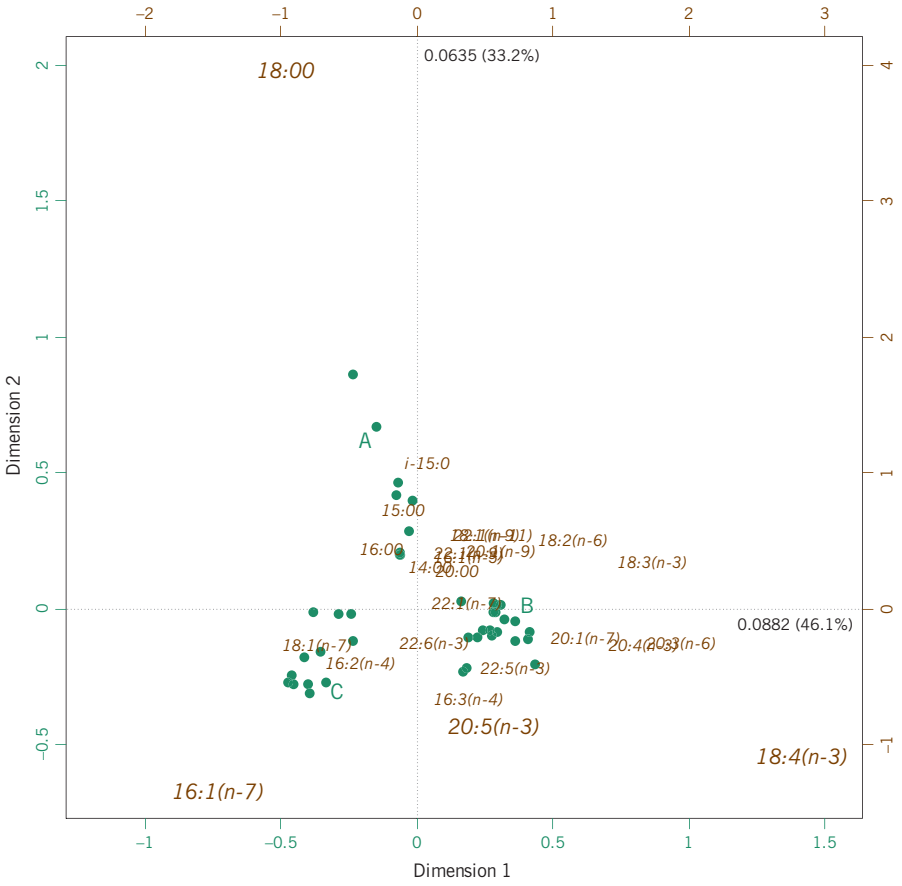


Exhibit 14.6:
Scatterplot of two log-ratios suggested by the biplot in Exhibit 14.5, perfectly separating the three groups of copepods. A third log-ratio combining the two describes a diagonal axis in the plot

plies the Box-Cox power transformation to the data (see Chapter 3 and definition (3.4)), with increasingly stronger power (for example, square root, then cube root, then fourth root, etc.), then the CA of the transformed data tends to LRA in the limit. Moreover, if the variance in the data is small, then the CA solution will be close to the LRA solution anyway. This means that CA is close to being subcompositionally coherent, and perhaps close enough for practical purposes. The CA biplot comparable to Exhibit 14.5 is given in Exhibit 14.7, and is indeed very similar. Here are some statistics comparing the two results:

	(Weighted) LRA	CA
Total variance (or inertia)	0.2260	0.1913
Variance, dimension 1	0.1375 (60.8%)	0.0882 (46.1%)
Variance, dimension 2	0.0539 (23.9%)	0.0635 (33.2%)
Percentage in two dimensions	84.7%	79.3%
Procrustes correlation between rows	0.950	
Procrustes correlation between columns	0.930	

Exhibit 14.7: *Row-principal CA biplot (asymmetric map) of "fatty acid" data. Explained variance is 79.3%*



Four out of the seven fatty acids previously highlighted in the LRA are singled out as high contributors in the CA. The three groups of copepods are separated in the same way, but the interpretation of the joint plot is different. Here, as for most biplots, the biplot axes are considered through the origin to each variable point. For example, if we draw a straight line from the bottom through the origin and up to fatty acid *18:00*, then the projections of the copepods on this axis should reproduce approximately the compositional values on this fatty acid. Exhibit 14.8 verifies this and also shows how close these projections are to the actual values. In fact, the original values show some overlap between the A group and the others, whereas the estimated values perfectly separate the A group. This is due to the fact that other fatty acids are operating in the biplot to separate the groups – so group A is separated not only because it is high on *18:00* but also low on *16:1(n-7)*, which brings us right back to the idea in LRA to work with ratios.

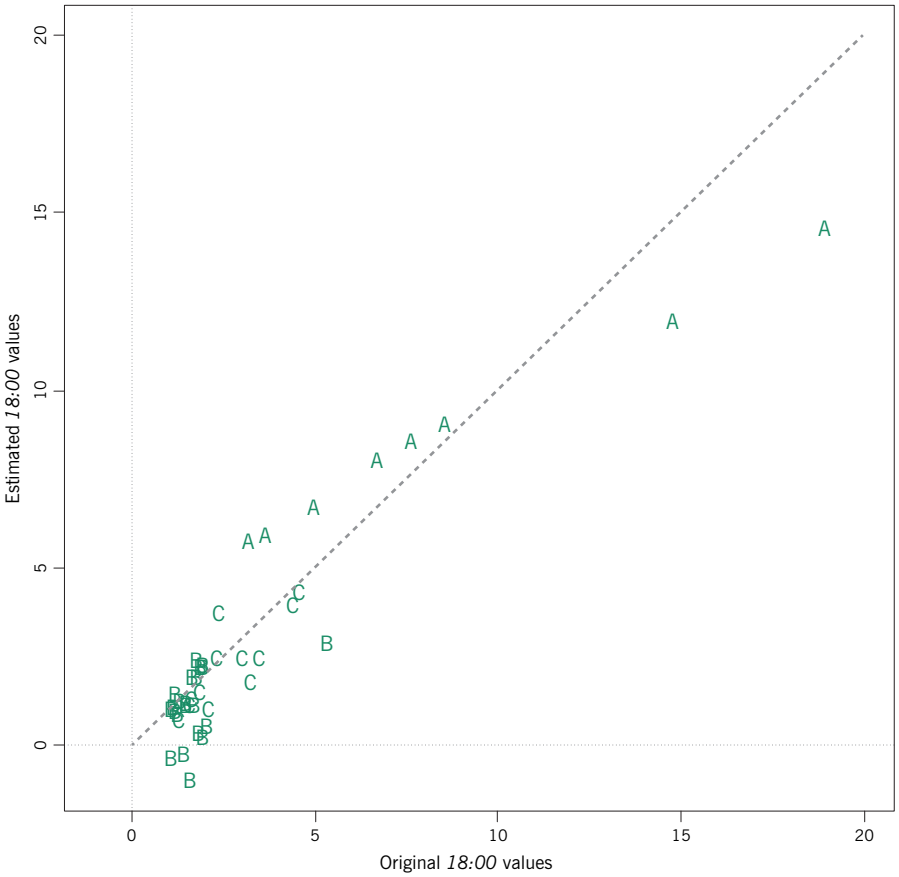


Exhibit 14.8: *Actual compositional value (as a percentage) of fatty acid 18:00 and estimated values from the CA biplot of Exhibit 14.7. The dashed line represents perfect reconstruction. The correlation is 0.928, thus the variance explained in 18:00 by the two dimensions is $0.928^2 = 0.861$, i.e. 86.1%*

Zeros in compositional data

Since LRA visualizes ratios, there should be no zero values in the data, as has been the case for the fatty acid data set used in this chapter so far. In fact, this data set, with 25 fatty acids, is a subset of a bigger one that does have an additional 15 fatty acids with some observed zeros. Collectively these 15 additional fatty acids account for between 3 and 4 percent of each sample, so they are rare fatty acids and thus sometimes observed as zeros. Let us call this data set with all 40 fatty acids the "complete fatty acid" data set, and consider how to analyse it. Zeros can arise for various reasons, one being that the presence of the fatty acid is below the detection limit of the measuring instrument. If one knows what this detection limit is, a value of half the detection limit, say, could be substituted for the zeros. This will create large log-ratios, and thus large variances, but because fatty acids are weighted in the analysis proportionally to their mean values, this will reduce the effect of these large variances in the rare fatty acids. Another option is to treat the zeros as missing values – there are ways for handling missing data by estimating values in the data table

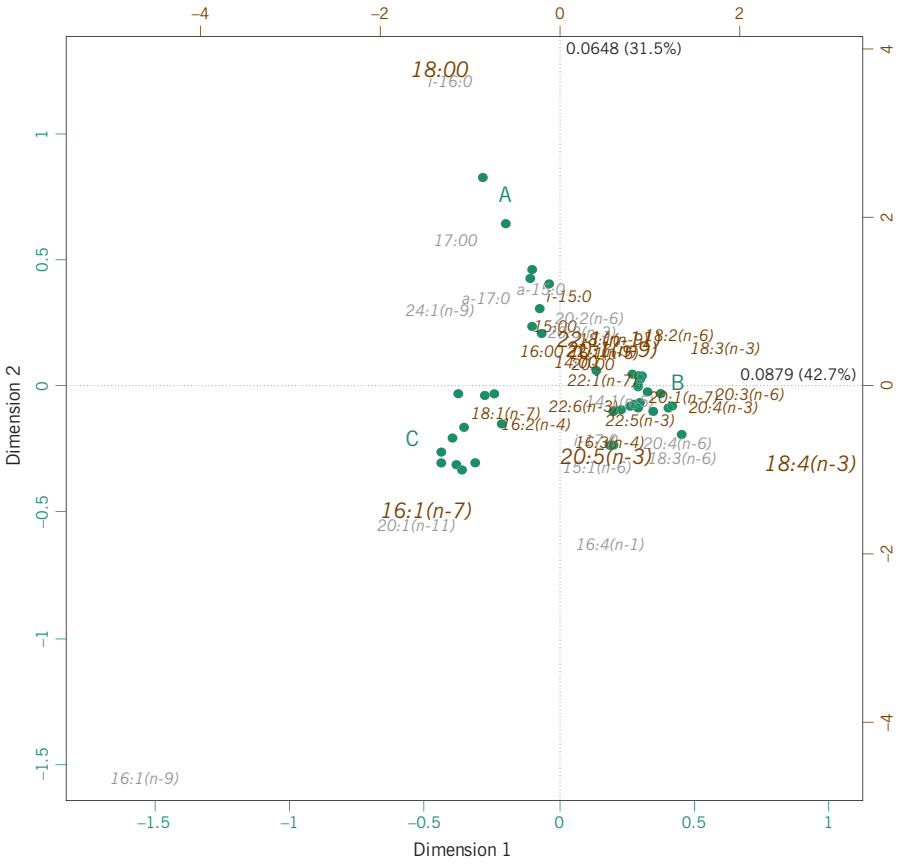


Exhibit 14.9: *CA of the "complete fatty acid" data set of 42 copepods and 40 fatty acids. The row-principal biplot is shown and the explained variance in this two-dimensional solution is 74.2%. Compared to Exhibit 14.7, the additional 15 fatty acids are coloured in gray*

from the biplot. An easier solution is to recognize that CA is a good approximation to LRA and close to having subcompositional coherence, and also has no problem with zeros in the data. Exhibit 14.9 is the CA of the complete data set of 40 fatty acids, and it is clear that the extra data have not changed the results that much. The samples are in an almost identical configuration, whereas the additional fatty acids are all low contributors. This biplot illustrates what often happens with low frequency variables, such as rare species or in this case fatty acids with low proportions. Some of these are in outlying positions in the biplot, for example *i-16:0* at the top and *16:1(n-9)* at bottom left. If one does not take into account the contributions, then one might think that *16:1(n-9)*, for example, is the most important fatty acid separating out group C, whereas it is in fact *16:1(n-7)*. This can be easily corroborated by making a scatterplot of these two fatty acids, shown in Exhibit 14.10.

A better way of showing the CA results in this case is in the form of a contribution biplot (Exhibit 14.11), where the low contributing variables shrink to the centre

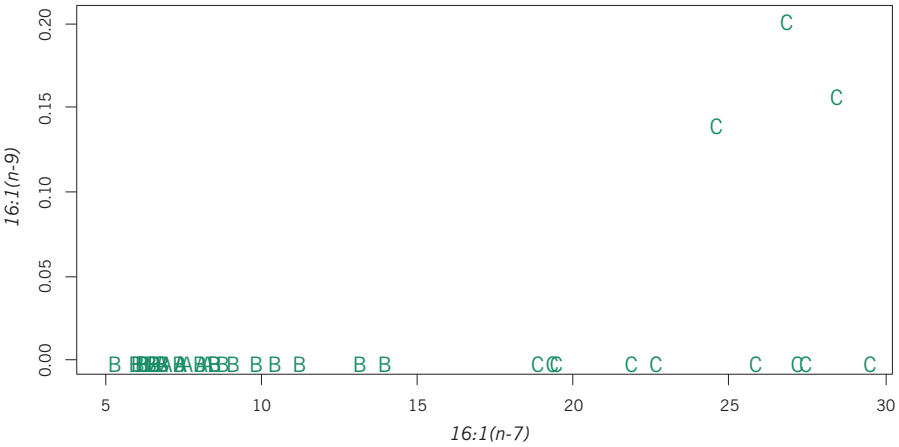


Exhibit 14.10: *Scatterplot of fatty acids 16:1(n-7) and 16:1(n-9) of the "complete fatty acids" data set, showing that 16:1(n-7) is the more important one for separating out group C of copepods. The rare fatty acid 16:1(n-9) has only three small positive percentages, coinciding with three copepods in group C*

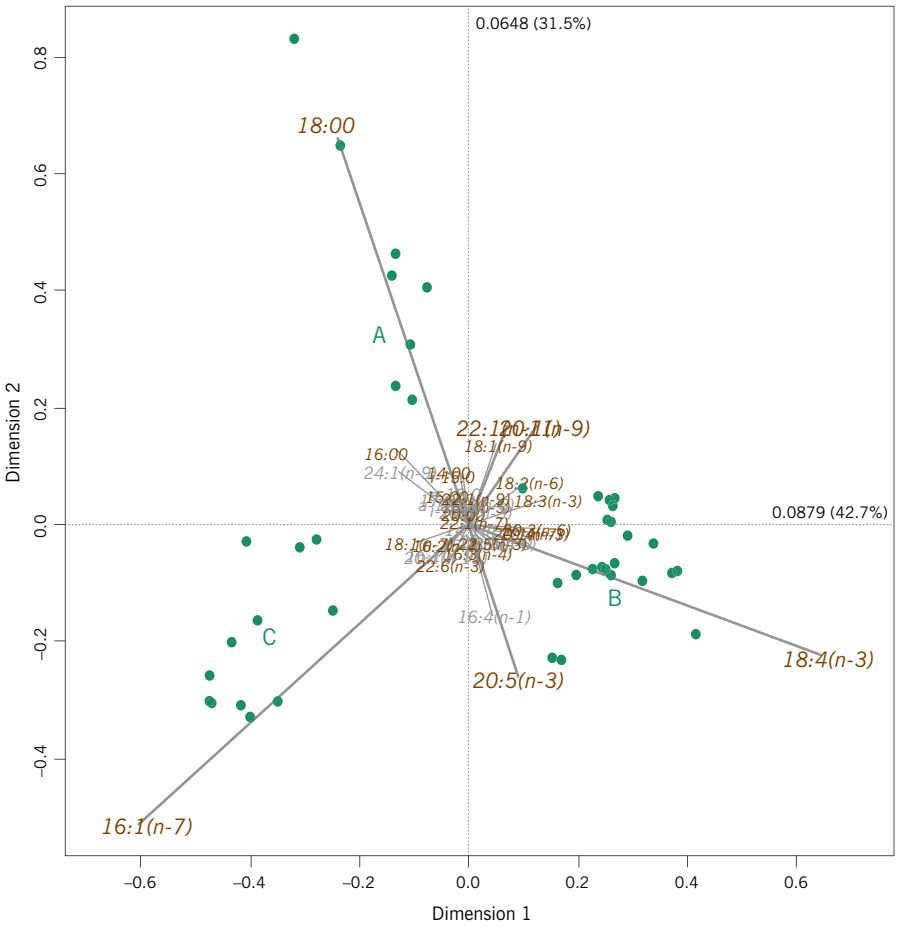


Exhibit 14.11: *CA contribution biplot of "complete fatty acid" data set. The six high contributing fatty acids stand out from the rest*

and the high contributors stand out according to their contribution. Notice too that only one scale is necessary in the contribution biplot (cf. Exhibits 14.5, 14.7 and 14.9 where there were separate scales for row and column points).

SUMMARY: Compositional data and log-ratio analysis

1. Compositional data have the property that for each sample its set of values, called *components*, sum to a constant, usually 1 (for proportions) or 100 (for percentages).
2. Because of this constant sum property, called the property of *closure*, many conventional statistics calculated on the components, such as the correlation coefficient, are inappropriate because they change when subcompositions are formed from a subset of components. Measures that do not change are said to have the property of subcompositional coherence.
3. The log-ratio transformation implies analysing all the pairwise ratios between components on a logarithmic scale. Ratios do not change in subcompositions and are thus subcompositionally coherent.
4. Log-ratio analysis (LRA) is a dimension reduction technique like PCA and CA that visualizes all the pairwise log-ratios in a biplot along with the sample points. Links between pairs of components in the biplot give directions of the log-ratio biplot axes, onto which samples can be projected to estimate the corresponding log-ratios.
5. CA turns out to have a strong theoretical link to LRA and, although not subcompositionally coherent, is close to being so. It provides a good alternative to LRA, especially when there are zero values in the data and the log-ratio approach can not be applied unless the zeros are substituted with positive values.
6. The contribution biplot is a valuable way to separate out components in the log-ratio analysis that are important for the interpretation.

CHAPTER 15

Canonical Correspondence Analysis

PCA, CA and LRA operate on a single data matrix, and have similar ways of reducing the high dimensionality of the data to a low-dimensional approximation for ease of interpretation. The low-dimensional views of the data are the best in terms of the least-squares criterion in each case, accounting for a maximum amount of variance while simultaneously minimizing the unexplained variance. Often additional data are available, which can be related afterwards to an existing ordination. One of the most common situations in ecology is when the data consist of biological measurements (e.g., species abundances) at different locations, and in addition there are various environmental variables observed at the locations. We have shown how biological data can be optimally displayed with respect to ordination axes and then how the environmental variables can be related to these dimensions. The reverse can also be done, first optimally displaying the environmental data and then fitting the biological data to the results. In either case these relationships might be weak. Ecologists may be more interested in that part of the biological information that is more directly related to the environmental information. This environmentally related part of the biological variance is also multidimensional, so we again resort to ordination to interpret it through dimension reduction. Methods that relate two sets of data are often described as *canonical* in statistics, and this chapter deals mainly with one of the most popular in ecology, canonical correspondence analysis.

Contents

Response and explanatory variables

In Chapter 10 we looked at the introductory data set "bioenv" in detail and made regression biplots using two of the environmental variables, pollution and depth, as the support of the biplot, or two so-called *canonical dimensions* that were obtained by maximizing the correlation between the biological and environmental data sets. Since then we have learned a bit more about analysing abundance data using CA, so in this chapter we will introduce a variant of CA, called *canonical correspondence analysis* (CCA), which is appropriate for this particular combination of biological and environmental measurements on the same samples. Recalling Exhibit 1.1, there were 30 samples and the five biological variables, regarded as response variables, accompanied by four environmental variables, of which three are on continuous scales and one on a categorical scale. The objective is to find out how much of the variance (in the CA sense, in other words, inertia) is accounted for by the environmental variables and to interpret the relationship. In this approach the two sets of variables are considered asymmetrically: the biological data are the responses (like the *"Y"* variables in a regression) and the environmental variables are the explanatory variables, or predictors (like the *"X"* variables). This is different from the canonical correlation analysis of Chapter 10, which treated the two sets of data symmetrically and would have been the same if the two sets of variables were interchanged.

Indirect gradient analysis

Before explaining CCA, let us first consider the CA of the 30×5 matrix of biological data and, as before, the ways of displaying the environmental variables on the CA biplot. Exhibit 15.1 shows the row-principal contribution biplot of the samples and species, which means that the distances between the samples are approximate chi-square distances between their profiles, and the standardized species (standardized in the CA sense – see Chapter 13) have been regressed by weighted least-squares on the two CA dimensions and are depicted by their regression coefficients. The total inertia of the biological data is equal to 0.544 and the axes account for 0.288 (53.0%) and 0.121 (22.2%) respectively, that is 75.2% of the total. Both 75.2% of the variance of the sample points and 75.2% of the variance of the species is explained by this solution.

While the variance explained for the species abundance data is the best possible according to the optimization criterion in CA, the regressions of the environmental variables are much lower and can vary a lot in terms of variance explained:

Depth	30.4%
Pollution	69.5%
Temperature	2.1%
C (clay)	3.4%
S (sand)	9.9%
G (gravel)	18.7%

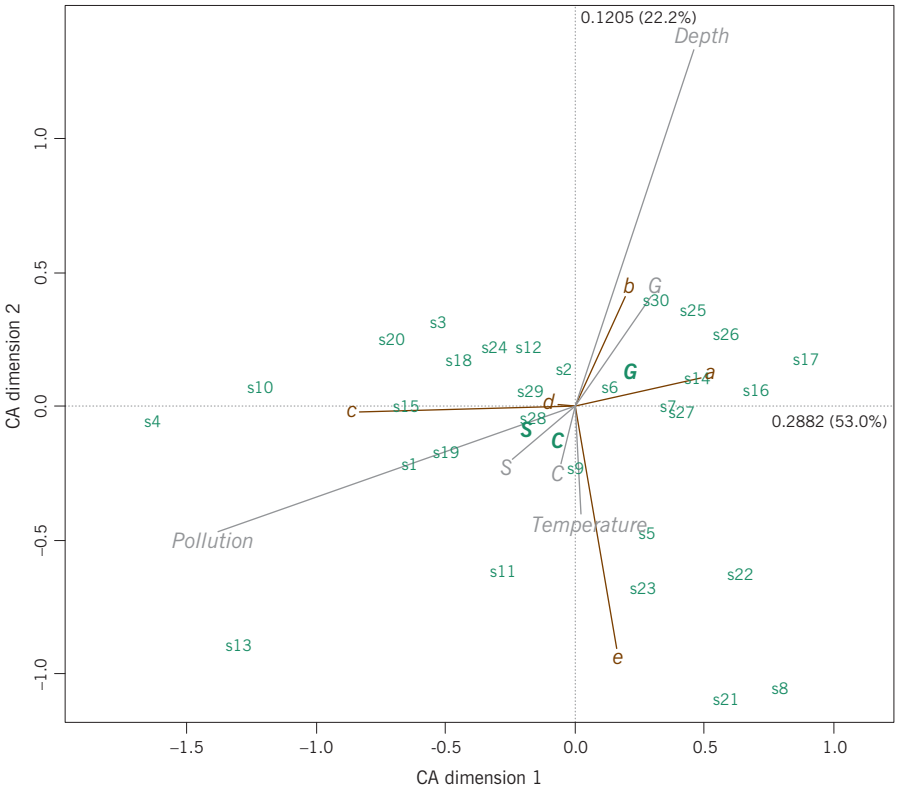


Exhibit 15.1: *CA biplot of the biological data in the "bioenv" data set, with samples in principal coordinates and species in contribution coordinates. The one discrete and three continuous environmental variables are shown according to their regression coefficients and the discrete variable's categories are additionally shown (in black) at the centroids of the samples in the corresponding categories*

Notice that dummy variables such as the sediment categories *C*, *S* and *G* (the gray points in Exhibit 15.1), with values of 0 and 1, will always have low variance explained. The other way of showing the categories is as centroids of the samples (the black points in Exhibit 15.1) – this can be achieved by adding three extra rows to the data matrix where the abundances of the species are aggregated across the samples for each sediment type, and declaring these additional rows as *supplementary points*. These additional rows are as follows:

	a	*b*	*c*	*d*	*e*
C	105	46	73	81	27
S	103	70	115	104	32
G	196	146	64	142	30

The (row) profiles of the sediment categories are exactly the centroids shown in Exhibit 15.1. These centroids do not lie on the same vector as the dummy vari-

ables, but there is a close mathematical relationship between these alternative sets of coordinates for category points added to the display, which depends on the mass of each category and the parts of inertia on each axis. In any case, what we are assured of is that the corresponding categories always lie in the same quadrant (one of the four regions defined by the two ordination axes), and if the parts of inertia are similar, then the centroids will lie very close to the dummy variable biplot axis.

This type of analysis in Exhibit 15.1 is called *indirect gradient analysis*: first an ordination is obtained optimally displaying the samples and response variables (here, the species), and then the explanatory variables are related to the ordination axes.

Direct gradient analysis

In indirect gradient analysis the relationship between the explanatory variables and the response variables is conditioned on the ordination of the response variables. One could imagine a situation in which the main dimensions of the responses have little relationship with the explanatory variables, because this relationship is to be found on less important dimensions of the response data. So, in order to focus specifically on the relationship between biological and environmental variables in this example, we first make a projection of the biological variables into the space of the environmental variables. This is also called *constrained* or *restricted* ordination, because a condition is introduced that the ordination axes must be linear functions of the environmental variables.

There are three continuous variables and three dummy variables, but – as in regression analysis – the dummy variables count for one less because of their interdependency, so there are five dimensions in the explanatory variable space. The first step then is to project the species response data into this space, which also means we eliminate all variance in the response data that is not correlated linearly with the explanatory variables – we are only interested in that part of the variance that is correlated with the environmental variables. The total inertia of the species data was, as we reported earlier, 0.544, and it turns out that the amount 0.249 of this inertia is linearly related to the environmental variables, i.e. 45.8% of the total. So from now on we are only interested in this constrained part of the inertia.

The analysis then continues as a regular CA in this restricted five-dimensional space, to find the axes that explain a maximum of this constrained inertia – Exhibit 15.2 shows the result of what is now a *canonical correspondence analysis* (CCA), in the form of a *triplot* of samples, species and environmental variables. Almost all (96.3%) of this constrained inertia of 0.249 is explained in the new ordination map. Pollution is the most important variable on the first axis,

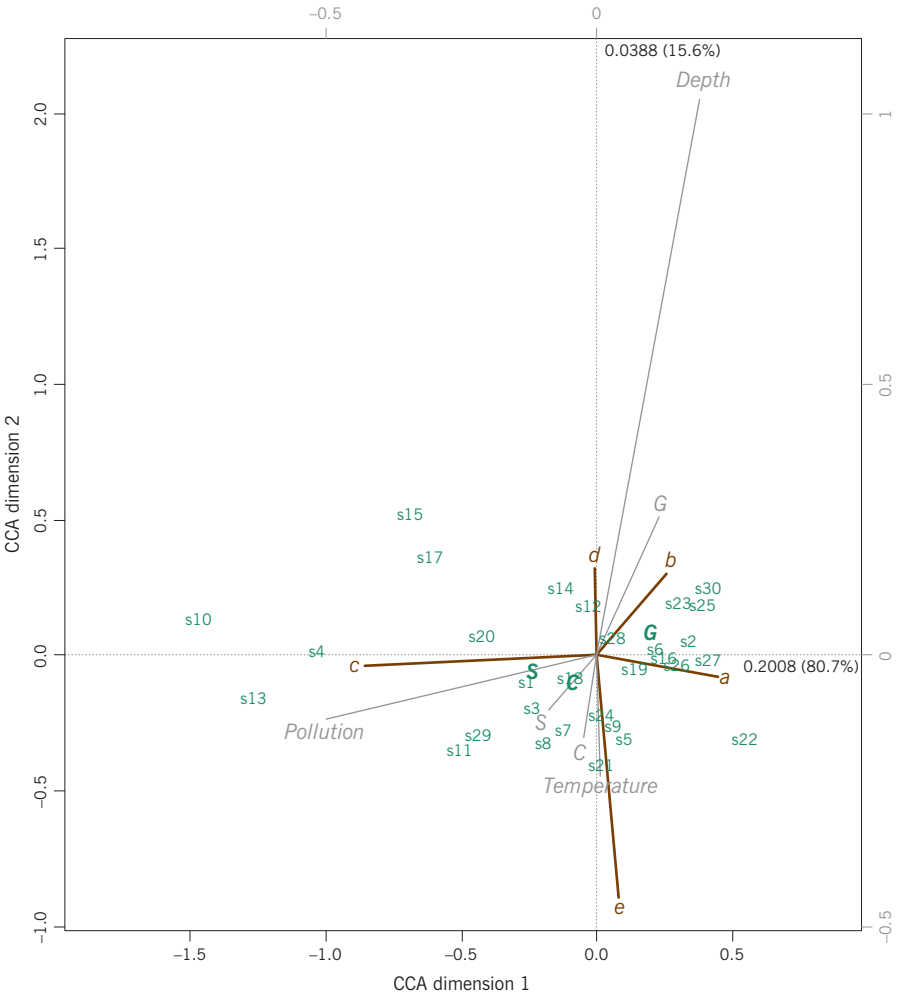


Exhibit 15.2: *Canonical correspondence analysis triplot of "bioenv" data. The row-principal scaling with species in contribution coordinates is again shown, as well as the environmental variables regressed onto the ordination axes. Percentages of inertia explained are with respect to the restricted inertia*

whereas depth is the most important on the second. Because the regressions of the environmental variables are performed on the sample principal coordinates that have much less variance on the second axis, depth's regression coefficient on the second axis is large and the variable gives the impression that it is more important than pollution. If we wanted comparability between the coordinates of the gradient vectors of the environmental variables, the biplot should be made using the standard coordinates of the samples, as in the regression biplots of Chapter 10.

The variances explained by the CCA axes of the environmental variables are now much higher than before, due to the constraining of the solution:

Depth	85.3%
Pollution	99.1%
Temperature	3.3%
C (clay)	8.8%
S (sand)	14.8%
G (gravel)	33.5%

Notice again that the dummy variables, by their very nature of having only values of 0 and 1, cannot attain a high percentage of variance explained – this issue is dealt with in the next chapter.

In the CCA described above, we have analysed the inertia of the abundance data in the space constrained by the environmental variables. The constrained space is formed by axes that are predictions based on linear regression on the environmental variables. In other applications researchers might be more interested in the inertia not correlated with a particular environmental variable or variables – for example, it may be of interest to partial out a known effect such as a latitudinal gradient. Exploring this unconstrained part is called *partial CCA*, which will be illustrated in the case study of Chapter 19.

Restricted ordination in PCA and LRA

The constrained version of CA illustrated above is similarly applicable to PCA and LRA. When the responses are continuous variables on an interval scale, then the version of PCA restricted in terms of a separate set of explanatory variables is called *redundancy analysis*. Similarly, when the responses are compositional data and LRA is applicable, it is possible to restrict the solution to be linearly related to predictor variables. The idea is the same in each case: project the data, with its particular distance function, into the space of the explanatory variables, and then carry on as before. We continue with CCA, which is the most popular of these options.

CCA as the analysis of weighted averages

There is another way of thinking about CCA, in terms of the weighted averages of the explanatory variables, using the relative abundances of the species as weights. Exhibit 15.3 shows this variables-by-species table, computed as follows. Take species *a* and variable *Depth* as an example. The relative frequencies of species *a* (i.e., the column profile) are 0, 26/404 = 0.0644, 0, 0, 13/404 = 0.0322, 31/404 = 0.0767, and so on (see Exhibit 1.1). These are used to compute a weighted average of the depth values at each site:

$$0 \times 72 + 0.0644 \times 75 + 0 \times 59 + 0 \times 64 + 0.0322 \times 61 + 0.0767 \times 94 + \cdots = 78.77$$

	a	*b*	*c*	*d*	*e*
Depth	78.77	81.16	72.58	79.27	70.17
Pollution	3.11	3.24	5.49	3.78	3.64
Temperature	3.03	3.06	3.04	3.06	3.11
C	0.26	0.18	0.29	0.25	0.30
S	0.26	0.27	0.46	0.32	0.36
G	0.49	0.56	0.25	0.43	0.34

Exhibit 15.3: *Weighted averages of the environmental variables, using the relative abundances of each species across the samples as weights*

So if the species tends to occur with higher abundance in deeper samples, then the weighted average will be high. Species c, for example, must be occurring in higher abundances in samples with high pollution, and species e in samples of lower depths, which can be verified in Exhibit 15.2. In the case of the dummy variables for sediment, the three values for each species sum to 1 and code the proportion of total abundance of that species in each category.

To obtain the equivalent result as a CCA by analysing the matrix in Exhibit 15.3 needs some technical explanation, since it involves the covariance matrix of the explanatory variables, but the point is that, once the appropriate transformations are made, the inertia in this table is identical to the restricted inertia of 0.249. Knowing this equivalence gives an extra way of thinking about the connection between the species abundances and environmental variables in the triplot.

Coding of explanatory variables

CCA (or the equivalent constrained methods in PCA and LRA) adds the condition that the ordination axes should be linearly related to the explanatory variables. Linearity of the relationship might not be realistic in most circumstances, so just like in regression analysis we can contemplate introducing transformations of the variables, for example, logarithmic transformation, or including polynomial terms, or fuzzy coding. In Chapter 11 we discussed an indirect gradient analysis of the "Barents fish" data set, coding the environmental variables either linearly or fuzzily, and also the geographical position of the samples either in a crisp way in terms of their regional location, or in a fuzzy way based on fuzzy latitude and longitude coordinates. In Chapter 13 various CAs of this same data set were considered. So now we can see how CCA performs on these data, and we will contrast the different ways of coding the environmental variables. Exhibit 15.4 shows the CCA triplot based on linear constraints on the two environmental variables and the 10 dummy variables for the spatial position. Of the total inertia of 2.781 in the abundance data, the environmental variables account for 1.618, that is 58.2%. Of this latter amount, 61.9% is displayed in Exhibit 15.4.

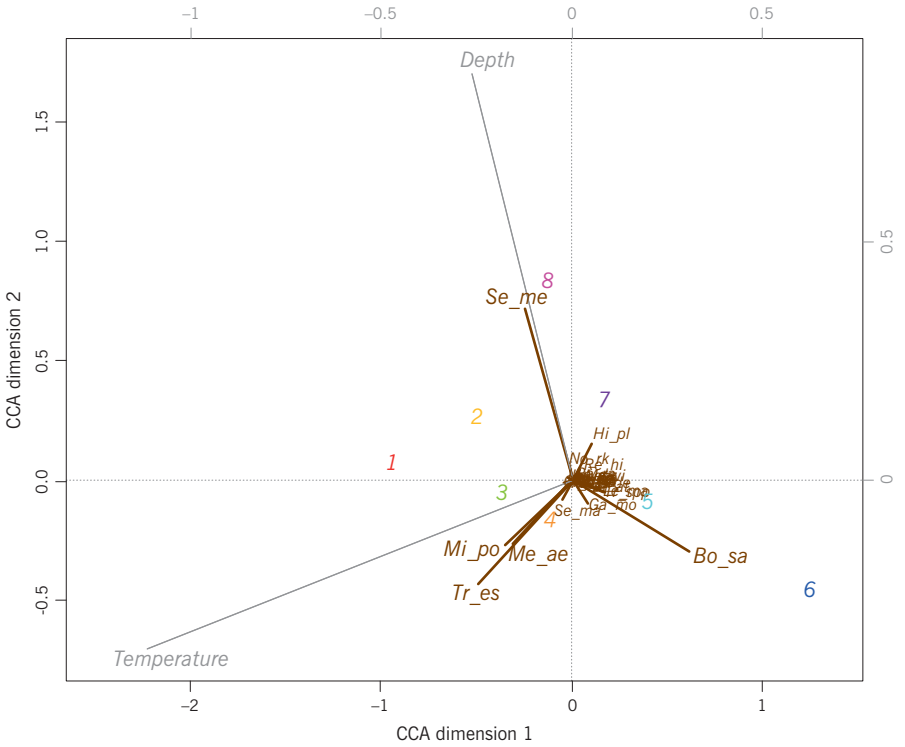


Exhibit 15.4: *CCA of "Barents fish" data, showing highly contributing species in larger font and the two continuous environmental variables according to their regressions on the axes. The 89 sampling sites are not shown, but their averages in the eight regional groupings are*

As a contrast, we try the CCA using depth and temperature coded into four fuzzy categories each, and the spatial position into nine fuzzy categories as explained in Chapter 11, leading to Exhibit 15.5. This time slightly more of the biological-environmental relationship is captured in the restricted space (1.700, i.e. 61.1%, compared to 58.2% previously), which is understandable since nonlinearities in depth and temperature can be picked up thanks to the fuzzy coding. In fact, temperature does appear to have a nonlinear pattern in the triplot, with low and high temperatures being associated with lower values of depth. 60.4% of this constrained ordination is explained in the triplot.

A distinct advantage of Exhibit 15.5 over Exhibit 15.4 is that all the environmental variables, including spatial ones, are displayed in the same way, in this case as weighted averages of the sample positions. To explain this weighted averaging more specifically, Exhibit 15.6 shows the positions of the 89 sample sites corresponding to the CCA result of Exhibit 15.5. Every station has associated with it the values of each fuzzy category, that is 17 values between 0 and 1 inclusive for the 4 fuzzy values of depth, 4 fuzzy values for temperature and 9 fuzzy values for spatial position. For example, for the fuzzy category *d4* the 89 samples have

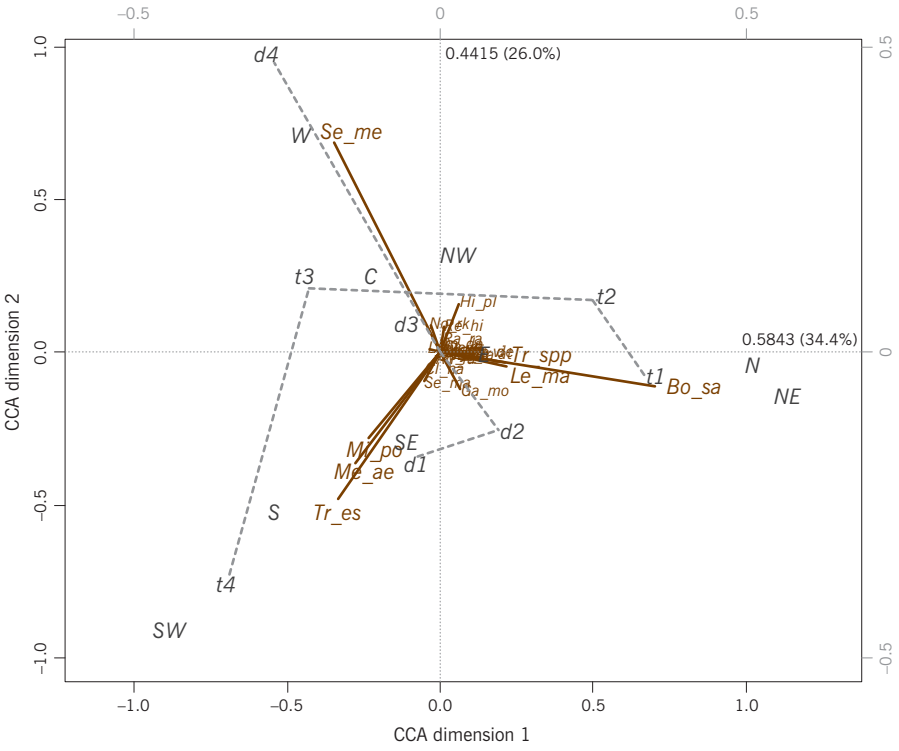


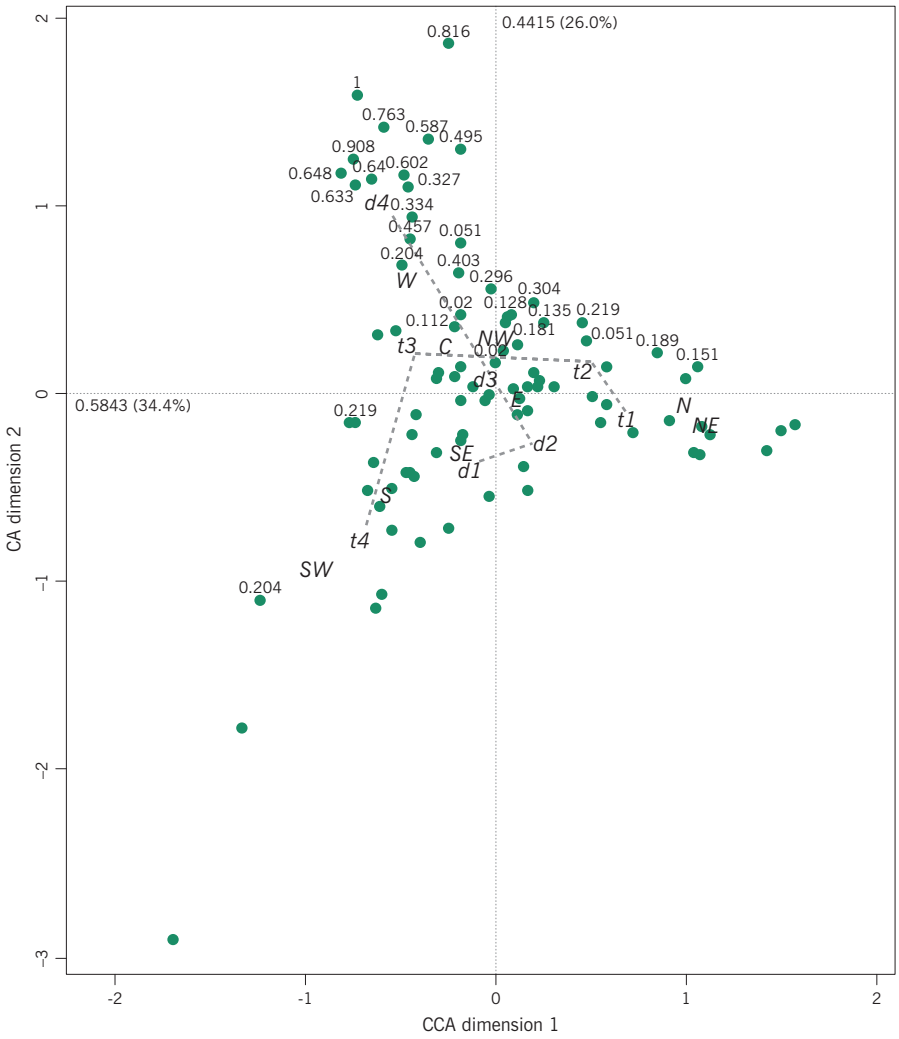
Exhibit 15.5:
CCA triplot of "Barents fish" data, with environmental variables coded into fuzzy categories. Again, sample sites are not shown (see Exhibit 15.6) but the weighted averages of all the fuzzy coded categories are, including the nine fuzzy spatial categories (eight compass points and central category)

59 values equal to 0 and 30 positive values varying from 0.020 to 1.000. These 30 values are shown in Exhibit 15.6 marking their corresponding sample positions, the remaining samples all having zero weights. These 30 samples are almost all in the upper and especially upper left of the ordination and the position of *d4* is at the weighted average of the positions of these 30 samples using those fuzzy values as weights. Especially in the upper left, those marked stations have high values of depth and thus high fuzzy values on category *d4* (the maximum depth is indicated by the value 1), and this leads to *d4* being where it is. In a similar way, all the other fuzzy categories have positions according to the weights placed on the sample points by the respective positive values in the fuzzy coding of the categories.

CCA as a discriminant analysis

The difference between CA of a table of abundances, say, and CCA of the same table constrained by some environmental variables, is that CCA tries to separate the samples on dimensions coinciding with the environmental variation. Thus, in Exhibits 15.5 and 15.6, which one can imagine overlaid to give the triplot of samples, species and variables, separation of the samples is achieved so that the categories of depth, temperature and spatial position are optimally separated in

Exhibit 15.6: *Positions of 89 samples in the CCA of Exhibit 15.5. Each category is at the weighted average of the sample positions, using the fuzzy values as weights. The positive values for category d4 are shown numerically at the respective sample positions*



the ordination. If there is a single environmental categorical variable, coded as a set of (crisp) dummy variables, then CCA with that variable as the constraining variable simplifies to a CA that can be thought of as a discriminant analysis between the categories. For example, suppose in the same data set we had a sediment type associated with each sample. Then all the abundances could be aggregated into each sediment type to obtain a sediment-by-species table where the (i,j)-th element would be the total abundance of species j in the sample with sediment type i. The CA of this aggregated table, with all the individual samples as supplementary points, is identical to the CCA with sediment as a categorical constraining variable.

SUMMARY: Canonical correspondence analysis

1. PCA, LRA and CA all have their constrained versions when additional information is available on each biological sample, and these additional variables are considered as predictors of the biological variation.

2. Canonical correspondence analysis (CCA) is the CA of an appropriate table regarded as responses (for example, an abundance matrix) where the dimensions of the result are constrained to be linear combinations of the predictor variables (for example, environmental variables). These predictor variables can be continuous or discrete.

3. CCA projects the response data onto the space of the predictors, and performs a CA in this restricted space. There is thus a splitting of the response inertia into two parts: the part related linearly to the predictors and the part unrelated to the predictors. The inertia of the former part becomes the new total inertia that is decomposed along ordination axes of the CCA. The biological variation that is unrelated to chosen predictors can also be of interest, especially when the variation due to a predictor variable needs to be partialled out of the analysis – this is then called *partial CCA*.

4. There are several advantages of coding the predictor variables fuzzily: nonlinear relationships between the ordination axes and the predictors can be handled, more of the response variable variance is usually explained, and the interpretation of the triplot is unified since all predictors are coded in a categorical way.

5. When there is just one predictor that is discrete, then the CCA constrained by this predictor is equivalent to a CA of the table of response data aggregated into the predictor categories, which in turn is a type of discriminant analysis between these categories.

INTERPRETATION, INFERENCE AND MODELLING

CHAPTER 16

Variance Partitioning in PCA, LRA, CA and CCA

Principal component analysis (PCA, log-ratio analysis (LRA) and correspondence analysis (CA) form one family of methods, all based on the same mathematics of matrix decomposition and approximation by weighted least-squares. The difference between them is the type of data they are applied to, which dictates the way distances are defined between samples and between variables. In all methods there is a measure of total variance as a weighted sum of squares of the elements of a matrix that is centred or double-centred – equivalently this total variance can be defined as a weighted sum of squared distances. This total variance can be broken down into various parts, parts for each row and for each column, parts along dimensions, and parts for each row and each column along the dimensions. This neat decomposition of variance provides several diagnostics to assist in the interpretation of the solution. The same idea applies to constrained analyses such as canonical correspondence analysis (CCA), where similar decompositions take place in the constrained space.

Contents

Total variance or inertia

PCA, LRA and CA all involve the decomposition of a matrix into parts across dimensions, from the most important dimension to the least important. Each method has a concept of total variance, also called *total inertia* when there are weighting factors. This measure of total variation in the data set is equal to the (weighted) sum of squared elements of the matrix being decomposed, and this total is split into parts on each ordination dimension. Let us look again at the matrix being decomposed as well as the total variance in each case.

The simplest case is that of PCA on unstandardized data, applicable to interval-level variables that are all measured on the same scale, for example the growths in millimetres of a sample of plants during the 12 months of the year (i.e., a matrix with 12 columns), or ratio-scale positive variables that have all been log-transformed. Suppose that the $n \times m$ cases-by-variables data matrix is $\mathbf{X}$, with elements x_{ij}, then the total variance of the data set, as usually computed by most software packages, is the sum of the variances of the variables:

$$\text{sum of variances} = \sum_j \left[\frac{1}{n-1} \sum_i (x_{ij} - \bar{x}_j)^2 \right] \tag{16.1}$$

Inside the square brackets is the variance of the j-th column variable of $\mathbf{X}$, and these variances are summed over the columns. Since we have introduced the concept of possible row and column weightings, we often prefer to use the following definition of total variance:

$$\text{total variance} = \frac{1}{mn} \sum_i \sum_j (x_{ij} - \bar{x}_j)^2 = \frac{n-1}{mn} \text{ (sum of variances)} \tag{16.2}$$

That is, we divide by n and not $n - 1$ in the variance computation, thus allocating an equal weight of $1/n$ to each row, and then average (rather than sum) the variances, thus allocating a weight of $1/m$ to each variable. So "total variance" here could rather be called *average variance*. This definition can easily be generalized to differential weighting of the rows and columns, so if the rows are weighted by $r_1, \ldots, r_n$ and the columns by $c_1, \ldots, c_m$, where weights are nonnegative and sum to 1 in each case, then the total variance would be:

$$\text{total variance} = \sum_i \sum_j r_i c_j (x_{ij} - \bar{x}_j)^2 \tag{16.3}$$

where the variable means $\bar{x}_j$ are now computed as weighted averages $\sum_i r_i x_{ij}$.

When continuous variables on different scales are standardized to have variance 1, then (16.1) would simply be equal to m, the number of variables. For our averaged versions (16.2) and (16.3) would be equal to $(n-1)/n$, or 1 if variance is defined using $1/n$ times the squared deviations, rather than $1/(n-1)$.

In the case of LRA of a matrix $\mathbf{N}$ of positive values, all measured on the same scale (usually proportions or percentages) and assuming the most general case of row- and column-weighting, the data are first log-transformed to obtain a matrix $\mathbf{L} = \log(\mathbf{N})$, and then double-centred using these weights. The total variance is then the weighted sum of squares of this double-centred matrix:

$$\text{total log-ratio variance} = \sum_i \sum_j r_i c_j (l_{ij} - l_{i\bullet} - l_{\bullet j} + l_{\bullet\bullet})^2 \tag{16.4}$$

where a subscript • indicates weighted averaging over the respective index. An equivalent formulation is to sum the squares of all the log odds ratios in the matrix formed by a pair (i, i') of rows and a pair (j, j') of columns, each weighted by the respective pairs of weights:

$$\text{total log-ratio variance} = \sum\sum_{i<i'} \sum\sum_{j<j'} r_i r_{i'} c_j c_{j'} \left[\log\left(\frac{n_{ij} n_{i'j'}}{n_{ij'} n_{i'j}} \right) \right]^2 \tag{16.5}$$

The *odds ratio* is a well-known concept in the analysis of frequency data: an odds ratio of 1 is when the ratio between a pair of row elements, say, is constant across a pair of columns. For example, in the fatty acid data set of Chapter 14, when two fatty acids j and j' in two different samples i and i' have the same ratio (e.g,. one is twice the other: $n_{ij} / n_{ij'} = n_{i'j} / n_{i'j'} = 2$), then $\log(1) = 0$ and no contribution to the total variance is made. The higher the disparities are in comparisons of this kind, the higher is the log-ratio variance.

Finally, in CA, which applies to a table **N** of nonnegative variables all measured on the same scale, usually count data, the total variance (called *inertia* in CA jargon) is closely related to the Pearson chi-square statistic for the table. The chi-square statistic computes expected values for each cell of the table using the table margins, and measures the difference between observed and expected by summing the squared differences, each divided by the expected value:

$$\text{chi-square statistic } \chi^2 = \sum_i \sum_j \frac{(\text{observed} - \text{expected})^2}{\text{expected}} \tag{16.6}$$

where the observed value = n_{ij} and the expected value = $n_{i+} n_{+j} / n_{++}$, and the subscript + indicates summation over the respective index. The chi-square statistic increases with the grand total n_{++} of the table (which is the sample size in a cross-tabulation), and the inertia in CA is a measure independent of this total. The relationship is simply as follows:

$$\text{total inertia} = \chi^2 / n_{++} \tag{16.7}$$

In spite of the fact that the definitions of total variance in PCA, LRA and CA might appear to be different in their formulations, they are in fact simple variations of the same theme. Think of them as measuring the weighted dispersion of the row or column points in a multidimensional space, according to the distance

function particular to the method: Euclidean distance (unstandardized or standardized) for PCA, log-ratio distance for LRA and chi-square distance for CA.

Variances on principal axes

The dimension-reduction step in all these methods concentrates a maximum amount of variance on the first ordination dimension, then – of the remaining variance – a maximum on the second dimension, and so on, until the last dimension, which has the least variance, in other words the dimension for which the dispersion of the rows or columns is closest to a constant. These dimensions, also called *principal axes,* define subspaces of best fit of the data – we are mainly interested in two-dimensional subspaces for ease of interpretation. We have seen various scree plots of these parts of variance on successive dimensions (Exhibits 12.6 and 13.3), which are eigenvalues of the matrix being decomposed by each method, and usually referred to as such in program results and denoted by the Greek letter λ, sometimes also called *principal variances* or *principal inertias.* What we are interested in now is the decomposition of these eigenvalues into parts for each row or each column.

Decomposition of principal variances into contributions

Thanks to the least-squares matrix approximation involved in this family of methods, there is a further decomposition of each ordination dimension's variance into part contributions made by each row and each column. The complete decomposition is illustrated schematically in Exhibit 16.1. Each eigenvalue can

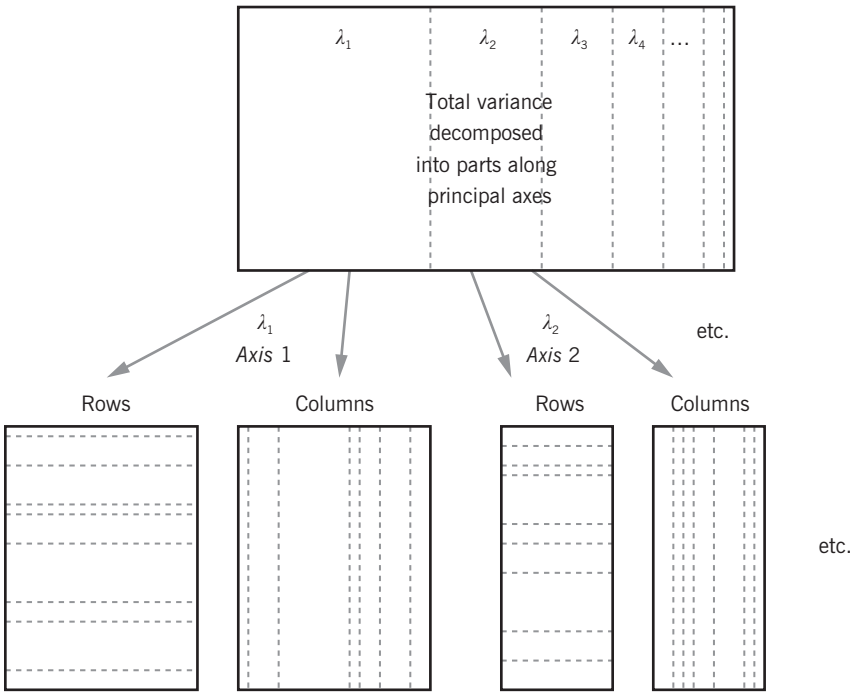


Exhibit 16.1: *Schematic explanation of the decomposition of total variance into parts. First, variance is decomposed from largest to smallest parts (λ_1, λ_2, ...) along successive principal axes. Then each λ can be decomposed into contributions either from the rows or from the columns. These part contributions to each axis provide diagnostics for interpretation of the results*

	dim1	*dim2*	*dim3*	*dim4*	*Sum*
a	0.0684	0.0013	0.0300	0.0025	*0.1022*
b	0.0107	0.0202	0.0108	0.0278	*0.0694*
c	0.2002	0.0001	0.0070	0.0012	*0.2086*
d	0.0013	0.0000	0.0248	0.0254	*0.0515*
e	0.0077	0.0989	0.0009	0.0045	*0.1120*
Sum	*0.2882*	*0.1205*	*0.0735*	*0.0614*	*0.5436*

Exhibit 16.2: *Tabulation of the contributions of five species in data set "bioenv" to the four principal inertias of CA: the columns of this table sum to the eigenvalues (principal inertias) and the rows sum to the inertia of each species*

be decomposed into parts, called *contributions,* either for the rows or for the columns. An example is given in Exhibit 16.2, for the CA of the "bioenv" data (see Exhibit 1.1), showing the contributions of each of the five species to the four dimensions of the CA. The grand total of this table, 0.5436, is equal to the total inertia in the data set. This is decomposed into four principal inertias, $\lambda_1 = 0.2882$, $\lambda_2 = 0.1205$, $\lambda_3 = 0.0735$, $\lambda_4 = 0.0614$ (column sums), while the values in each column are the breakdown of each principal inertia across the species. The row sums of this table are the inertias of the species themselves, and the sum of their inertias is again the total inertia.

The actual values in Exhibit 16.2 are by themselves difficult to interpret – it is easier if the rows and columns are expressed relative to their respective totals. For example, Exhibit 16.3 shows the columns expressed relative to their totals. Thus, the main contributors to dimension 1 are species *a* and *c*, accounting for 23.7% and 69.5% respectively, while species *e* is the overwhelming contributor to dimension 2, accounting for 82.1% of that dimension's inertia. To single out the main contributors, we use the following simple rule of thumb: if there are 5 species, as in this example, then the main contributors are those that account for more than the average of $1/5 = 0.2$ of the inertia on the dimension. In the last column the relative values of the inertias of each point is given when summed over all the dimensions, so that *c* is seen to have the highest inertia in the data matrix as a whole.

	dim1	*dim2*	*dim3*	*dim4*	*All*
a	0.2373	0.0110	0.4079	0.0410	*0.1880*
b	0.0370	0.1676	0.1463	0.4527	*0.1277*
c	0.6947	0.0005	0.0959	0.0200	*0.3837*
d	0.0045	0.0001	0.3373	0.4130	*0.0947*
e	0.0266	0.8208	0.0126	0.0733	*0.2060*
Sum	*1.0000*	*1.0000*	*1.0000*	*1.0000*	*1.0000*

Exhibit 16.3: *Contributions of the species to the principal inertias and the total inertia (Exhibit 16.2 re-expressed as values relative to column totals)*

The alternative way of interpreting the values in Exhibit 16.2 is to express each of the rows as a proportion of the row totals, as shown in Exhibit 16.4. This shows how each species' inertia is distributed across the dimensions, or in regression terminology, how much of each species' inertia is being explained by the dimensions as "predictors". Because the dimensions are independent these values can be aggregated. For example, 66.9% of the inertia of species *a* is explained by dimension 1, and 1.3% by dimension 2, that is 68.2% by the first two dimensions. Notice in the last row of Exhibit 16.4 that the eigenvalues are also expressed relative to their total (the total inertia), showing overall how much inertia is explained by each dimension. For example, 53.0% of the total inertia of all the species is explained by dimension 1, but dimension 1 explains different percentages for each individual species: 66.9% of *a*, 15.4% of *b*, etc.

Contribution coordinates

Exhibit 16.3 shows the contributions of each species to the variance of each dimension. These contributions are visualized in the contribution biplot, described in Chapter 13. Specifically, the contribution coordinates are the square roots of these values, with the sign of the respective principal or standard coordinate. For example, the absolute values of the contribution coordinates for species *a* on the first two dimensions are $\sqrt{0.2373} = 0.487$ and $\sqrt{0.0110} = 0.105$ respectively, with appropriate signs. Arguing in the reverse way, in Exhibit 13.5 the contribution coordinate on dimension 1 of the fish species *Bo_sa* is –0.874, hence its contribution to dimension 1 is $(-0.874)^2 = 0.763$, or 76.3%.

Correlations between points and axes

Just like the signed square roots of the contributions of points to axes in Exhibit 16.3 have a use (as contribution coordinates), so the signed square roots of the contributions of axes to points in Exhibit 16.4 also have an interesting interpretation, namely as correlations between points and axes, also called *loadings* in the factor analysis literature. For example, the square roots of the values for species *a* are $\sqrt{0.6690} = 0.818$, $\sqrt{0.0130} = 0.114$, $\sqrt{0.2934} = 0.542$, $\sqrt{0.0246} = 0.157$. These are the absolute values of the correlations of species *a* with axis 1, with signs depending on the sign of the respective principal or standard coordinates. Species *a* is thus highly correlated with axis 3. Remember that the four

Exhibit 16.4: *Contributions of the dimensions to the inertias of the species (Exhibit 16.2 re-expressed as values relative to row totals). In the last row the principal inertias are also expressed relative to the grand total*

	dim1	*dim2*	*dim3*	*dim4*	*Sum*
a	0.6690	0.0130	0.2934	0.0246	*1.0000*
b	0.1536	0.2910	0.1550	0.4004	*1.0000*
c	0.9600	0.0003	0.0338	0.0059	*1.0000*
d	0.0251	0.0002	0.4820	0.4928	*1.0000*
e	0.0684	0.8832	0.0083	0.0402	*1.0000*
All	*0.5302*	*0.2216*	*0.1352*	*0.1129*	*1.0000*

ordination dimensions function as independent predictors of the rows or columns of the data set, with the first dimension explaining the most variance, the second dimension the next highest, and so on. Each value in Exhibit 16.4 is a squared correlation and can be accumulated to give an R^2 coefficient of determination, thanks to the zero correlation between the dimensions. Summing these for the first two dimensions in Exhibit 16.4 gives R^2's for the five species of 68.2%, 44.5%, 96.0%, 2.5% and 95.2% respectively – clearly, species *d* is very poorly explained by the first two dimensions, and from Exhibit 16.4 it is mainly explained by the last two dimensions.

Contributions in constrained analyses

In canonical correspondence analysis the dimensions are constrained to be linear combinations of external predictors. This restricts the space in which dimension reduction will take place. As seen in Chapter 15, for the "bioenv" data set, the amount of inertia in this constrained space is equal to 0.2490, compared to the 0.5436 in the unconstraineded space. This lower value of 0.2490 now becomes the total inertia that is being explained, otherwise everything is as before. Exhibit 16.5

(a)

	dim1	*dim2*	*dim3*	*dim4*	*Sum*
a	0.0394	0.0002	0.0025	0.0003	*0.0424*
b	0.0131	0.0035	0.0029	0.0005	*0.0201*
c	0.1470	0.0001	0.0001	0.0002	*0.1474*
d	0.0000	0.0040	0.0000	0.0022	*0.0061*
e	0.0013	0.0310	0.0006	0.0001	*0.0330*
Sum	*0.2008*	*0.0388*	*0.0060*	*0.0033*	*0.2490*

(b)

	dim1	*dim2*	*dim3*	*dim4*	*All*
a	0.1962	0.0064	0.4127	0.0818	*0.1703*
b	0.0652	0.0910	0.4860	0.1614	*0.0807*
c	0.7321	0.0015	0.0093	0.0682	*0.5919*
d	0.0001	0.1018	0.0003	0.6526	*0.0247*
e	0.0063	0.7993	0.0916	0.0360	*0.1324*
Sum	*1.0000*	*1.0000*	*1.0000*	*1.0000*	*1.0000*

(c)

	dim1	*dim2*	*dim3*	*dim4*	*Sum*
a	0.9291	0.0058	0.0587	0.0064	*1.0000*
b	0.6519	0.1757	0.1459	0.0266	*1.0000*
c	0.9977	0.0004	0.0004	0.0015	*1.0000*
d	0.0042	0.6438	0.0003	0.3516	*1.0000*
e	0.0386	0.9410	0.0168	0.0036	*1.0000*
All	*0.8066*	*0.1559*	*0.0242*	*0.0133*	*1.0000*

Exhibit 16.5:
(a) Raw contributions to four dimensions by the species in the CCA of the "bioenv" data (see Chapter 15). The row sums are the inertias of the species in the restricted space, while the column sums are the principal inertias in the restricted space. (b) The contributions relative to their column sums (which would be the basis of the CCA contribution biplot. (c) The contributions relative to their row sums (i.e., squared correlations of species with axes)

shows the tables corresponding to Exhibits 16.2, 16.3 and 16.4 for the CCA. Having restricted attention to that part of the variance correlated with the external predictors (depth, pollution, temperature and sediment), it now appears that a much larger percentage of the inertia is explained on the first two axes. Thus the species are better explained now according to Exhibit 16.5(c), but we should emphasize that we are explaining only the restricted part of the species inertia in the space of the environmental predictors.

Finally, each of the external predictors can be correlated with the axes, and the proportion of the predictors' variance explained by the dimensions can be computed. Exhibit 16.6 shows the squared correlations with each axis, where the sites are weighted by their usual masses in the computation of the correlation. Squared correlations can be accumulated to give proportions of variance explained.

Exhibit 16.6: *Squared correlations of each predictor variable with each CCA ordination axis. In computing the correlations the weights of the cases (sites in this example) are used. The values can be accumulated across the columns of this table to give proportions of variance explained by sets of dimensions*

	dim1	*dim2*	*dim3*	*dim4*	*Sum*
Depth	0.129	0.723	0.081	0.004	*0.938*
PolIn	0.980	0.011	0.008	0.000	*0.999*
Temp	0.000	0.033	0.496	0.462	*0.992*
C	0.010	0.078	0.511	0.128	*0.727*
S	0.118	0.030	0.012	0.040	*0.200*
G	0.169	0.165	0.269	0.249	*0.852*

Exhibit 16.7: *Improved squared correlations of sediment categories with the four ordination axes of the CCA, considering them as supplementary row points that aggregate the species abundances in the sites corresponding to each category*

	dim1	*dim2*	*dim3*	*dim4*	*Sum*
C	0.023	0.079	0.897	0.001	*1.000*
S	0.186	0.284	0.526	0.004	*1.000*
G	0.571	0.226	0.191	0.012	*1.000*

For example, depth has $0.129 + 0.723 = 0.852$, i.e. 85.2%, of its variance explained by the first two dimensions. Pollution is almost 100% explained by the first two dimensions, while temperature is very poorly explained. The dummy variables for sediment are also poorly explained, but this is mostly due to the fact that they take on only two values. An improved measure of fit can be obtained by considering the categories rather as groupings of cases (i.e., supplementary row points rather than dummy variable column points in the CCA), just like we displayed categories as centroids of the site points corresponding to the respective categories. Exhibit 16.7 shows the squared correlations of the sediment categories

as row profiles with the four dimensions of the CCA solution. Thus gravel has a variance explained by the first two dimensions of 79.7%, compared to 33.4% according to Exhibit 16.6.

SUMMARY: Variance partitioning in PCA, LRA, CA and CCA

1. All these methods that analyse a rectangular data table share a common theory and a common objective – what differentiates them is the distance function inherent in structuring the space of the rows and columns, which in turn is a function of the type of data being analysed. Weights for the rows and columns are implicit in all the methods, even when the weights are equal.
2. In each method, based on the weights and the metric, a matrix is formed for dimension reduction, and the total variance of the data is measured by the sum of squares of that matrix. In PCA it is the sum (or average) of the variances of the variables; in LRA it is the weighted sum of all logarithms of the odds ratios in the matrix; in CA it is the total inertia, the chi-square statistic divided by the grand total of the table; in CCA it is that part of the total inertia that is projected onto the space of the explanatory variables.
3. The total variance or inertia is decomposed along principal axes, in decreasing parts, such that the part accounted for by the first axis is the maximum, and then of the remaining inertia the second axis accounts for the maximum, and so on.
4. Each part of variance on the principal axes is decomposed in turn into contributions by the rows (usually cases) or by the columns (usually variables) of the data table.
5. These contributions can be used as diagnostics in two ways: interpreting how each axis is built up from the rows or from the columns, or interpreting how each row or column is explained by the axes. The relative contributions of the axes to the variances of a row or column are squared correlations and can be summed to obtain a R^2 measure of explained variance of the row or column.
6. The contributions of the columns to the axes are what are visualized in the contribution biplot, because the columns usually define the variables (e.g., species) of the table. In principle, one can define contribution coordinates for the rows as well.
7. All of the above applies similarly to constrained forms of these methods, where the total variance is restricted to the part that is directly related to a set of explanatory variables.

CHAPTER 17

Inference in Multivariate Analysis

We have presented multivariate analysis as the search for structure and relationships in a complex set of data comprising many sampling units and many variables. Groupings are observed in the data, similarities and differences reveal themselves in ordination maps, and the crucial question then arises: are these observed patterns just randomly occurring or are they a signal observed in the data that can be considered significant in the statistical sense? In this chapter we shall tackle this problem in two different ways: one is using bootstrapping, to assess the variability of patterns observed in the data, analogous to setting confidence intervals around an estimated statistic, and the other is using permutation tests in order to compute *p*-values associated with the testing of different hypotheses. We will illustrate these computational approaches to statistical inference in two different situations, where group differences or associations between variables are being assessed. Before tackling the multivariate context we shall treat more familiar univariate and bivariate examples in each respective situation.

Contents

Univariate test of group difference

One of the simplest statistical tests to perform is the two-group *t*-test of difference in means between two populations, based on a sample from each population. Taking our "bioenv" data set as an example, suppose we aggregate the samples corresponding to clay (C) and sand (S) sediment (labelled as group CS), to be compared with the gravel sediment sample (G). We want to perform a hypothesis test to compare the pollution values for the 22 sites of CS with the 8 sites of G. The mean pollution in each group is computed as:

$$\bar{x}_{CS} = 5.18 \qquad \bar{x}_{G} = 2.70$$

Performing the usual *t*-test and not worrying for the moment whether the assumptions of this test have been satisfied, we obtain a *t*-statistic (with 30 – 2 = 28 degrees of freedom) of 3.22 with associated *p*-value for a two-sided test of 0.0032. Thus we would conclude that there is truly a difference in pollution between clay/sand and gravel samples, with gravel samples having less pollution on average. The estimated difference in the mean pollution is 2.48 and a 95% confidence interval for the difference in the means is [0.90, 4.05]. Now this simplest form of the *t*-test assumes that the data are normally distributed and that the variances are equal in the two groups. An alternative form of the *t*-test, known as the *Welch test,* does not assume equal variances and obtains a *t*-statistic of 4.62, a lower *p*-value of 0.00008 and a much narrower 95% confidence interval of [1.38, 3.58]. If we examine the normality by making a normal quantile plot and using the *Shapiro-Wilks test*[1] in these quite small samples, there is no strong evidence that the data are not normal.

An alternative distribution-free approach to this test, which does not rely on the normality assumption, is to perform a *permutation test.* Under the hypothesis of no difference between the two groups it is assumed they come from one single distribution of pollution, so any observation could have been in the clay/sand group or the gravel group. So we randomly assign the 30 pollution observations to a sample consisting of 22 of the values, with the remaining 8 values in the other sample, and recompute the difference in the group means. The number of ways we can randomly separate the 30 values into two samples of 22 and 8, is:

$$\binom{30}{22} = \binom{30}{8} = 5{,}852{,}925$$

that is, almost 6 million ways. In such a situation we do this random allocation a large number of times, typically 9,999 times, plus the actual observed samples, giving 10,000 permutations in total. The distribution of these 10,000 values, called the *permutation distribution,* is given in Exhibit 17.1 – it is the distribution of the difference in means under the null hypothesis of no difference. To obtain a *p*-value we see where our observed difference of 2.48 lies on the distribution, counting how many of the random permutations give differences higher or equal to 2.48, as well as lower or equal to −2.48, since the test is two-sided. There are 29 permutations outside these limits so the *p*-value is 29/10,000 = 0.0029, which is compatible with the *p*-value calculated initially for the regular *t*-test.

[1] See Appendix C for descriptions of the functions used, and the online R code.

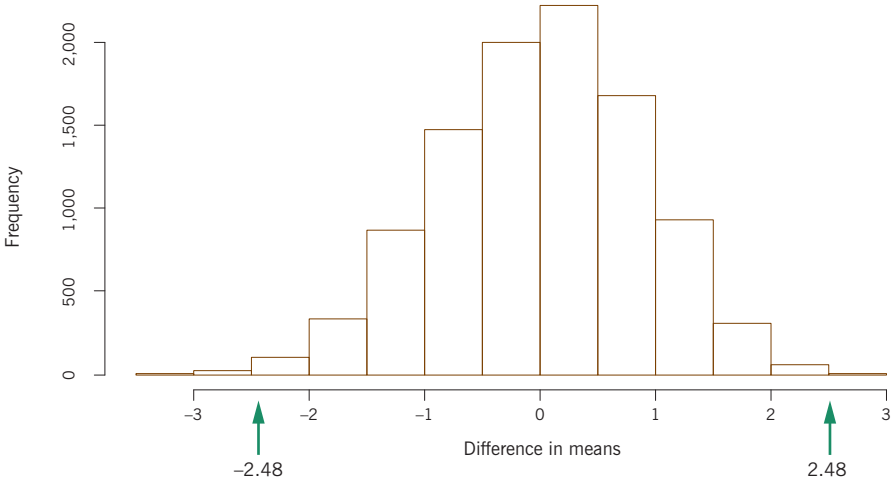


Exhibit 17.1: *Permutation distribution for test of difference in means of two populations based on samples of size 22 and 8. Of the 10,000 permutations 29 lie outside the limits of ± 2.48, hence the estimated* p*-value is 0.0029*

To estimate the variability of the estimated difference correctly, without recourse to distributional assumptions, we would need repeated pollution samples of size 22 and 8 respectively from the populations of clay/sand and gravel locations from which the original data were obtained, which is clearly not possible since we only have one set of data. To simulate data from these two populations we can resort to *bootstrapping* the data. Samples are taken from the two sets of data, with replacement, which means that the same observation can be chosen more than once and some not at all. We do this repeatedly, also 10,000 times for example, each time computing the difference in means, leading to the *bootstrap distribution* of this difference, shown in Exhibit 17.2, alongside the separate bootstrap distributions of

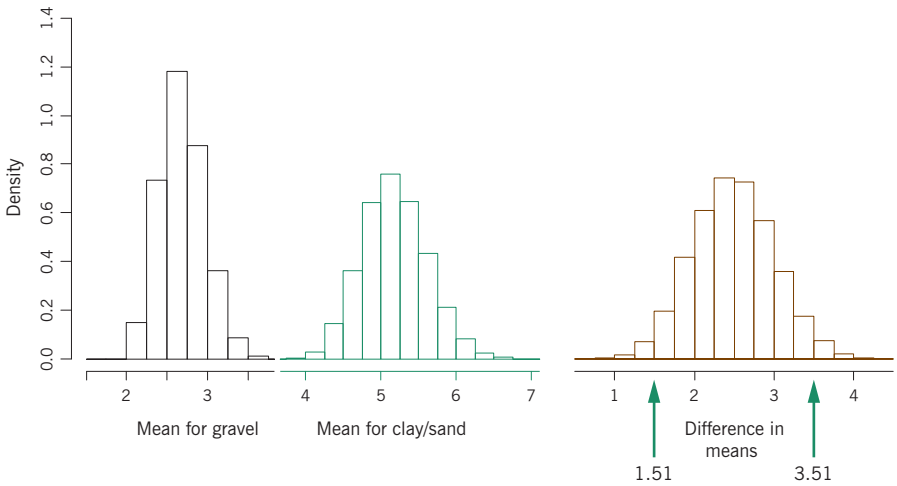


Exhibit 17.2: *Bootstrap distributions of the mean pollution for the gravel and clay/sand groups, based on 22 samples and 8 samples respectively, drawn with replacement 10,000 times from the original data. The right hand histogram is the bootstrap distribution of the differences, showing the limits for a 95% confidence interval*

the two populations. The bootstrap distribution is an estimate of the true distribution of the differences, so to estimate the 95% confidence interval, we cut off 2.5% (i.e., 250 values out of 10,000) on each side of the distribution, obtaining the interval [1.51, 3.51]. This is more in line with the confidence interval obtained by the Welch method.

Test of association between two variables

Another common situation in statistical inference is to test an association, for example correlation, that is measured between two variables. In Chapter 1 we calculated a correlation between pollution and depth in the "bioenv" data set of –0.396 and a p-value of 0.0305 according to the two-tailed t-test for a correlation coefficient. This test relies on normality of the data but a distribution-free permutation test can also be conducted, as follows. Under the null hypothesis of zero correlation there is no reason to link any observation of depth with the corresponding observation of pollution in a particular sample, so we can randomly permute one of the data vectors. We do this 9,999 times, computing the correlation coefficient each time, and Exhibit 17.3 is the permutation distribution. The observed value of –0.396 is exceeded in absolute value by 315 of these randomly generated ones, and so the estimated p-value is 315/10,000 = 0.0315, almost the same as the t-test result.

Bootstrapping can be performed to obtain a confidence interval for the correlation coefficient. Now the pairs of depth and pollution values are kept together, and the sampling is done from their bivariate distribution by taking 30 samples at a time from the data set, with replacement (again, some samples are chosen more

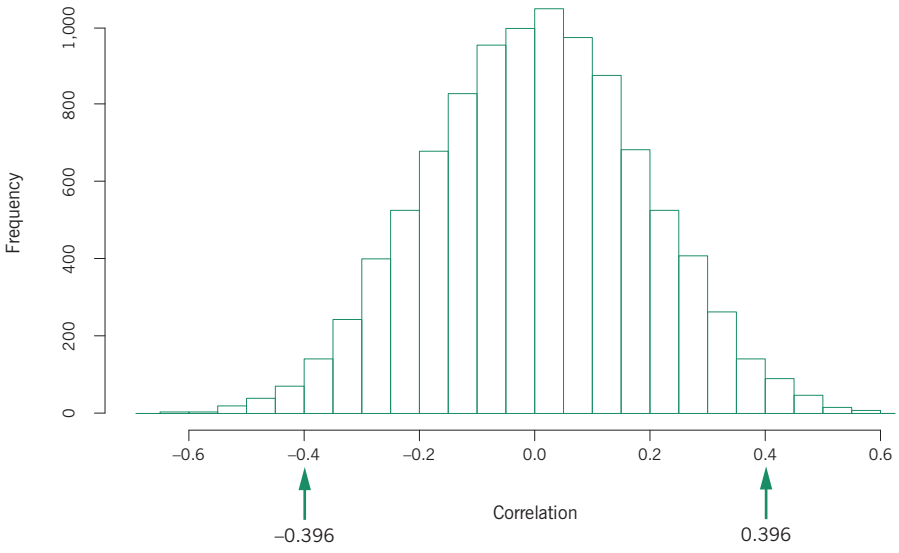


Exhibit 17.3: *Permutation distribution based on 9,999 estimates of the correlation between depth and pollution, under the null hypothesis of no correlation, together with the observed value of –0.396. The values ± 0.396 are indicated – there are 315 values equal to or more extreme, hence the* p*-value is 0.0315*

than once, some not at all). This is done 10,000 times, and each time a correlation coefficient is calculated, and Exhibit 17.4 shows their distribution. Cutting off 2.5% of the values on each tail gives a two-sided confidence interval for the correlation of [–0.704, –0.044]. Notice that the distribution in Exhibit 17.4 is not symmetric, and that this 95% confidence interval does not include zero, which is another way of saying that the observed correlation is significant at the 5% level of significance.

We have not exhausted all the possible alternative approaches in the last two sections. For example, a nonparametric Kruskal-Wallis rank test can be performed to test the difference in pollution between clay/sand and gravel samples, leading to a p-value of 0.0017. Or a Spearman rank coefficient can be computed between depth and pollution as –0.432 and its p-value is 0.021. Both these alternative approaches give results in the same ball-park as those obtained previously. Having shown these alternative ways of assessing statistical significance, based on statistical distribution theory with strong assumptions on the one hand, and using computationally intensive distribution-free methods on the other hand, the question is: which is preferable? It does help when the different approaches corroborate one another, but there is no correct method. However, we can eliminate methods that clearly do not fit the theory, for example normality-based methods should not be used when the data are clearly not normal. When we come to the multivariate case, however, the situation is much more complex, and in the absence of a theoretical basis for statistical testing, we rely more on the distribution-free approaches of permutation testing and bootstrapping.

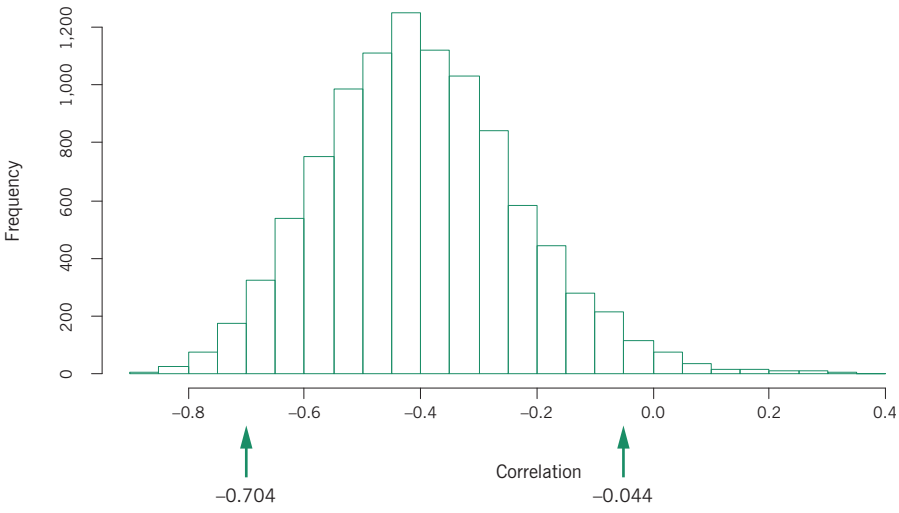
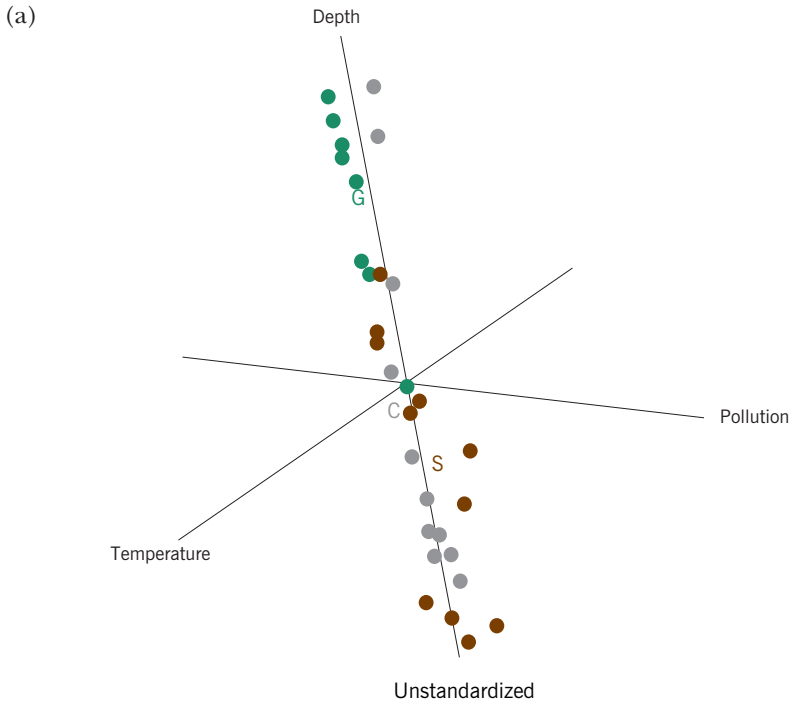
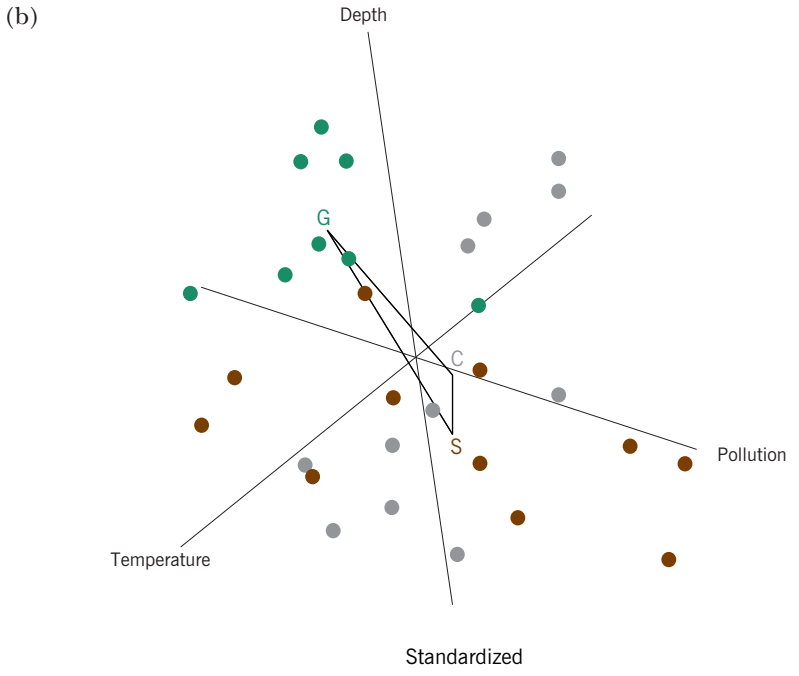


Exhibit 17.4: *Bootstrap distribution of the correlation coefficient, showing the values for which 2.5% of the distribution is in each tail*

Exhibit 17.5: *Three-dimensional views of the 30 samples in the unstandardized (a) and standardized (b) Euclidean space of the three variables. Clay, sand and gravel samples are colour coded as gray, brown and green respectively, and their group average positions denoted by* C, S *and* G. *Since depth has a much higher range of numerical values than the other two variables, it would dominate the computation of inter-group difference if the data were not standardized in some way*





Multivariate test of group difference

Suppose now that we wanted to test the differences between the sediment groups based on all three continuous environmental variables depth, pollution and temperature. This time let us also keep the three sediment groups separate. Even in the univariate case, when we pass to more than two groups, the notion of a negative difference no longer exists – any measure of difference will be a strictly positive number. Furthermore, when we pass to more than one variable then the issue of standardization is crucial in the measure of group difference, since the variables have to contribute equitably to the measure of difference. To rub this point in even further, consider the positions of the samples in unstandardized and standardized coordinates in Exhibit 17.5. The centroids of the three sediment groups are also shown, and it is clear that standardization is necessary, otherwise depth would dominate any measure of intergroup difference. We are going to measure the difference between the three sediment groups by the lengths of the sides of the triangle in the standardized space – see the triangle in Exhibit 17.5(b). If these lengths are large then the group means are far apart, if they are small then the means are close together. The question then is whether they are significantly far apart. The sum of these three lengths turns out to be 6.303. To obtain a p-value a permutation test is performed by randomly allocating the C, S and G labels to the data samples many times, and each time computing the same statistic, the sum of the distances between group means. The permutation distribution is shown in Exhibit 17.6, and the observed statistic lies well into the tail of the distribution, with a p-value of 0.0032. Notice that now it is only the right tail that is counted, since the value of 0 on the left side of the distribution indicates the null hypothesis of no difference.

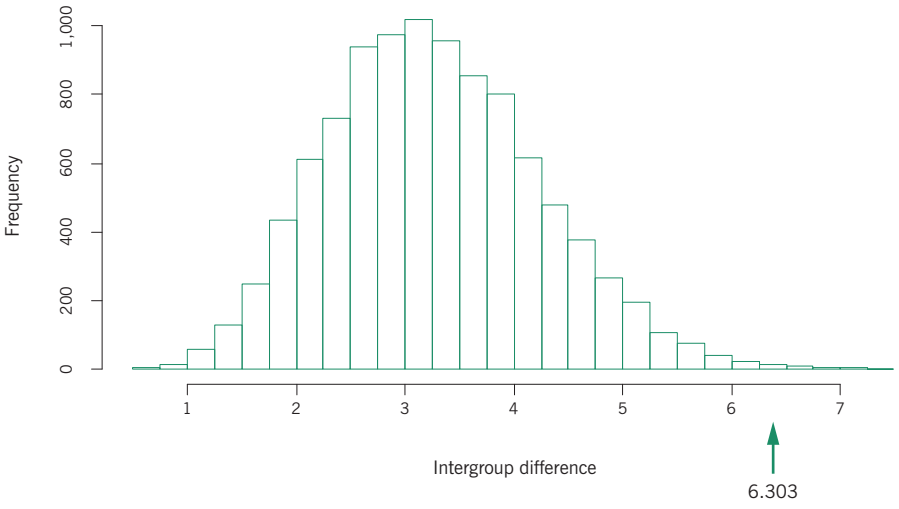
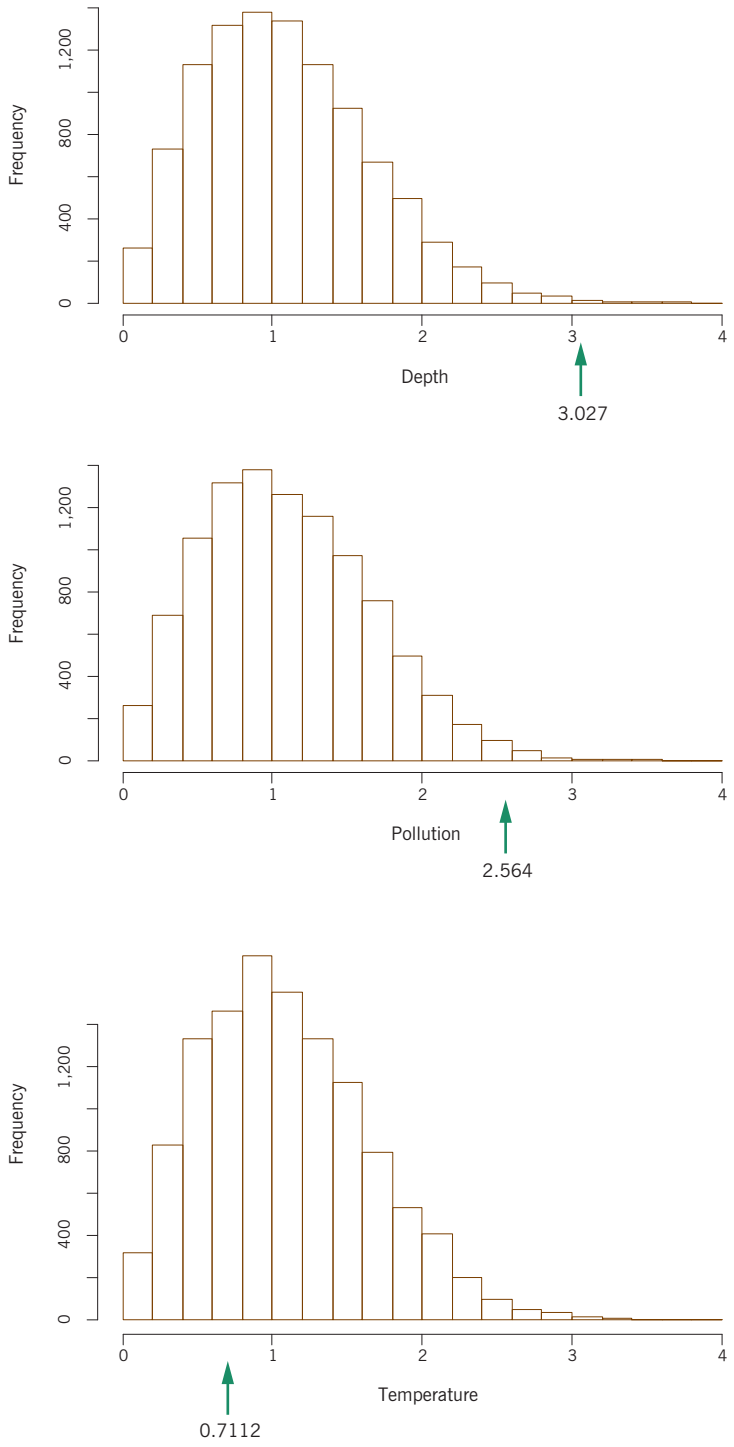


Exhibit 17.6: *Permutation distribution of measure of intergroup difference in standardized multivariate space. There are 32 of the simulated values greater than or equal to the observed value of 6.303, hence the* p*-value is 32/10,000 = 0.0032*

Exhibit 17.7: *Permutation distributions for measure of intergroup difference based on single variables. The observed difference is indicated each time and the* p*-values are 0.0032, 0.0084 and 0.7198 respectively.*



Having concluded that the group differences are significant there are two further aspects to be considered, since there are three variables and three groups involved: first, are all groups significantly different from one another? and second, which variables contribute mostly to the difference? These two questions are related, since it may be that a group may be different from another on only one or two of the variables, whereas the third group may be different from the other two on all three variables. Anyway, let us consider the latter question first: are the groups significantly different on all three variables? We can perform the same test three times using one variable at a time – here it would not matter if the data were standardized or not, but we continue to use the standardized form since it is easier to compare the three results. Exhibit 17.7 shows the three permutation distributions and it is clear that temperature is not at all different between the three sediment groups, so we could drop it from further consideration.

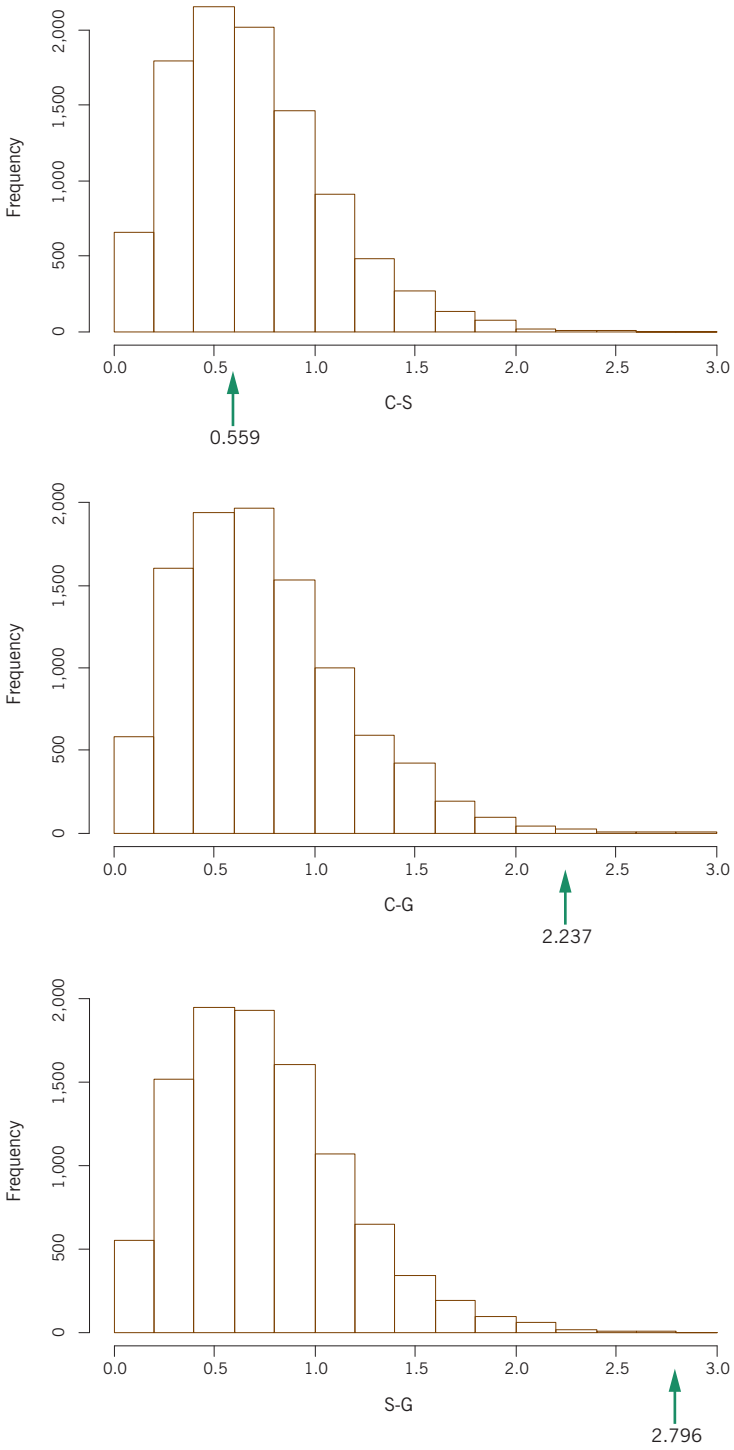
Next, we examine whether the groups are all different from one another, based on just depth and pollution. The differences between C and S, S and G and C and G are computed, similar to the multiple comparison procedure in ANOVA, and their permutation distributions generated, shown in Exhibit 17.8. It is clear that there is no significant difference between clay and sand groups (hence our aggregating them in the initial example of this chapter), whereas they are both highly significantly different from the gravel group.

Test of association between groups of variables

The multivariate equivalent of testing a correlation coefficient is when there are several predictor variables being used to explain one or more response variables. The most relevant case to us is in canonical correspondence analysis (CCA), when many biological variables are being related to several environmental variables, for example, via a process of dimension reduction in the biological space. There are two ways to assess this relationship: one way is to simply include all the environmental variables in the model and test for their overall significance, while another more laborious way is to look for a subset of significant environmental predictors, eliminating the insignificant ones. We shall illustrate these two strategies again using the simple "bioenv" data set, leaving a more substantial application to the case study of Chapter 19.

The inertia of the biological data in this example (species *a*, *b*, *c*, *d* and *e*) is 0.544 (see Chapter 13). When using depth, pollution and temperature as environmental predictors in a CCA, the inertia accounted for is 0.240, or 44.1%. We can generate a permutation distribution to test whether this percentage is significant. As in the case of the correlation of two variables, under the null hypothesis of no relationship, the biological and environmental data vectors can be randomly paired, keeping the biological vectors (with

Exhibit 17.8: *Permutation distributions for measure of pairwise intergroup differences based on depth and pollution. The observed difference is indicated each time and the* p*-values are 0.5845, 0.0029 and 0.0001 respectively*



five abundance values) and the environmental vectors (with three continuous measurements) intact. If we do this 9,999 times, we do not get any case that improves the figure of 44.1% inertia explained, so the p-value is 0.0001, highly significant.

Dropping one variable at a time and repeating this exercise, we obtain the following percentages of explained inertia and p-values for the three different pairs of variables:

Depth and pollution:	42.7% ($p = 0.0001$)
Depth and temperature:	11.4% ($p = 0.1366$)
Pollution and temperature:	37.4% ($p = 0.0001$)

and for a single variable at a time:

Depth:	10.0% ($p = 0.0366$)
Pollution:	36.3% ($p = 0.0001$)
Temperature:	1.1% ($p = 0.8642$)

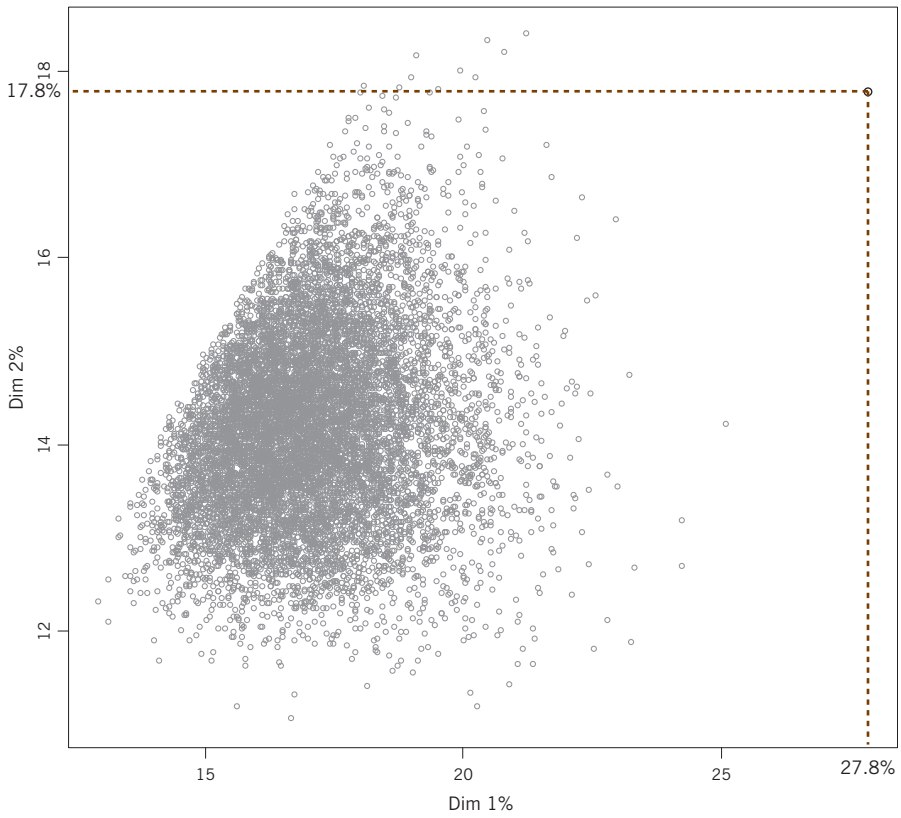
So it seems clear that temperature has no predictive role and can be dropped. Pollution is the best predictor and if a forward stepwise process were followed, then pollution would be the first to enter the model. The only question remaining is whether depth adds significantly to the model that is driven mainly by pollution. This can be tested by generating a permutation distribution with pollution as a predictor, unpermuted, while just the depth values are permuted randomly. After the usual 9,999 permutations of the depth vector, the result is that the percentage of inertia explained by depth and pollution, seen above to be 42.7%, is the 222nd highest value in the sorted list of 10,000, so the p-value for the additional explained inertia of depth is 0.0222, significant at the 5% level. The final model would thus include pollution and depth.

Other permutation tests for ordination

In any dimension-reduction technique to establish an ordination of a data set, the objective is to separate what is "signal", that is true structure, from "noise", that is random variation. In Chapter 12 we discussed an informal way of judging which dimensions are "significant" from the appearance of the scree plot (see Exhibit 12.6 and related description). A permutation test can make a more formal decision about the dimensionality of the solution. In a PCA, the correlations between the variables combine to form the principal

axes of an ordination, so if there is no correlation between the variables then there is no structure. Hence a permutation test for the principal axes of the "climate" data set can be performed by generating several random data sets under the null hypothesis of no correlation between the variables by randomly permuting the values down the columns of each variable. The eigenvalues of these randomly generated data sets yield permutation distributions of the eigenvalues under the null hypothesis. Since the total variance in a PCA of standardized data is a fixed number, it is equivalent to look at the percentages of variance explained on the axes. Exhibit 17.9 shows the scatterplot of the first and second percentages of variance for the 9,999 permuted data sets, along with the actual values of 27.8% and 17.8% in the original data set. The p-values are again calculated by counting how many of the values are greater than or equal to the observed ones, only 1 for the first dimension (the observed value itself) and 13 for the second, hence the p-values are 0.0001 and 0.0013 respectively. Continuing with the third and higher dimensions, the p-values are 0.0788, 0.2899, 0.9711 and so on, none of which is significant.

Exhibit 17.9: *Scatterplot of percentages of variance on the first two dimensions of 10,000 PCAs, one of which is based on the observed data set "climate" and the other 9,999 are computed using random matrices obtained by permutation*



Hence, the two-dimensional solution, accounting for 45.6% of the variance, is considered the appropriate solution, and the remaining 54.4% of the variance is regarded as random variation.

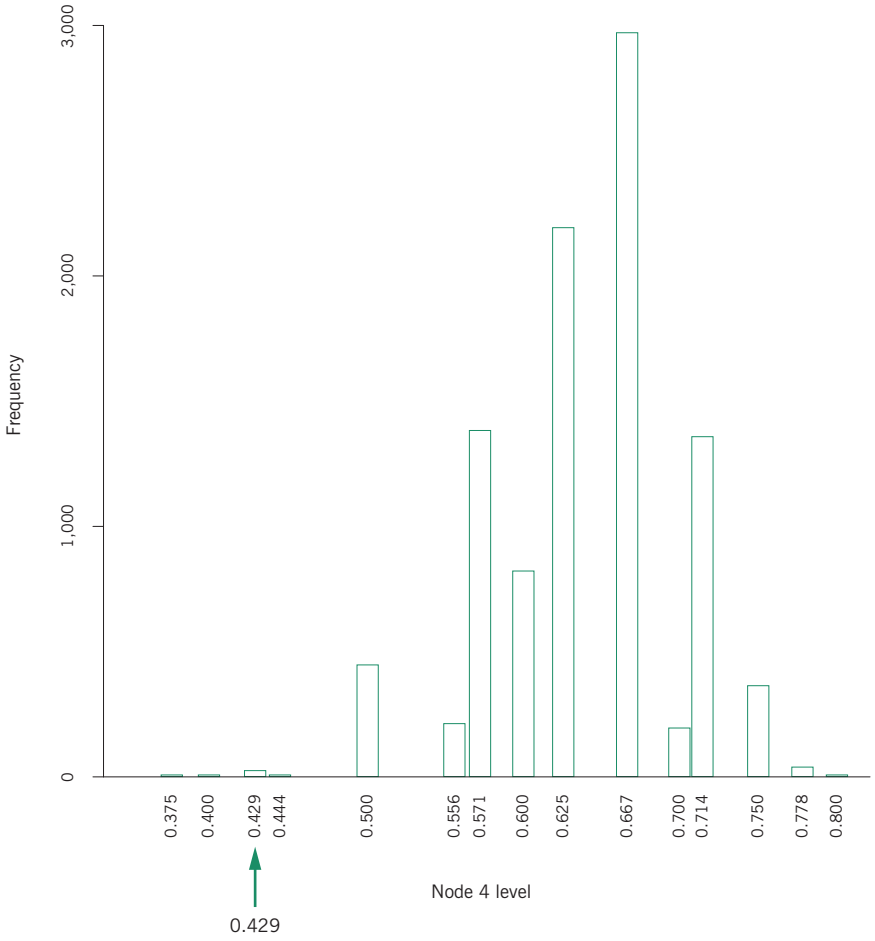
The situation is slightly different for methods like CA, LRA and CCA that do not have a fixed total variance, and have weights attached to the rows and columns. Whereas in PCA it is equivalent to consider the eigenvalues or their percentages relative to the total variance, in these other methods such as CA, for example, the total inertia of an abundance matrix can be very high compared to the variance of the permuted matrices under a null model of no relationship between the species. So we would base our decision about significance of the dimensions purely on the percentages of inertia. An example will be given in the case study of Chapter 19.

A permutation test for clustering

The same issues arise when performing a cluster analysis: for example, which are the "significant" clusters, or in a hierarchical clustering at what level should the dendrogram be cut for the identification of "significant" clustering? This is a difficult question and we present one possible approach to identifying a significant cutpoint. The levels of the nodes at which clusters are formed are first saved for the original dendrogram, for example in the dendrogram on the right of Exhibit 7.8, based on clustering zero/one data using the Jaccard index, the levels of the nodes are (from the bottom up): 0.200, 0.250, 0.333, 0.429, 0.778 and 1.000. Now we generate a permutation distribution for these levels by randomly permuting the columns of the data matrix, given in Exhibit 5.6, so that we have a large number of simulated values (again, 9,999) under the hypothesis of no relationship between the species. For each permuted matrix the node levels are stored, and then for each level we count how many are less than or equal to the originally observed node level. For significant clustering we would expect the node level to be low. The p-values associated with each node are (again, from bottom up): 0.1894, 0.0224, 0.0091, 0.0026, 0.7329, 1.000, so that node level 4, which cuts the sample into three clusters, is the most significant. Exhibit 17.10 shows the permutation distribution for the node 4 levels and the observed value of 0.429. There are only 26 permutations where the node level is lower than or equal to 0.429, hence the p-value of 0.0026.

As a contrasting example, the same strategy was applied to the dendrogram of Exhibit 7.10 that clusters 30 samples based on their standardized Euclidean distances using variables depth, pollution and temperature. None of the p-values for the 29 nodes in this example are less than 0.05, which indicates that there are no real clusters in these data, but rather a continuum of dispersion in multivariate space.

Exhibit 17.10: *Permutation distribution of node 4 levels, corresponding to a three-cluster solution, for the presence-absence data of Exhibit 5.6 – see the dendrogram on the right of Exhibit 7.8. There are 26 permutations (out of 10,000) that are less than or equal to 0.429, the value of the level in the dendrogram*



SUMMARY: Inference in multivariate analysis

1. Conventional statistical inference relies on assuming an underlying distribution of the data under a null hypothesis (e.g., a hypothesis of no difference, or of no correlation), called the *null distribution*. The unusualness of the observed value (e.g., a difference or a correlation) is then judged against the null distribution and if its probability of occurring (i.e., *p*-value) is low, then the result is declared statistically significant.

2. Distribution-free methods exist that free the analyst from assuming a theoretical distribution: null distributions can be generated by permuting the data under the null hypothesis, and the variability of observed statistics can be estimated using bootstrapping of the observed data.

3. In the multivariate context, where theory is much more complex, we shall generally rely purely on computer-based permutation testing and bootstrapping.

4. To assess the significance of group differences, the null hypothesis of no difference implies that we can allocate observations to any groups. A statistic measuring group difference is postulated, and then the same statistic is measured on data that have been randomly permuted a large number of times by randomly assigning the group affiliations to the observations. The statistic computed on the original data is then judged against the permutation distribution to obtain a *p*-value.

5. To assess the association between two sets of variables, a statistic measuring this association is first postulated. Under the null hypothesis of no difference we can randomly connect the first and second sets of data, and doing this many times generates a null distribution of the association measure. The observed association measured on the original data is once again judged against the permutation distribution to obtain a *p*-value.

6. The parts of variance/inertia, or eigenvalues, can also be assessed for statistical significance by generating a null distribution of their percentages of the total, under an hypothesis of no relationship between the variables (usually columns of the data matrix), in which case the values for each variable can be permuted randomly to generate a null distribution of each eigenvalue.

7. We propose a similar procedure for hierarchical cluster analysis, where clusteredness is indicated by low node levels. The data for each variable are permuted randomly and each time the same clustering algorithm performed, generating a permutation distribution for each level under the null hypothesis. Observed node levels that are in the lower tail of these permutation distributions will indicate significant clustering.

CHAPTER 18

Statistical Modelling

As we said in Chapter 2, the principal methodologies of statisticians are functional methods that aim to explain response variables in terms of a set of explanatory variables (also called *predictors),* i.e. the typical regression situation. In this book we have been concentrating on structural methods of particular importance for ecological data analysts – this is mainly because of the large numbers of response variables observed in ecological studies, and the ensuing need for dimension reduction. In this chapter we intend to give a short overview of the sort of functional methods that are in use today when there is one response variable of interest, emphasising that this is a very brief description of a topic that deserves a book by itself. We start with several variants of linear regression, gathered together under the collective title of generalized linear models. These approaches all achieve a mathematical equation that links the mean of the response variable with a linear function of the explanatory variables. We shall also give some glimpses of two alternative nonparametric approaches to modelling: generalized additive modelling, which replaces the linear function of the predictors with a much freer set of functional forms, and classification and regression trees, which take a radically different approach to relating a response to a set of predictors and their interactions, in the form of a decision tree.

Contents

Multiple linear regression

Multiple regression is a model for the *conditional* mean of a response variable y given a set of explanatory variables x_1, x_2, …, x_m. To explain this statement and the assumptions of the regression model, we consider the simple case when there is only one explanatory variable x. Exhibit 18.1 shows an explanatory variable x (which could be depth, for example) that can take on the integer values from 50 to 100, and for each value of x there are many values of y (which could be pollution). The brown line trajectory connects the means of y in every subsample of points corresponding to a given value of x. For each x we can imagine the total population of values of y, and each of these populations has a probability distribution, called a *conditional distribution* because it depends on x. Each of these conditional distributions has a mean and a variance and if we connected the means of all these conditional distributions (as opposed to the sample means that are connected by the brown lines) we would have what is called the *regression* of y on x, denoted by $\mu(x)$. Multiple linear regression has the following assumptions:

1. The regression function (i.e., means of the conditional distributions for all values of x) is linear, in this case a straight line: $\mu(x) = a + bx$.

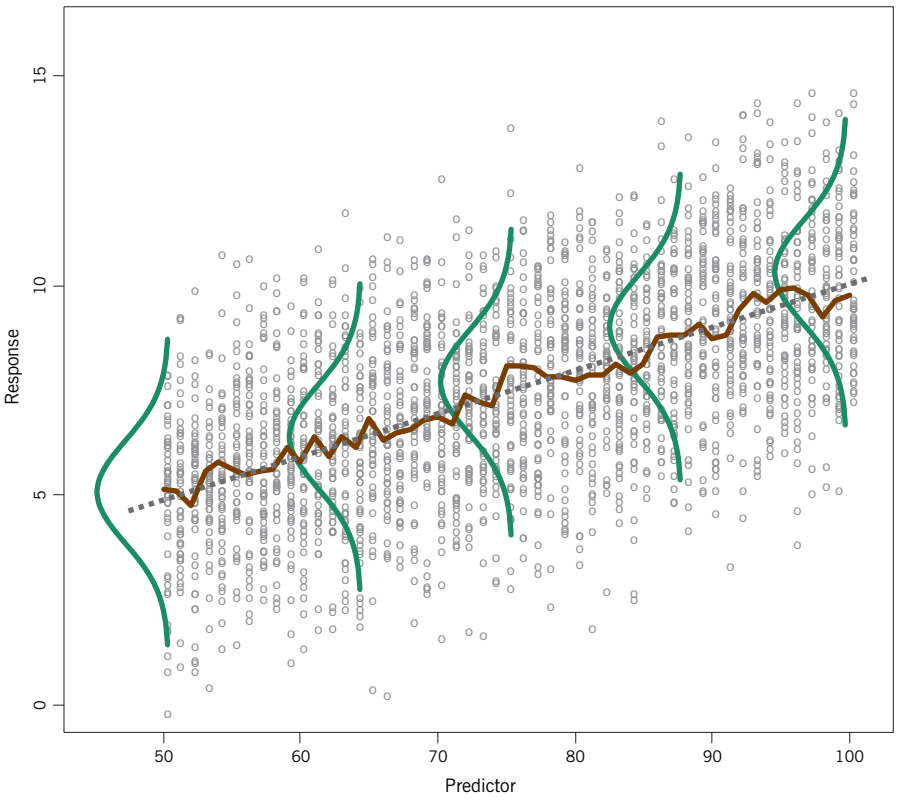


Exhibit 18.1: *Many observations of a response variable y for different integer values of a predictor x. Every set of y values for a given x is a sample conditional distribution, with a mean and variance, and the brown trajectory links the conditional sample means. Multiple linear regression assumes that the data are from normal distributions conditional on x (a few are shown in green), with conditional means modelled as a straight line and with constant variances*

2. The conditional distribution for any x is normal, with mean equal to a linear function of x (i.e., assumption 1), and variance equal to a constant value across all the x values (we say that the variances are *homoscedastic*, as opposed to *heteroscedastic* if the variances change with x).

In Exhibit 18.1 the estimated linear regression function is shown by a dashed line, as well as a few examples of the conditional distributions – since the y values are on the vertical axis the distribution functions are shown on their sides. In this example there would be a conditional distribution for each value of x and the regression line is the assumed model for the means of these distributions, called the *conditional means*.

There are more than 2,000 sample points in Exhibit 18.1 and we would seldom get such a large sample – rather, we would get a sample of size 50, say, as shown in Exhibit 18.2, but the assumptions of linearity and variance homogeneity remain exactly the same. The analysis estimates the linear regression relationship, as shown, which is used firstly for interpreting the relationship between

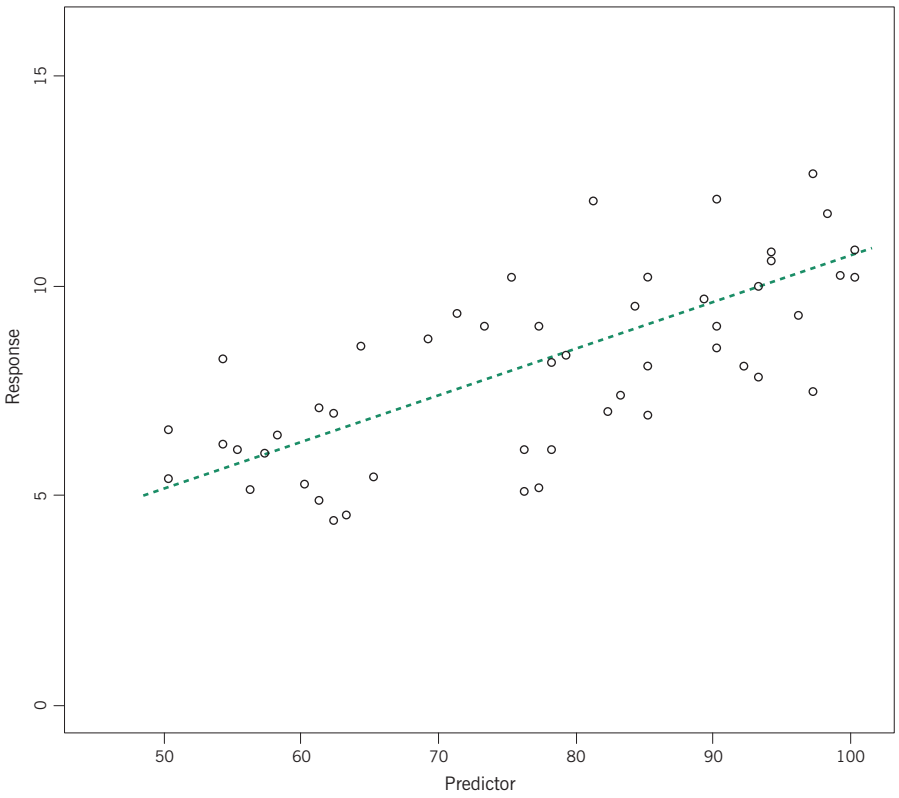


Exhibit 18.2:
A sample of 50 observations of the response y *and predictor* x*, showing the estimated regression line*

response and predictor, and secondly for predicting y for given x. Notice that when a response is predicted from a given value of x, it is the mean of the conditional distribution that is being predicted. In Exhibit 18.2, for example, an x value of 80 predicts a y value of about 7.89. One has then to imagine all the possible y values that constitute the population, possibly infinite, that could be observed for $x = 80$, and then the value 7.89 is the predicted mean of these. Regression analysis includes the technology for setting confidence limits around these predicted means.

Performing the regression on the above sample of 50 observations gives the following regression equation with standard errors (in square brackets) and p-values (in round brackets) associated with each coefficient:

$$\begin{array}{rcccc} \text{mean of } y & = & 0.392 & + & 0.0937x \\ & & [1.587] & & [0.0218] \\ & & (p = 0.81) & & (p < 0.0001) \end{array} \qquad (18.1)$$

The statistical conclusion would be that the constant term (i.e. intercept) is not significantly different from zero, while the predictor (i.e. slope) is highly significant. This agrees with the way these data were simulated: the conditional mean of y was set as the linear function $0.1x$ with no constant term, and the confidence interval for the coefficient of x, based on the estimate and the standard error (where the confidence interval is about 2 standard errors about the mean), does include the true value 0.1.

Poisson regression

The regression model can be generalized to the situation where the responses are of different data types, for example count and categorical variables. This general family of methods is called *generalized linear modelling* (GLM, for short). We first consider the case of a count response, which can be modelled using *Poisson regression*. Exhibit 18.3 shows some count data (for example, abundance counts, where only counts of 0 to 5 were observed here) recorded for different values of the predictor x. Theoretically again, we could have an infinite number of count observations for each x value, and the assumption is that for each x the counts follow the natural distribution for count data, the Poisson distribution, with a mean that depends on x (three examples of conditional distributions are shown in Exhibit 18.3, for $x = 50$, 75 and 100). Because a count variable is considered to be a ratio variable, it is the logarithm of the conditional means of the Poisson distribution that is modelled as a linear function: $\log(\mu(x)) = a + bx$ (see Chapter 3 where we discussed relationships of this type, where an additive change in x would imply a multiplicative change in the mean count). Notice in Exhibit 18.3

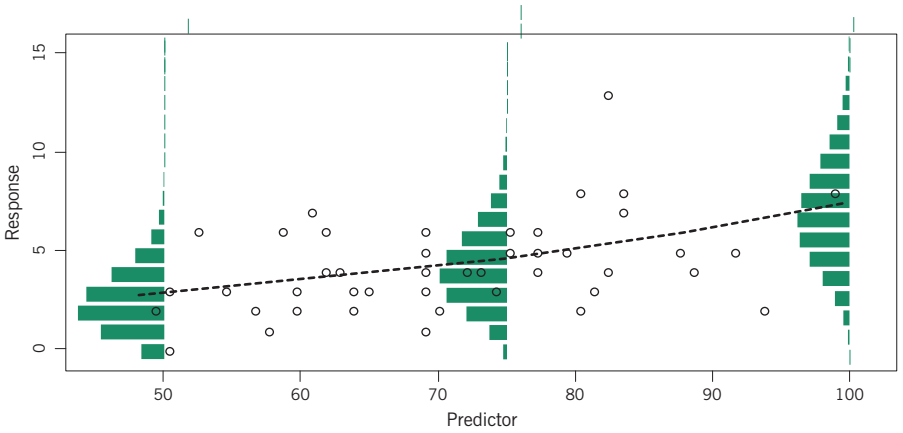


Exhibit 18.3: *Poisson regression, showing some observed response count data, three examples of the conditional distributions for x = 50, 75 and 100, assumed to be Poisson, shown in green, with the dashed line showing the estimated regression relationship of the means, where log mean is modelled as a linear function of the predictor*

that the conditional means are increasing and that the variance of the Poisson distributions is increasing accordingly.

Performing the Poisson regression on the above sample of 50 observations gives the following regression equation with standard errors (in square brackets) and p-values (in round brackets) associated with each coefficient:

$$\begin{array}{rccc} \log(\text{mean of } y) = & 0.0822 & + & 0.0183x \\ & [0.3785] & & [0.0046] \\ & (p = 0.83) & & (p < 0.0001) \end{array} \qquad (18.2)$$

Again, the statistical conclusion would be that the constant term is not significantly different from zero, while the predictor is highly significant. This agrees with the way these data were simulated: the log of the conditional mean of y was set as the linear function $0.02x$ with no constant term, and the confidence interval for the coefficient of x, based on the estimate and the standard error, does include the true value 0.02. The interpretation of the estimated coefficient 0.0183 is that for every unit increment in x, the log mean is increased by 0.0183, that is the mean is multiplied by $\exp(0.0183) = 1.0185$, or an increase of 1.85%. Notice how the slope of the regression curve is increasing (i.e., the curve is convex) due to the multiplicative effect of the predictor.

Logistic regression

As a final example of a generalized linear model, consider a dichotomous response variable, for example presence/absence of a species, and consider observations of this response along with associated predictor values, shown in Exhibit 18.4. Now the

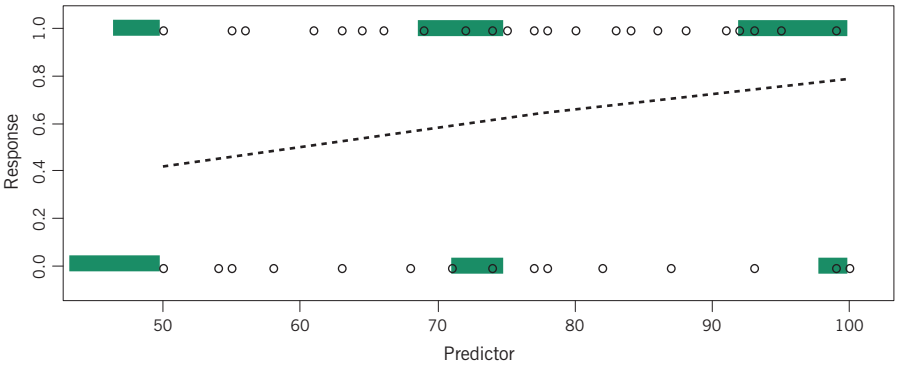


Exhibit 18.4: *Logistic regression, showing some observed dichotomous data, three examples are shown in green of the conditional probabilities of a 1 and a 0, for* x = *50, 75 and 100 (remember again that these probability distributions are shown on their sides, in this case there are only two probabilities for each distribution, the probability of a 0 and the probability of a 1). The dashed line shows the estimated regression of the means, in this case probabilities, where the logits of the probabilities are modelled as linear functions of the predictor*

observations are only 0s (absences) and 1s (presences) and there could be some repeated observations.

Performing the logistic regression on the above sample of 50 observations gives the following regression equation with standard errors (in square brackets) and *p*-values (in round brackets) associated with each coefficient:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \begin{matrix} -3.568 \\ [1.523] \\ (p = 0.019) \end{matrix} + \begin{matrix} 0.0538x \\ [0.0219] \\ (p = 0.014) \end{matrix} \qquad (18.3)$$

The statistical conclusion would be that both the constant term and the predictor are significant. This agrees with the way these data were simulated: the conditional mean of y was set as the linear function $-2 + (1/30)x$, where $1/30 = 0.0333$, and the confidence intervals for both coefficients do include the true values, but do not include 0.

Explained variance, deviance and AIC

The measure of explained variance in linear regression is well-known and we have used the concept many times in other contexts as well, for example the variance (or inertia) explained by the dimensions of a PCA, LRA, CA or CCA solution. *Deviance* is the generalization of this concept when it comes to GLMs. Without defining deviance mathematically, it functions in the same way: first, there is the concepts of the *full* (or *saturated*) *model*, where the response is fitted perfectly, and the *null model*, where no explanatory variables are fitted at all, with just the constant term being estimated. This difference is used as a measure of total variance and is, in fact, equal to the total variance in the case of linear regression. Deviance is used to measure the difference between models (i.e. hypotheses) in

their ability to account for variation in the response variable. For example, if an explanatory variable is introduced in the GLM, deviance measures the amount of variance accounted for by this new variable compared to the null model. If a second variable is then introduced, one can compute the deviance between the one-variable model and the two-variable model to ascertain if it is worth introducing the second variable. Deviance has an approximate chi-square distribution, with degrees of freedom depending on the difference in numbers of parameters between two models. Let us take the logistic regression just performed above as an example. The *null deviance* for a model with no variables turns out to be equal to 68.99, and the *residual deviance*[1] for a model with the single predictor x is equal to 61.81.The deviance is thus the difference 68.99 − 61.81 = 7.18, which is from a chi-square distribution with 1 degree of freedom (the model has one additional parameter), for which the p-value is 0.007. Note that this p-value is not exactly the same as the one ($p = 0.014$) reported for the coefficient in (18.3), which was computed using the normal distribution.

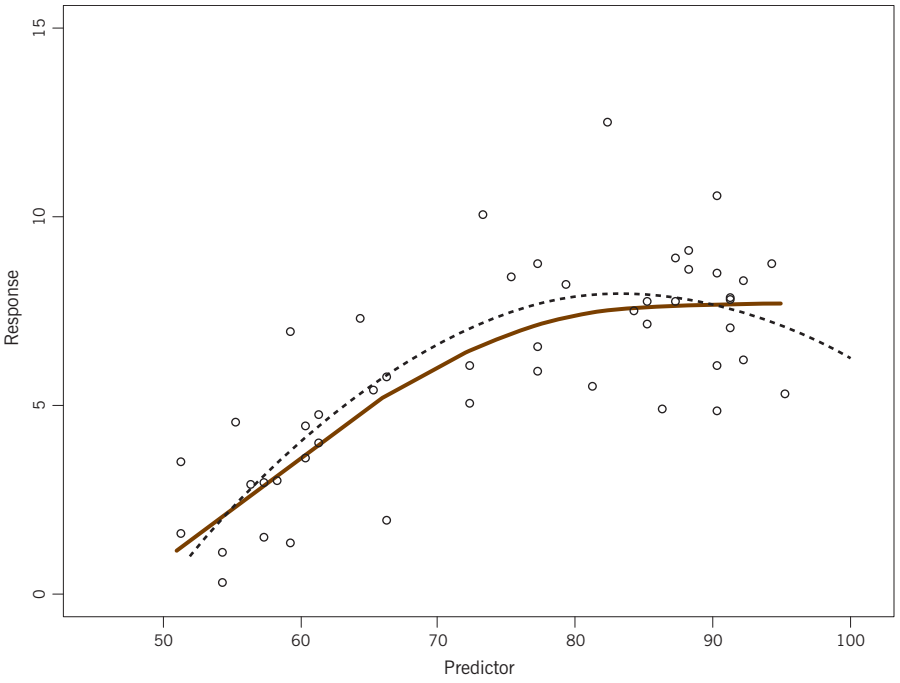
When comparing the performance of alternative models (i.e. testing alternative hypotheses) it is important to consider both their goodness of fit (e.g., analysis of deviance) and complexity, that is the number of explanatory variables (and associated parameters) included in each model. Akaike's information criterion (AIC) is a way of comparing models with different numbers of parameters that combines goodness of fit and complexity considerations. Since adding explanatory variables to a model, whether they are relevant or not, will always explain more variance, i.e. reduce the residual deviance, the AIC criterion adds a penalty for additional parameters equal to two times the number of parameters. Here the constant term is included, hence in our logistic regression model (18.3) the AIC is equal to the residual deviance plus 4 (2 times the number of parameters), i.e. 61.81 + 4 = 65.81. Models with the lowest AIC are to be preferred. For example, if we added another explanatory variable to the model and reduced the residual deviance to 60.50, say, then the AIC would be 60.50 + 2 × 3 = 66.50, and the first model with one parameter is preferable because it has lower AIC.

Nonlinear models

In the above examples, although the link function that transforms the mean of the response changes according to the response variable type, the way the predictors are combined is always a linear function, which is quite a simplistic assumption. Transformations can also be made of the predictors to accommodate nonlinear relationships. For example, Exhibit 18.5 shows another set of observations, and a scatterplot smoother has been added (in brown) indicating the possibility

[1] This is the way the R function `glm` reports the deviance values.

Exhibit 18.5: *A scatterplot of a continuous response* y *and a predictor* x*, showing a scatterplot smoother (brown line) which suggests a nonlinear relationship. The estimated quadratic relationship is shown by a dashed curve*



that the relationship is not linear, but reaching a peak or possibly an asymptote. In order to take into account the curvature, an additional predictor term in x^2 can be included so that a quadratic is postulated as the regression function (polynomial regression), and the result is as follows:

$$\begin{array}{lccccccc} \text{mean of } y = & -39.91 & + & 1.139x & - & 0.0068x^2 & & (18.4) \\ & [9.09] & & [0.258] & & [0.0018] & & \\ & (p < 0.0001) & & (p < 0.0001) & & (p = 0.0004) & & \end{array}$$

All terms are significant and the confidence intervals for the coefficients all contain the true values used in the simulated formula, which is $-32 + 0.9x - 0.005x^2$. The estimated regression function in (18.4) is shown with a dashed line.

Since we have introduced fuzzy coding, it is interesting to compare the results using this alternative. Fuzzy coding of the predictor variable with three fuzzy categories allows for a curve with one turning point, which is what we need, so three categories were created, x_1, x_2, x_3, and the following regression function resulted:

$$\begin{array}{rccccc} \text{mean of } y = & 1.03x_1 & + & 7.95x_2 & + & 7.56x_3 \\ & [0.59] & & [0.62] & & [0.61] \\ & (p = 0.08) & & (p < 0.0001) & & (p < 0.0001) \end{array} \qquad (18.5)$$

(notice that we show the result without the constant for the three fuzzy dummies). The variance explained is 65.6%, only slightly less than the 66.0% for the quadratic; both have three coefficients and thus the same degrees of freedom in the regression. Exhibit 18.6 shows that with the fuzzy coding the relationship is estimated as two linear functions, because of the triangular membership functions used to create the fuzzy categories. To capture a curved relationship we should use different membership functions, of which there are many possibilities, for example Gaussian (i.e., normal) membership functions. If interest is not only in diagnosing a smooth relationship (like the scatterplot smoother visualizes) but also in testing it statistically, then the section on generalized additive models later in this chapter provides a solution.

Multiple predictors and interactions

Most often there are many explanatory variables, therefore we need a strategy to decide which are significant predictors of the response, and whether they interact, so as to choose the best model (hypothesis) given the data. As an illustration

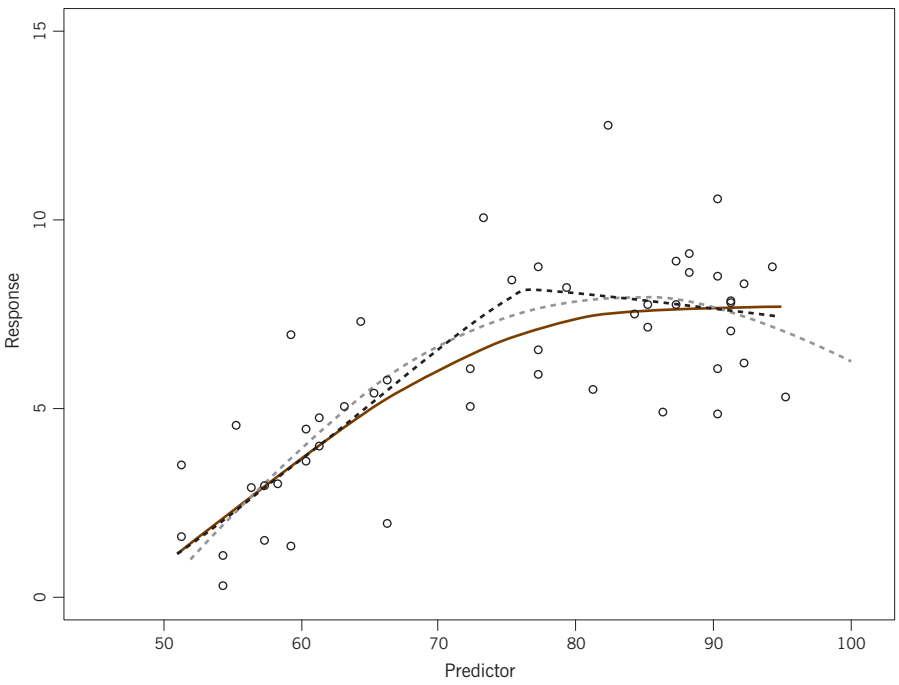


Exhibit 18.6: *Same data as in Exhibit 18.5, with the estimated quadratic relationship in gray, and the relationship according to (18.5) shown by black dashed lines*

we return to the "Barents fish" data set studied in Chapters 1, 13 and 15. To convert the abundances of the 30 fish species, effectively a 30-variable response, to a single response, we computed the Shannon-Weaver diversity H' for each of the 89 stations, using the formula:

$$H' = -\sum_j p_j \log(p_j)$$

where p_j is the proportional abundance of species j, $j = 1,\ldots 30$. This well-known diversity index is at its lowest value, 0, when there is only a single species observed in the sample, and reaches its maximum if all species are contained in the same proportion. Amongst possible explanatory variables of diversity we have the bottom depth and temperature at the station trawled. If depth and temperature are entered into a linear regression, then the estimated effects of each variable do not depend on the other variable. In this example the effects of depth and temperature in a model as separate linear terms are not statistically significant. Significant results are found, however, when the interaction term is included, i.e. the product of the two variables, which allows for the possibility that the relationship with temperature depends on the depth and vice versa. The regression, explaining only 6.6% of the variance, but still significant, is:

$$\begin{array}{lccccc} \text{mean } H' = & 0.466 & + 0.00243\,\text{depth} & + 0.493\,\text{temp.} & - 0.00152\,\text{depth} \times \text{temp.} & (18.6) \\ & [0.470] & [0.00152] & [0.218] & [0.00072] & \\ & (p = 0.32) & (p = 0.11) & (p = 0.026) & (p = 0.038) & \end{array}$$

The interaction term implies that the relationship with depth varies according to the temperature – notice that we would retain the linear term in depth even though it is insignificant, because the interaction term which involves depth is significant. Exhibit 18.7 shows the linear relationships with depth for three different temperatures that are chosen in the temperature range of the observed data.

Generalized additive models

Generalized additive models (GAM for short) are a very flexible framework for taking care of nonlinearities in the data. The approach is more complex but the benefits are great. Without entering too much into technicalities, we show the equivalent GAM analysis used to estimate the regression of diversity as a function of depth and temperature, in the previous example. If we enter depth and temperature as separate variables, the GAM results show that depth is significant with a clear nonlinear relationship ($p < 0.0001$) but not temperature ($p = 0.25$) – see Exhibit 18.8. In a GAM model the form of the relationship

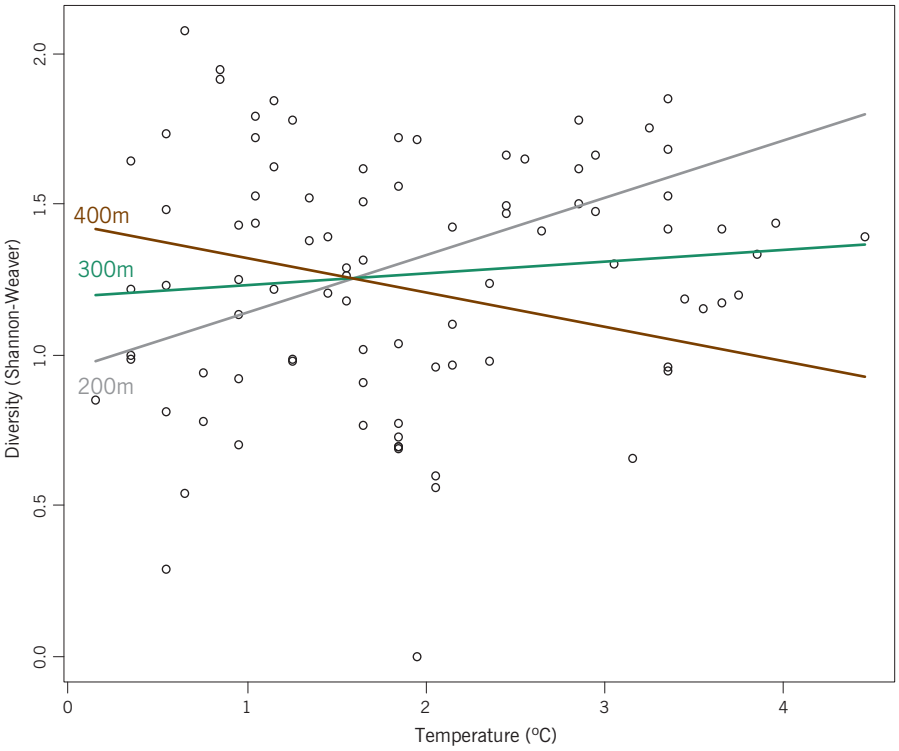


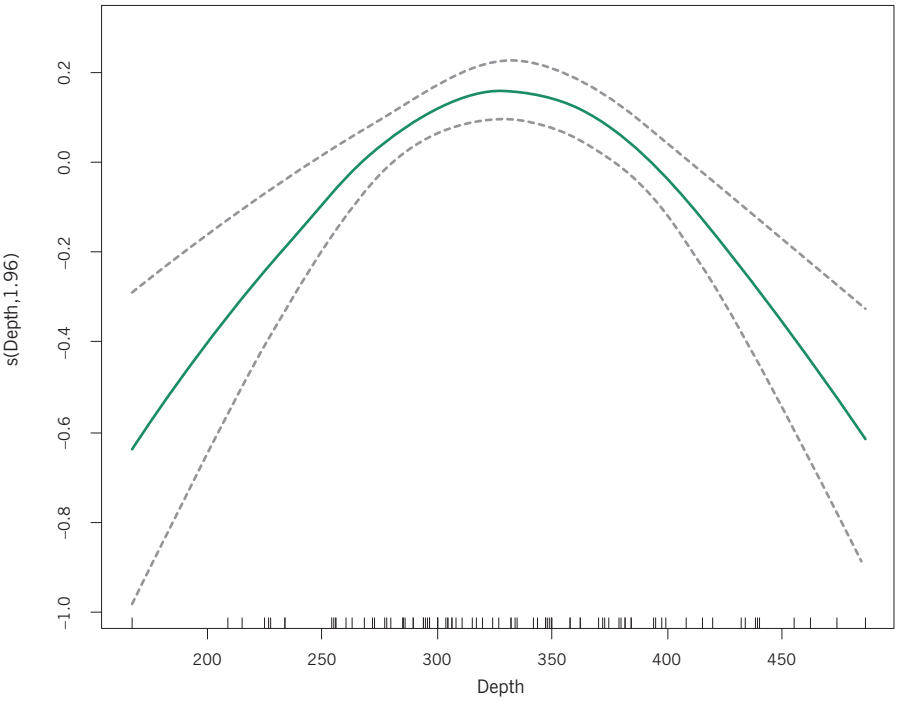
Exhibit 18.7: *Linear regression relationships (18.6) between diversity and temperature that depend on the depth (illustrated for three values of depth), showing an interaction effect (regression lines with different slopes). If the interaction effect were absent, the three lines would be parallel and the effect of temperature would be the same (i.e., not contingent on depth)*

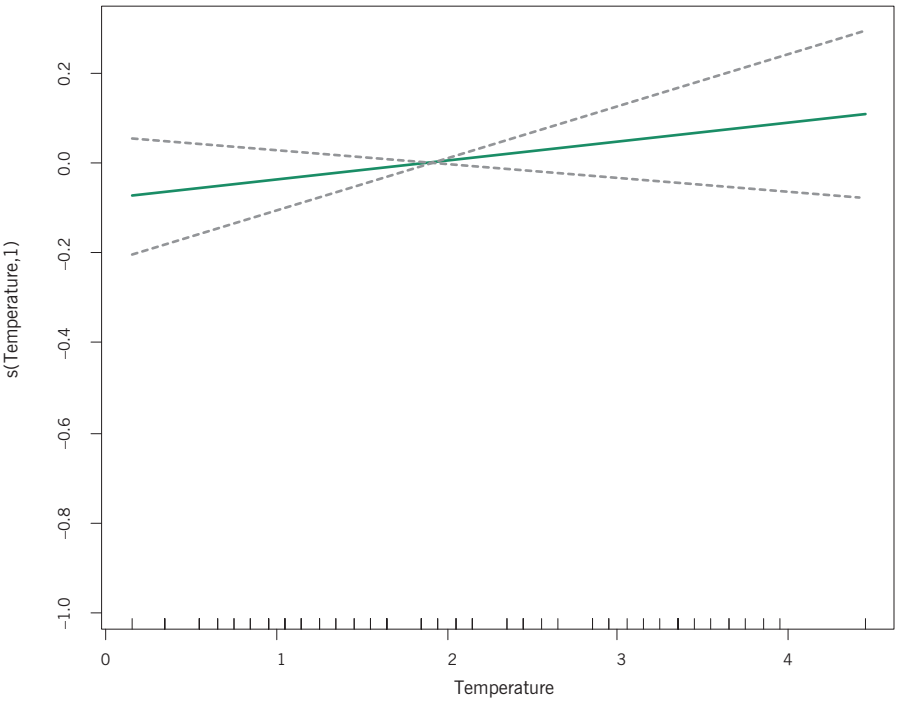
is diagnosed for the researcher, and the number of degrees of freedom of the relationship is estimated: 1.96 (say 2) for depth, and 1 for temperature. There is no mathematical model for the regression here, so we cannot write down a formula as before. But, since the depth relationship looks quadratic, we could try adding a quadratic term to the model, and return to using conventional regression:

$$\begin{aligned} \text{mean } H' = {} & -2.22 + 2.15\ \text{depth} - 0.0000326\ \text{depth}^2 + 0.0463\ \text{temp.} \qquad (18.7) \\ & [0.80] \quad [0.0048] \quad [0.0000071] \quad [0.0374] \\ & (p = 0.007)\ (p < 0.0001) \quad (p < 0.0001) \quad (p = 0.22) \end{aligned}$$

To choose between models (18.6) and (18.7) we can compare the AIC in each case: 94.9 for (18.6) and 79.7 for (18.7). The difference in AIC between the parametric model in (18.7) and the GAM model summarized in Exhibit 18.8, which has an AIC of 79.5, is tiny. Model (18.7) could thus be further improved by dropping the insignificant temperature term:

Exhibit 18.8: *Generalized additive modelling of diversity as smooth functions of depth and temperature: depth is diagnosed as having a significant quadratic relationship, while the slightly increasing linear relationship with temperature is non-significant. Both plots are centred vertically at mean diversity, so show estimated deviations from the mean. Confidence regions for the estimated relationships are also shown*





$$\begin{array}{lccccc} \text{mean } H' = & -2.11 & + & 2.13 \text{ depth} & - & 0.0000324 \text{ depth}^2 \\ & [0.80] & & [0.0048] & & [0.0000071] \\ & (p = 0.01) & & (p < 0.0001) & & (p < 0.0001) \end{array} \qquad (18.8)$$

in which case the AIC is 79.3, and explained variance is 19.9%.

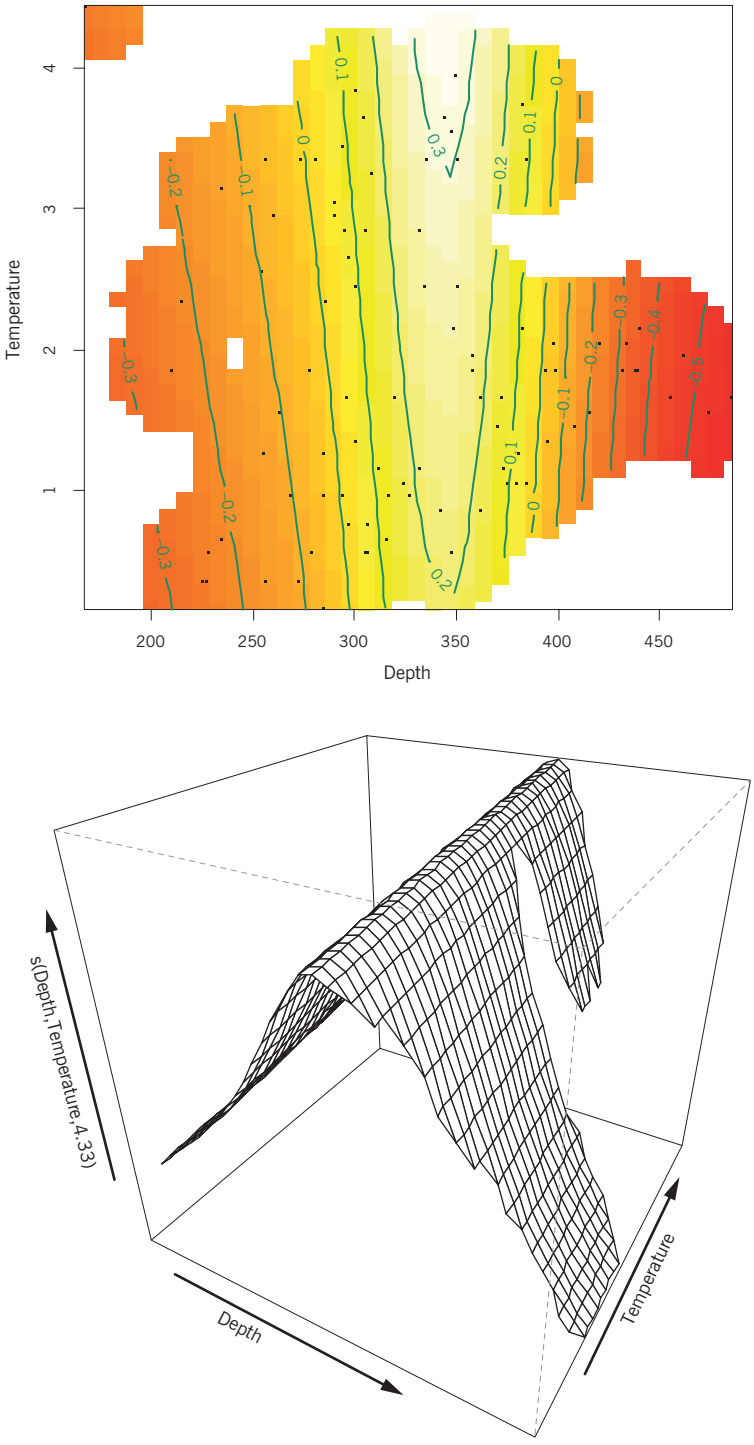
As a final illustration of the power of GAMs, we can make a model with a smooth interaction of depth and temperature. This has an even lower AIC value 70.7, and now to visualize the diagnosed relationship requires making a contour plot of the model values in the space of the depth and temperature variables, or making a perspective plot in three dimensions – see Exhibit 18.9.To test whether the interaction is significant we can compare the residual deviances for the model shown in Exhibit 18.8 (85.04) and the one in Exhibit 18.9 (83.67), i.e. a difference of only 1.37 units of deviance, which is not significant.[2] All these results and considerations lead us to the conclusion that the parametric model (18.8) with depth modelled as a quadratic is the one of choice – it has few parameters, is a function that can be easily interpreted and computed and does almost as well as several competing models that are more complex. Here we have demonstrated how GAM can help to suggest a nonlinear model for a regression. We will return to GAM modelling in Chapter 20 where we show that it is a convenient and flexible approach for taking into account the effect of spatial position.

Classification trees

We close this chapter on statistical modelling by showing a completely different approach to modelling a continuous or categorical response variable, by constructing a type of decision tree with the goal of predicting the continuous response variable (regression trees) or categorical response category (classification trees). We consider the latter case first, and take as example the presence/absence of polar cod *(Boreogadus saida)* in a sample. In the data matrix there are 21 samples with polar cod and 68 without, so the response data consist of 21 ones and 68 zeros. Applying a classification tree algorithm, with two predictors, depth and temperature, produces the tree model of Exhibit 18.10. The 89 samples are notionally fed down the tree and are split by the decisions at each branch, where each decision indicates the subsample that goes to the left hand side. For example, samples going to the left at the top of the tree satisfy the condition

[2] Here we have not entered into the aspect of the degrees of freedom for this comparison of GAM models, nor how *p*-values are computed. In GAM the degrees of freedom are not integers, but estimates on a continuous scale. Hence, comparing models leads to differences in degrees of freedom that are also not whole numbers – in this particular case the degrees of freedom associated with the deviance difference of 1.37 are 1.01, close enough to 1 for all practical purposes.

Exhibit 18.9: *Contour plot (upper) and perspective plot (down) of the diagnosed interaction regression surface of depth and temperature, predicting the deviations from mean diversity. The concave relationship with depth is clearly seen as well as the slight relationship with depth. The difference between the model with or without interactions is, however, not significant*



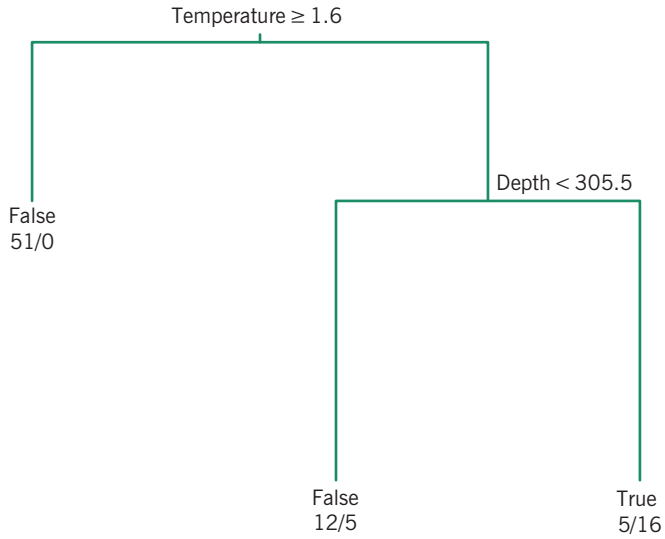


Exhibit 18.10: *Classification tree model for predicting the presence of polar cod. The one branch which predicts their presence gives the rule: temperature < 1.6°C and depth ≥ 306 m. This rule correctly predicts the presence of polar cod in 16 samples but misclassifies 5 samples as having polar cod when they do not*

that temperature is greater than or equal to 1.6°C, while the others for which temperature is less than 1.6°C go to the right. Of the 89 samples, 51 go to the left, and all of them have no polar cod, so the prediction is False (i.e., no polar cod). The remaining 38 samples that go to the right are optimally split into two groups according to depth, 305 m or less to the left, and 306 m or more to the right. Of 38 samples, 17 go to the left and of these 12 have no polar cod, so False is predicted, while 21 go to the right, and a majority has polar cod so polar cod is predicted (True). The final branches of the tree, where the final predictions are made, are called *terminal nodes*, and the objective is to make them as concentrated as possible into one category.

The beauty of this approach is that it copes with interactions in a natural way by looking for combinations of characteristics that explain the response, in this case the combination of lower temperature (lower than 1.6°C) and higher depths (greater than or equal to 306 m) is a prediction rule for polar cod, otherwise no polar cod are predicted.

As a comparison, let us perform a logistic regression predicting polar cod, using depth and temperature. Both variables are significant predictors but result in only 12 correct predictions of polar cod presence. The misclassification tables for the two approaches are given in Exhibit 18.11.

Regression trees

The same style of tree model can be constructed for a continuous response. In this case the idea is to arrive at terminal nodes with standard deviations (or

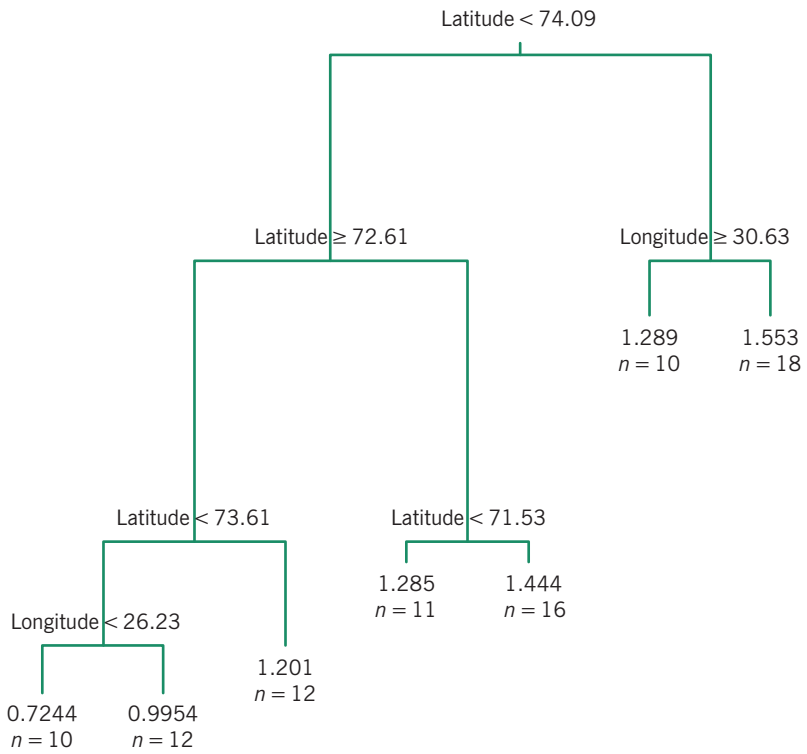
Exhibit 18.11: *Comparison of misclassification rates for the classification tree of Exhibit 18.10, compared to that for logistic regression, using the same predictors. The classification tree correctly predicts presence and absence in 79 of the 89 samples, while logistic regression correctly predicts 74*

		Classification tree		Logistic regression	
		Truth		*Truth*	
		Polar cod	No polar cod	Polar cod	No polar cod
Predicted	Polar cod	16	5	12	6
	No polar cod	5	63	9	62

any other appropriate measure of variability for the response) as low as possible. As an example, we return to the diversity response, this time choosing time latitude and longitude coordinates as the predictors in order to classify the samples into regions of homogeneous diversity. The result is given in Exhibit 18.12.

The regression tree partitions the sampling area and can be drawn on the map in Exhibit 18.13. The most diverse area is in the north-west, while the least diverse is in the central western block.

Exhibit 18.12: *Regression tree predicting fish diversity from latitude and longitude of sample positions. The terminal nodes give the average diversity of the samples that fall into them. This tree yields the spatial classification of the sampling region given in Exhibit 18.13*



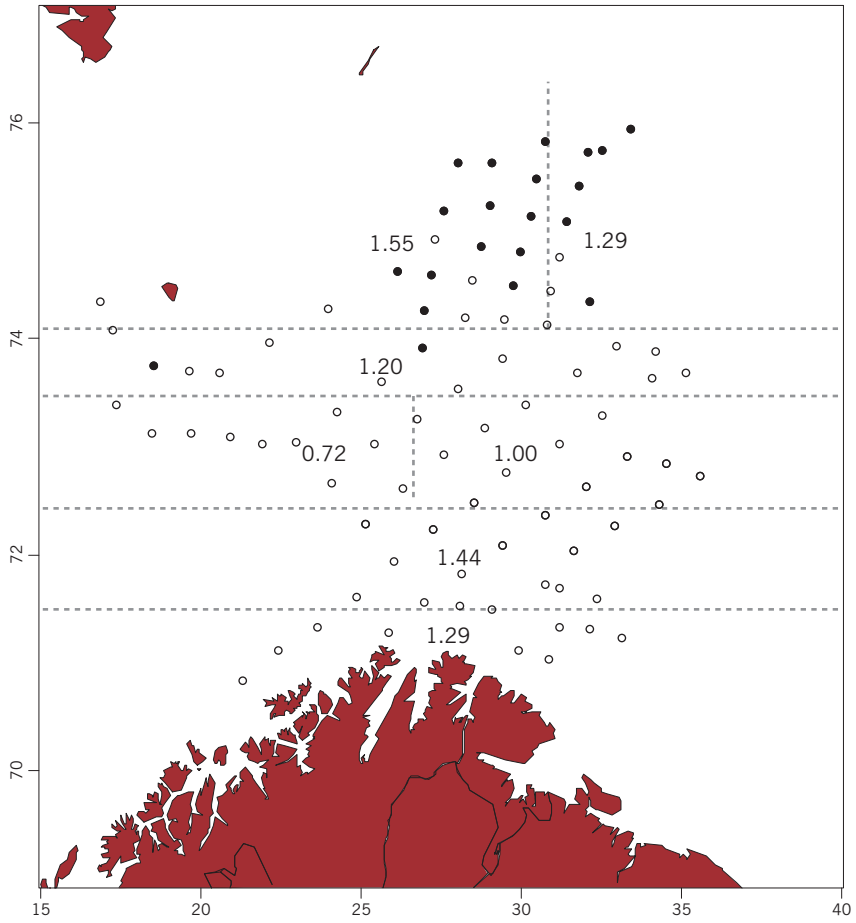


Exhibit 18.13: *Map of Barents Sea showing the locations of the 89 sampling sites (see Exhibit 11.1) and the slicing up of the region according to the regression tree of Exhibit 18.12, and the average fish diversities in each block. Most of the slices divide the latitudes north to south, with just two east-west divisions of the longitudes. Dark locations show the 21 sites where polar cod were found*

SUMMARY: Statistical modelling

1. The family of generalized linear models (GLMs) includes multiple linear regression, Poisson regression, and logistic regression, when the response variable is continuous, count or categorical, respectively, for which the assumed conditional distributions given a set of explanatory variables (or predictors), are normal, Poisson and binomial respectively.

2. Each of these models assumes that a transformation of the mean is a linear function of the explanatory variables. This transformation is called the *link function*. In multiple regression there is no transformation, and the link is thus the identity. In Poisson regression the link is the logarithm, and in logistic regression it is the logit function, or log-odds.

3. To take into account nonlinearities, polynomial functions of the explanatory variables or fuzzy coding into several categories can be used.

4. Generalized additive models (GAMs) are even more general than GLMs, allowing considerable flexibility in the form of the relationship of the response with the explanatory variables.

5. Both GLM and GAM environments allow interaction effects to be included and tested.

6. Classification and regression trees are an alternative that specifically look at the interaction structure of the predictors and come up with combinations of intervals that predict either categorical or continuous responses with minimum error.

TWO CASE STUDIES

CHAPTER 19

Case Study 1: Temporal Trends and Spatial Patterns across a Large Ecological Data Set

The examples presented in previous chapters have generally been on small- to medium-sized data sets that are good for teaching and understanding the basic concepts of the methodologies used. We conclude with two chapters detailing much larger studies that take full advantage of multivariate analysis to synthesize complex phenomena in a format that is easier to interpret and come to substantive conclusions. The two chapters treat the same set of data, a large set of samples of fish species in the Barents Sea over a six-year period, where the spatial location of each sample is known as well as additional environmental variables such as depth and water temperature. In the present chapter we shall study the temporal trends and spatial patterns of the fish compositions and also try to account for these patterns in terms of the environmental variables. But before applying multivariate analysis to data across time and space, we have to consider carefully the areal sampling across the years and reweight the observations to eliminate sampling bias.

Contents

Sampling bias

In this chapter we shall be considering samples in different regions over time over an area of interest. An important consideration is whether data have been collected in a balanced way over time in each region. This is important if one wants to summarize the data over the whole area and make temporal comparisons. If

sampling is more intense in some regions in some years and less intense in other years, this can lead to what is called *sampling bias*. Consider in Exhibit 19.1(a) the hypothetical layout of numbers of samples taken over an area divided into three regions, for three consecutive years.

Let us assume for the moment that the total number of samples over the three years is representative of the sizes (or some measure of importance in the study) of the three regions surveyed, that is region 2 is the largest, followed by region 1 and then region 3. Then, for the sampling to be balanced over the years the numbers of samples should follow the proportions 0.300, 0.433 and 0.267, as seen in the last line of Exhibit 19.1(a). Computing expected proportions for each year, in exactly the same way as one computes expected frequencies in a chi-square test, the table in Exhibit 19.1(b) is obtained. If this latter table of expected frequencies is now divided, cell by cell, by the former table of actual frequencies, a table of *weights* is obtained in Exhibit 19.1(c), reflecting the imbalances.

In Exhibit 19.1(c) the column of ones for region 1 shows that the sampling was in perfect proportion to the expected number. In contrast, region 2 is under-

Exhibit 19.1: *(a) Actual number of sample taken in three regions over a three-year period, with overall proportions of samples in each region over the whole period. (b) Expected number of samples if in each year sampling had taken place in accordance with the overall proportions. (c) The weights computed by dividing the values in table (b) by those in table (a)*

(a)

	Region 1	*Region 2*	*Region 3*	*Sum*
Year 1	30	20	50	*100*
Year 2	15	30	5	*50*
Year 3	45	80	25	*150*
All years	90	130	80	*300*
Prop'n	0.30	0.433	0.267	

(b)

	Region 1	*Region 2*	*Region 3*	*Sum*
Year 1	30	43.3	26.7	*100*
Year 2	15	21.7	13.3	*50*
Year 3	45	65	25	*150*
All years	90	130	80	*300*
Prop'n	0.30	0.433	0.267	

(c)

	Region 1	*Region 2*	*Region 3*
Year 1	1	2.167	0.533
Year 2	1	0.722	2.667
Year 3	1	0.813	1.600

sampled in year 1 and over-sampled in years 2 and 3. In year 1 this region has only 20 samples whereas the expected proportion of 43.3% of 100 is 43.3. The weight of 2.167 is then used to scale up the abundances of observed species in this region. In year 2 the 30 actual samples represent an over-sampling compared to the expected value of 21.7, and these 20 samples are thus down-scaled by a factor of 0.722, and so on. Each of the 300 samples thus receives a weight in Exhibit 19.1(c) according to its year and region, some up-weighted, others down-weighted – notice that the sum of the weights allocated to the 300 samples is equal to 300.

This reweighting is not necessary if the regions are studied one by one, for example average abundances or measures of diversity can be compared within a region using the original unweighted data. However, whenever the regions are put together to estimate a value over the whole area, the weighting will be necessary. Consider, for example, if in region 3 a certain species where particularly abundant. Since this region is heavily sampled in year 1, almost twice as much compared to the expected proportion, the unweighted data in year 1 could show a difference with the other years which is due to this oversampling. Of course, we are assuming that the proportions in the last row of Exhibit 19.1(a) reflect the "population" proportions, but these can be determined by an external criterion such as the area of each region.

Data set "Barents fish trends"

In Chapter 11 the data set "Barents fish" was introduced, a relatively small data set of fish abundances of 30 species at 89 sites in the Barents Sea, during a sampling period in 1997 (Exhibit 11.2). The geographical location was handled in different ways, first by defining a spatial grouping of the samples (Exhibit 11.1), second using latitude and longitude as continuous variables (Exhibits 11.5 and 11.6) and third by defining fuzzy positions with respect to eight compass points and a central category (Exhibits 11.8 and 11.9). In this case study we extend the data set to six consecutive years of data, from 1999 to 2004, called *Barents fish trends*, thus introducing a temporal component into the study. A total of 600 samples are included. We will implement a reweighting scheme in this application, explaining how the previous argument for "crisp" regions can be extended quite naturally to our fuzzy coding of the spatial positions.

Reweighting samples for fuzzy coded data

We are going to use fuzzy coding again to code the geographical position of each sample, as described at the end of Chapter 11. If each of the 600 samples had been allocated "crisply" to one of 9, say, regions, then we would proceed as just explained by counting how many samples were in each region in each year to check if proportionally the same number of stations were sampled from year to year. The situation is hardly different for the fuzzy coding, thankfully, since we can sum the fuzzy values and not the zero-one dummy variables for the region

categories. Exhibit 19.2 shows the sums for each year and for the whole period along with the overall proportions for each region.

To balance the allocation to each region each should follow the overall proportions, that is SW should have 2.5% of the 88 samples, i.e. 2.20 (so it is slightly over-represented, since the actual value is 2.33, W should have 15.5% of 88, i.e. 13.64 (again under-represented, actual value is 11.71), and so on. If we continue computing the expected values and comparing them with the observed ones in Exhibit 19.2, the ratios expected/observed give a matrix of weighting factors in Exhibit 19.3.

Exhibit 19.2: *Sums of fuzzy-coded regional categories for each year and for all years. Columns are the eight compass points and a central region (C)*

	SW	W	NW	S	C	N	SE	E	NE	*Sum*
1999	2.33	11.71	1.16	10.12	31.25	12.57	3.57	11.29	4.00	*88*
2000	4.60	18.86	1.47	14.32	39.27	11.65	2.47	11.43	2.93	*107*
2001	0.64	11.14	1.15	6.60	31.93	12.73	3.06	12.18	4.57	*84*
2002	2.46	15.83	1.41	11.83	37.44	12.50	2.22	11.01	4.30	*99*
2003	2.02	16.44	1.36	14.93	38.55	6.63	6.28	12.04	1.75	*100*
2004	2.72	18.87	1.61	15.46	45.42	14.56	4.17	14.37	4.80	*122*
All years	14.76	92.85	8.17	73.24	223.85	70.64	21.79	72.33	22.36	*600*
Prop'n	0.025	0.155	0.014	0.122	0.373	0.118	0.036	0.121	0.037	

Exhibit 19.3: *Weights for data according to year and fuzzy region*

	SW	W	NW	S	C	N	SE	E	NE
1999	0.929	1.163	1.033	1.062	1.051	0.824	0.895	0.940	0.820
2000	0.572	0.878	0.991	0.912	1.017	1.081	1.573	1.128	1.361
2001	3.229	1.167	0.994	1.554	0.981	0.777	0.997	0.831	0.685
2002	0.990	0.968	0.956	1.022	0.987	0.932	1.619	1.084	0.858
2003	1.218	0.941	1.001	0.818	0.968	1.776	0.578	1.001	2.130
2004	1.104	1.001	1.032	0.963	1.002	0.987	1.062	1.023	0.947

So in category SW (south-west) 2000's samples must be downweighted by a factor of 0.572, whereas 2001's samples must be upweighted by 3.229.

Since the samples do not fall strictly into a region, how can these weights be applied? For example a particular sample in 2000 is coded spatially as follows:

SW	W	NW	S	C	N	SE	E	NE
0.125	0.862	0.000	0.002	0.011	0.000	0.000	0.000	0.000

(19.1)

i.e., it is mostly in the western section, but a bit towards south-west, and quite far from the centre (remember that you can find the exact position of this station from the fuzzy coding). Now each value that we observe in this sample, for example an abundance value of 19 for the species *Sebastes mentella* (*Se_me,* beaked redfish) is split between the fuzzy categories in the above proportions, after which the weights for the year 2000 in Exhibit 19.2 are applied. This means that we can compute a sample-specific weight as the weighted average of the weighting factors:[1]

$$0.125 \times 0.572 + 0.862 \times 0.878 + 0.002 \times 0.912 + 0.011 \times 1.017 = 0.8413 \quad (19.2)$$

Hence, all the abundance data for this sample are downscaled by the factor 0.8413, e.g., for the *Se_me* value of 19, $0.8413 \times 19 = 15.99$. Weights for all the samples are computed in the same way, and the sum of these sample weights is equal to the sample size, 600 in this case. The weights can be used to adjust the abundances as well as in computing regression relationships, using weighted regression, or in computing overall measures over the whole study area such as means and diversity measures, for example, which are then appropriately re-weighted to compensate for sampling biases in different areas.

In most applications such as this one, where the sampling is not drastically out of proportion from year to year, it is not going to make a big difference to the results of a multivariate analysis whether one uses the original abundance matrix or the reweighted one – nevertheless, reweighting is an insurance against possible sampling bias. The negative side of this approach, however, is that in a severely under-sampled region such as the south-western region in 2001 (Exhibit 19.2) there might be some unusual samples that then become up-weighted and thus over-emphasize the species (or lack of species) in that region, so we should still have a certain minimum sample size in each area and each year to avoid estimation bias. In what follows, we will consistently use the reweighted data set and can report in passing that the results are very similar when compared to those of the unweighted data.

Correspondence analysis of reweighted data

Exhibit 19.4 shows the CA of the abundance matrix, first the samples (gray circles) and species (brown abbreviated labels) and then an enlargement of the central area showing the centroids of the year points and all the fuzzy categorical variables. Six species contribute more than average to the axes, shown with bigger labels. The first CA axis separates species dominating in cold Arctic waters from species found

[1] There are three different weights here: (1) the fuzzy coded values in (19.1) that add up to 1, which will be used as weights in the weighted averaging (only four of them are nonzero); (2) the fuzzy-region-specific weighting factors in Exhibit 19.2; and (3) the final value of 0.8413 which is a weight to apply to the abundances of this particular sample.

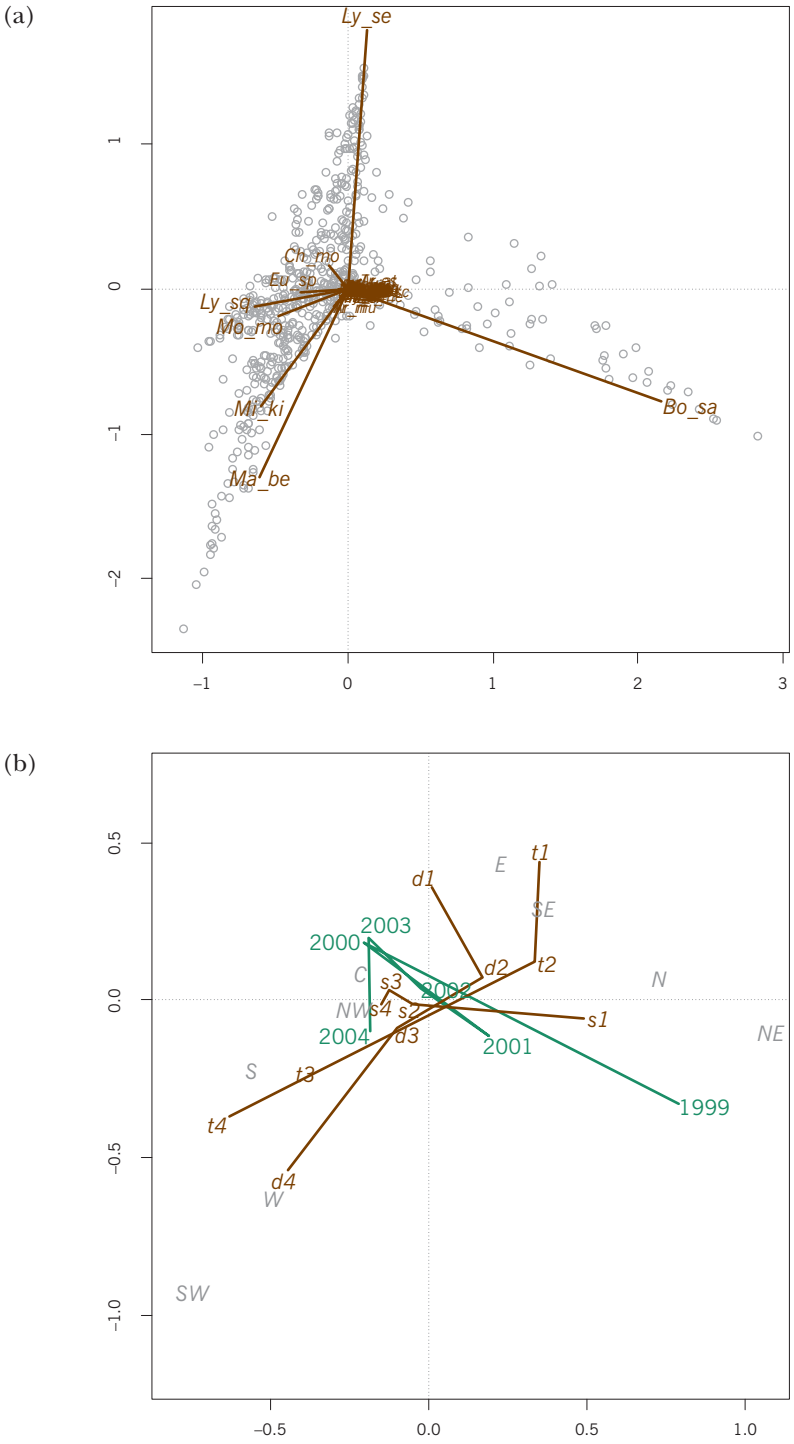


Exhibit 19.4: *Correspondence analysis contribution biplot of the "Barents fish trends" data set. The upper plot shows the active data, the samples and species (high-contributing species are shown with bigger labels).The lower plot shows the centroids of all the categories, linking together categories of ordinal variables. 32.6% of the total inertia of 4.017 is explained by these two first dimensions*

in warmer Atlantic waters, in other words a latitudinal effect. The second axis separates species in shallower waters from those found in deeper waters.

Canonical correspondence analysis of reweighted data

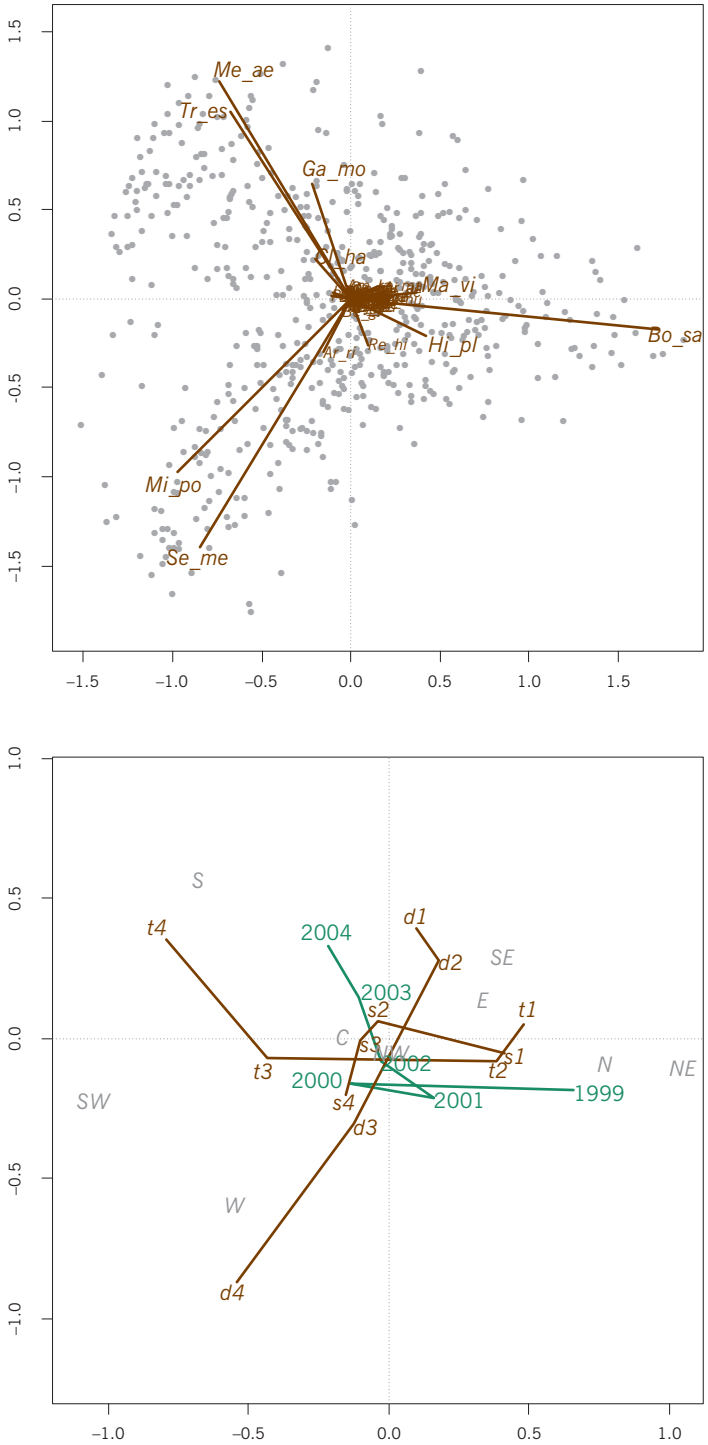
The CA will show the main dimensions of the 600 samples in the abundance matrix without specific reference to the interannual differences and differences across temperature, depth, slope and in space. The CCA will look at the dimensions of abundance that are in the space of the all these explanatory variables. Each set of dummy or fuzzy variables contributes one less than its number of categories to the dimensionality of the restricted space in which CCA operates: time (6 – 1 = 5), space (9 – 1 = 8) and depth, temperature and slope (4 – 1 = 3, each), totalling 22. This 22-dimensional restricted space contains 37.7% of the total inertia of the data, in other words there is 62.3% of the inertia that is unrelated to the explanatory variables. Exhibit 19.5 shows the result in the same format as Exhibit 19.4. If the scale of the lower centroid plot is compared to that of the centroid plot in Exhibit 19.4, it is clear that the categories are more spread out, which is the objective of the CCA to discriminate maximally between the categories. The high-contributing fish species have changed now, apart from *Bo_sa* which still maintains its important position on the first dimension, separating year 1999. The cloud of samples in upper left are associated with species extending out in that direction of the ordination, found in warm water coming from the Atlantic in the south, while the cloud of samples at bottom left is associated with deep water species found in the western area. The temporal trend is now clearer, with years 2003 and 2004 tending even more towards the warmer area of the map.

Each year points shows the centroid of all the samples for a particular year, grouping all the fuzzy regions. A trajectory for each region can be indicated as well, this time as supplementary points – that is, we fix the CCA solution and compute centroids for regional subsets of samples over the years. Exhibit 19.6 shows the regional trajectories for the categories N, E, W and S as well as their overall spatial and time centroids that were shown in Exhibit 19.5. It can be seen now that, of these four regions shown, it is mainly the southern and eastern regions that continue moving towards the "warm" region of the map in 2003 and 2004, whereas in the northern and western regions the warming trend stops from 2003 to 2004. In this way one can interpret the interaction between space and time, seeing the difference in trends between regions, or equivalently the difference in spatial patterns over time, while the six year points and nine region points show the average time trend and spatial pattern.

Some permutation tests

We can conduct various permutation tests to make conclusions about the statistical significance of the CCA results. A first test can be to confirm, as we surely believe, that the association between the abundance data and all the environmental data is significant. The environmental data set is kept fixed

Exhibit 19.5:
Canonical correspondence analysis of the "Barents fish trends" data. The format is the same as Exhibit 19.4, with the samples and species plotted in the upper biplot and an enlarged version of the category centroids in the lower plot. 58.5% of the restricted inertia is explained by these two dimensions



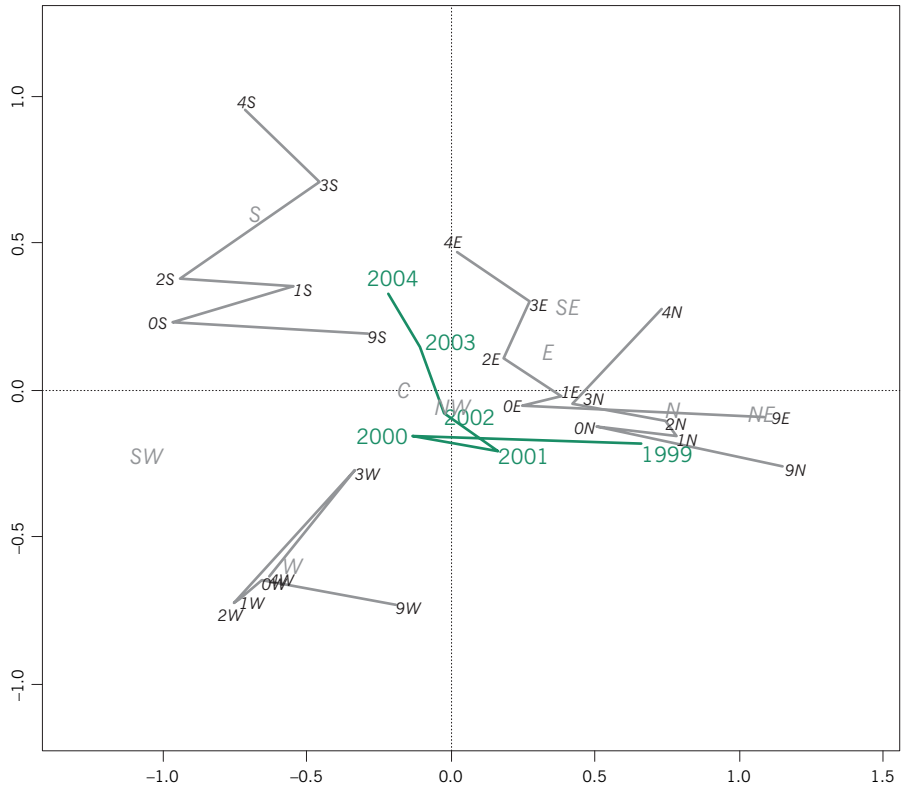


Exhibit 19.6: *Temporal trajectories in regional categories north, east, south and west. Time and regional centroids are at (weighted) averages of the corresponding category points: for example, S is at the average of the six points making up the trajectory for south, while 2004 is at the average of all the 2004 points (for all nine regions, only four shown here)*

and the samples in the abundance data set are permuted many times. In 1,000 permutations the highest inertia explained is by the original data, so the significance is $p = 0.001$ at most. What is more interesting is to see the significance of individual variables. Using them one at a time as constraining variables, the associated p-values are all highly significant ($p = 0.001$) except for slope ($p = 0.12$). Ordering them by explained inertia, Exhibit 19.7 shows the percentage of variance explained, denoted by R^2 because it is the direct analogue of the coefficient of determination in regression, as well as an analogue of the

Variable	k	R^2	Adjusted R^2
Spatial	9	0.275	0.242
Temperature	4	0.119	0.104
Depth	4	0.086	0.070
Year	6	0.057	0.030
Slope	4	0.024	0.007

Exhibit 19.7: *In descending order, the proportion of inertia explained, R^2, and adjusted R^2, of the five categorical environmental variables;* k *is the number of categories*

adjusted R^2 which takes into account the number of categories (see Appendix A and B for details). The spatial position of the sample has by far the most explanatory power.

Another statistical aspect of the constrained ordination solution that requires investigation is the dimensionality of the solution. The scree plot of inertias on successive dimensions is shown in Exhibit 19.8, suggesting a three-dimensional solution. A permutation test for the percentages of constrained inertia, as described in Chapter 18, but applied to the inertias on the constrained dimensions, confirms without any doubt that there are actually three significant dimensions in the constrained space. On the supporting website of this course there is a video of the three-dimensional ordination, which gives an idea of this additional dimension and the 19.6% additional inertia it accounts for.

Isolating the spatial part of the explained inertia

Because the spatial component is intimately related to the environmental variables, especially temperature, it is possible to use CCA to isolate which part of the constrained inertia is purely due to the spatial component and not confounded with the environmental variables. A partial CCA is used, which involves first partialling out the effect of one set of variables, and then doing a CCA on the residuals using a different set of constraining variables. The steps in separating contributions to inertia of inter-correlated variables are as follows:

– Perform the CCA with all constraining variables, in this case environmental, temporal and spatial: the inertia in the constrained space is 1.4746, i.e. 36.7% of the total inertia of the abundance data of 4.0170.

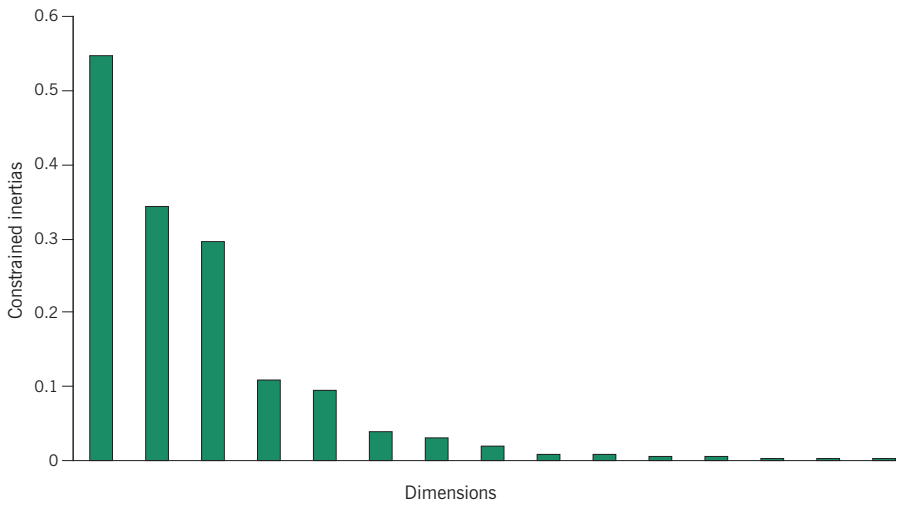


Exhibit 19.8: *Scree plot of the inertias of successive dimensions in the constrained space of the CCA of the "Barents fish trends" data. The first three dimensions clearly stand out from the rest*

– Perform the CCA with the environmental and temporal variables constraining: the inertia in the constrained space is now 0.9761, i.e. 24.3% of the total.

– Perform the CCA with the spatial variables constraining: the inertia in the constrained space is now 1.1044, i.e. 27.5% of the total. It is clear that there must be some confounding between the spatial variables and the others, because 24.3% + 27.5% is much higher than 36.7%, that is the total constrained inertia including all variables in a CCA.

– Perform the CCA with the environmental and temporal variables constraining after partialling out the spatial variation: the inertia in this constrained space is 0.3703, i.e. 9.2% of the total. This 9.2% of the inertia is due to the environmental variables only in a space uncorrelated with the spatial variation.

– Similarly, perform the CCA with the spatial variables constraining after partialling out the other variables: the inertia in the constrained space is now 0.4982, i.e. 12.4% of the total.

– From the last two calculations it must be that 36.7% – 9.2% – 12.4% = 15.1% is inertia due to that common effect of the spatial with the other environmental and temporal variables.

This set of inertia components can be depicted in compositional form as shown in Exhibit 19.9. This figure shows how much of the variation is unexplained and how the part that is explained is divided between the spatial and environmental predictors (where "environmental" includes the temporal trend in this case).

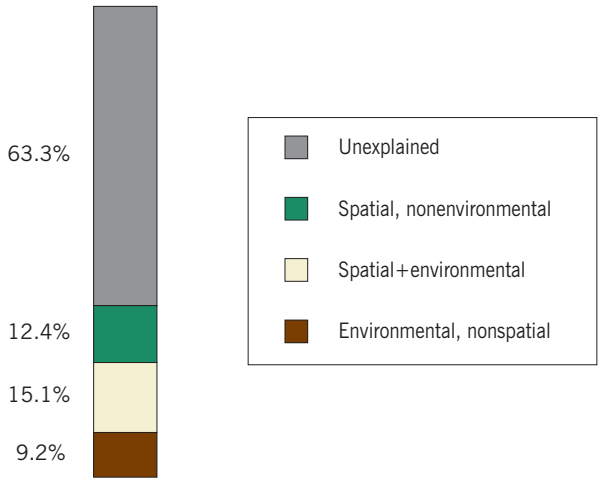


Exhibit 19.9:
Partitioning the total inertia in the abundance data into parts due to the spatial variables and other variables separately, and their part in common

Remember that, of the 36.7% explained variance, the two-dimensional CCA ordination of Exhibits 19.5 and 19.6 only accounts for 58.5%, that is 21.5% of the total variation in the fish abundances. If we take into account the third dimension (see Exhibit 19.8), this brings the explained constrained inertia up to 58.5 + 19.6 = 78.1%, which is 28.7% of the total inertia. Hence, in summary, using the available explanatory variables, depth, temperature, spatial position and year, we can give a statistically justifiable explanation of 28.7% of the variation in the species abundances.

SUMMARY: Case Study 1: Temporal trends and spatial trends across a large ecological data set

1. This case study involved a large data set of fish abundances from trawl samples taken in the Barents Sea, over a six-year period. In addition to the fish data, the environmental variables bottom depth, water temperature and slope of sea-bed were available for each sampling site, as well as latitude and longitude coordinates.

2. In studies such as these that involve sampling across a region over time it can happen that there is unrepresentative sampling in certain areas at different time periods. Conclusions about temporal trends, for example, can become biased due to these sampling imbalances.

3. Samples can be reweighted to be in line with some fixed distribution. In this study we took the distribution over the whole six-year period as the target distribution and reweighted the samples in nine fuzzy regions to be in line with this distribution, thereby eliminating bias in the estimates.

4. Sample weights can be used to reweight the abundance data, after which ordination by CA or CCA, for example, continues as before. In computing average temperatures or diversity measures across the whole region, weighted averages are used.

5. Permutation testing is useful for verifying that the relationship between the fish abundances, regarded as responses, have a statistically significant relationship with the environmental variables and to confirm temporal trends. Similarly, we can test how many dimensions in the solution are nonrandom.

6. The overall variation in the abundance data can be partitioned into a part explained by the environmental and spatial variables. The environmental and spatial predictors are confounded, however, but we can quantify the parts of variation that are purely environmental, purely spatial and a confounding of environmental and spatial.

CHAPTER 20

Case Study 2: Functional Diversity of Fish in the Barents Sea

The ability of a marine ecosystem to withstand environmental changes depends on its adaptability. Biodiversity makes an ecosystem more adaptable and thereby less vulnerable to change since a high number of species can perform a wide range of ecosystem functions present in the community. Diversity can be measured at the taxonomic level, the phylogenetic level or the functional level, and it is the object of this case study to investigate the last option in the same data set studied in Chapter 19. In order to measure functional diversity, species need to be coded in terms of their functional traits. There are then two alternative ways of proceeding: either create groups of species with similar functional traits and then measure diversity of the functional groups, or use a diversity measure which depends on the particular mix of species present at a site, and how far apart they are in terms of their trait characteristics. Both these approaches will be illustrated in this case study, as well as their relationships to environmental, spatial and temporal variables.

Contents

The functional trait matrix

The starting point for a study of functional diversity is the definition of a set of attributes, called *functional traits*, that define the functioning of the species. These can be the type of feeding, movement and reproductive behaviour, for example.

[1] We are indebted to Magnus Wiedmann of the University of Tromsø for his agreement to use these data, which are part of his PhD thesis and an article in the journal *Marine Ecology Progress Series* (see Bibliographical Appendix).

Exhibit 20.1 shows a part of the trait[1] matrix for the 62 Barents Sea fish species studied in Chapter 19. The total list of traits is as follows:

Diet: three-category variable, multiple responses possible

Habitat: two-category variable

Average fecundity: continuous variable, highly positively skew

Offspring size: three-category variable

Offspring behaviour: three-category variable

Maximum size: continuous variable, highly positively skew

Shape: five-category variable

Salinity range: three-category variable

Temperature range: three-category variable

Depth range: three-category variable

Thus, there are 10 traits, 8 categorical and 2 continuous.

As can be seen in Exhibit 20.1, the categorical options are coded as zeros and ones, and the first variable (diet) can have more than one option indicated as a trait (for

Exhibit 20.1: *Part of the trait matrix coding the various functional characteristics of Barents Sea fish species*

Species		Functional traits								
		Diet			Habitat		Fecundity	Offspring		
Name	Abbrevn	benthivorous	ichthyivorous	planktivorous	demersal	pelagic	(mean)	small	medium	large ⋯
Amblyraja hyperborea	Am_hy	1	1	0	1	0	30	0	0	1 ⋯
Amblyraja radiata	Am_ra	1	1	0	1	0	26.5	0	0	1 ⋯
Anarhichas denticulatus	An_de	1	1	1	1	0	46,500	0	1	0 ⋯
Anarhichas lupus	An_lu	1	0	0	1	0	12,740	0	1	0 ⋯
Anisarchus medius	An_me	1	0	0	1	0	700	1	0	0 ⋯
Anarhichas minor	An_mi	1	0	0	1	0	19,700	0	1	0 ⋯
Artediellus atlanticus	Ar_at	1	1	0	1	0	117.5	0	1	0 ⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

example, the first species is indicated as both benthivorous and ichthyivorous for diet), whereas the others only allow one option; for example, offspring can only be in one category of small, medium or large).

Distances between species based on the traits

There are several approaches to defining a distance, or dissimilarity, between pairs of fish, based on mixed-scale data such as these. Our choice here will be to code each of the continuous variables into three fuzzy categories so that the whole trait matrix can be treated as a set of categorical data. Because the two continuous variables are highly skew, it is important to first log-transform them before the fuzzy coding. Several distance functions are possible: simply computing the sum of absolute differences between the traits of pairs of fish, or applying a distance like the chi-square distance that will normalize the traits according to their average appearance in all the fish. We chose the former approach, so that the distance between fish would not depend on the particular sample of fish included in this study (the chi-square distance would depend on the marginal trait averages). Nevertheless, to get an idea of the relationship between this set of fish and the traits, CA using the chi-square distance is still of interest, as shown in Exhibit 20.2. In the upper right corner, for example, we find fish that must have some of the following characteristics: small (ML1) and bottom dwelling (Demersal) benthivorous species, with strange shapes (*Shape_eellike* or *Shape_deep_short*), having few (FM1), medium-sized *(Medium_offspring)*, demersal eggs *(Egg_dem)* and moderate tolerance to variations in abiotic factors such as temperature and salinity.

Hierarchical clustering of fish using trait distances

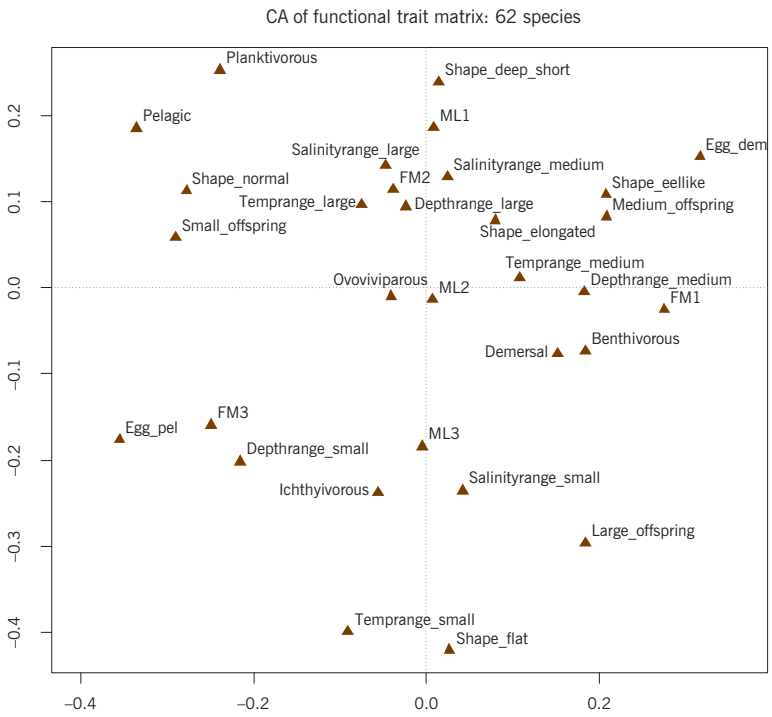
Having defined a distance between the fish, the next step is to perform a clustering of the fish into groups that are relatively homogenous with respect to the traits. Again several choices are available: complete or average linkage or Ward clustering. To ensure a certain level of compactness of the clusters we chose complete linkage – see Exhibit 20.3. Notice that the distance measure has been rescaled so that 1 equals maximum distance.

There are two approaches to defining functional diversity that we shall investigate here. The first way involves defining functional groups, using the results of the hierarchical clustering. Using the permutation test for clustering described in Chapter 17, we obtain the following estimates of p-values for significant clustering, from 2 to 12 groups:

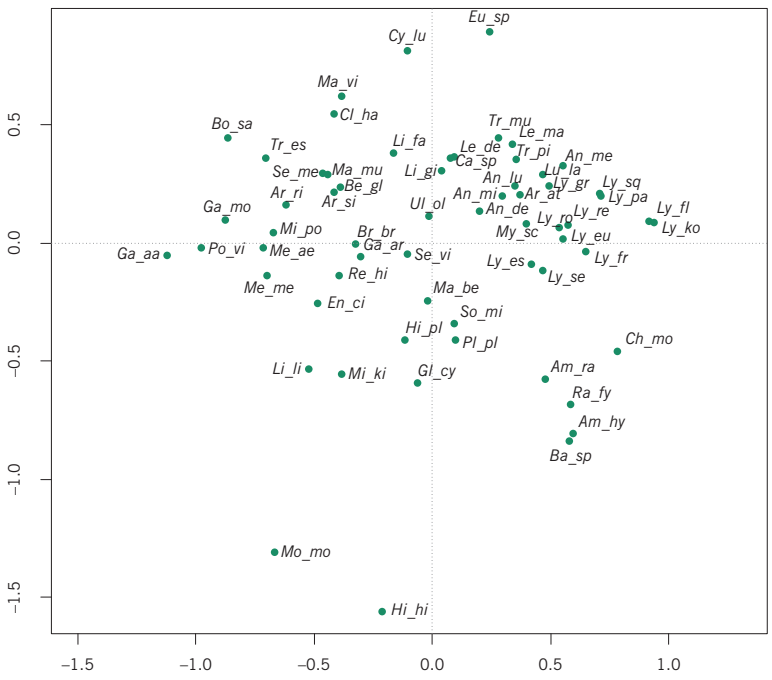
2 groups: $p = 0.989$	6 groups: $p = 0.021$	10 groups: $p = 0.048$
3 groups: $p = 0.975$	7 groups: $p = 0.177$	11 groups: $p = 0.082$
4 groups: $p = 0.354$	8 groups: $p = 0.001$	12 groups: $p = 0.019$
5 groups: $p = 0.821$	9 groups: $p = 0.006$	

Exhibit 20.2: *CA of the trait matrix, part of which is shown in Exhibit 20.1. Traits are shown in principal coordinates in (a) and the fish species in principal coordinates in (b). 27.4% of the inertia is displayed*

(a)



(b)



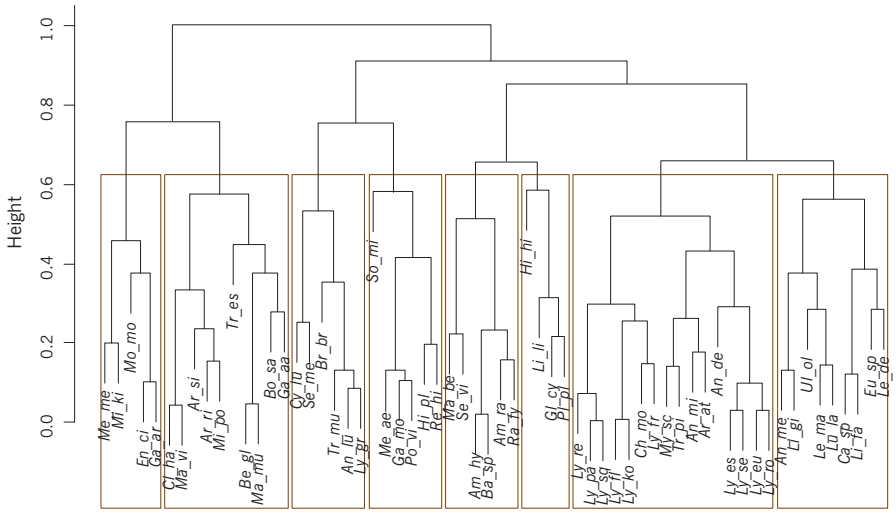


Exhibit 20.3:
Hierarchical clustering of fish based on distances between fish, showing boxes indicating eight clusters

We choose the eight-cluster solution, which is the most significant ($p = 0.001$), indicated in Exhibit 20.3. A six-group solution ($p = 0.021$) is another possibility if fewer groups are required, but we preferred more groups that are internally more homogeneous. Notice in the dendrogram that the nine-group solution ($p = 0.006$) would split off one species on its own, which is not desirable. Hence, the decision about the number of groups is based on statistical significance as a guideline, but also the nature of the dendrogram and substantive biological knowledge.

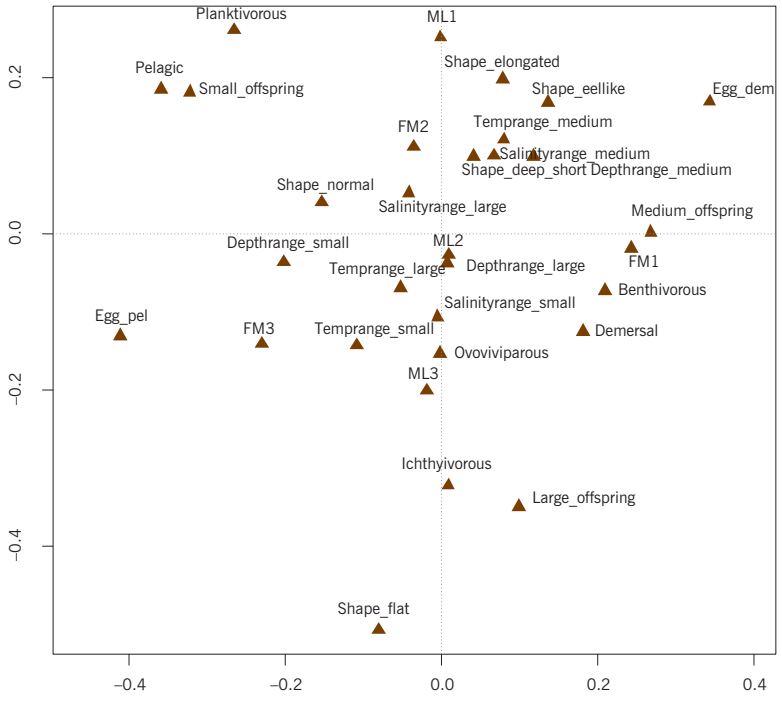
Once functional groups have been defined another CA is possible, at the group level, to interpret these definitions. The trait values for each fish group are aggregated, so we reduce the 62-row trait matrix in Exhibit 20.1 to a 8-row matrix – see Exhibit 20.4. As mentioned in Chapter 16, the CA of this aggregated matrix is a type of discriminant analysis between the fish groups, or alternatively a CCA of the original trait matrix with fish group as a constraining variable. In Exhibit 20.4(b) we show the fish group centroids as well as the convex hulls around the fish species in each group. There is some overlap because not all of the inter-group variance, contained in seven dimensions (one less than the number of groups), can be shown in the two-dimensional map.

Definition of functional diversity

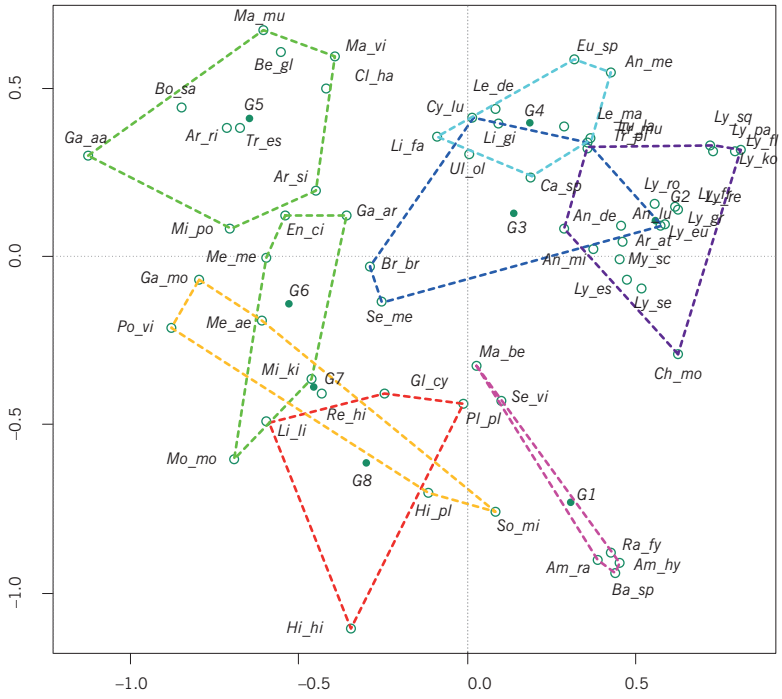
Once the tree, or dendrogram, given in Exhibit 20.3 is established, there are two ways to define functional diversity at a sampling site, one of which depends on having decided on the number of groups, as we have already done above, and the other which only needs the tree. The former is simple to understand: given a sample of fish at a site along with their abundance values, they are classified into groups and their abundance values are aggregated. Then a standard

Exhibit 20.4: *CA of the trait matrix aggregated according to the fish groups (G1 to G8) that were defined in Exhibit 20.3. The solution optimizes the group differences, although the basic configuration is similar to that of Exhibit 20.2 which optimized the fish differences. The functional traits are displayed in contribution coordinates in (a). 52.4% of the inertia between fish groups is displayed*

(a)



(b)



diversity measure is computed, for example the Shannon-Weaver index, denoted by H':

$$H' = -\sum_g p_g \log(p_g)$$

where p_g is the proportional abundance of functional group g.

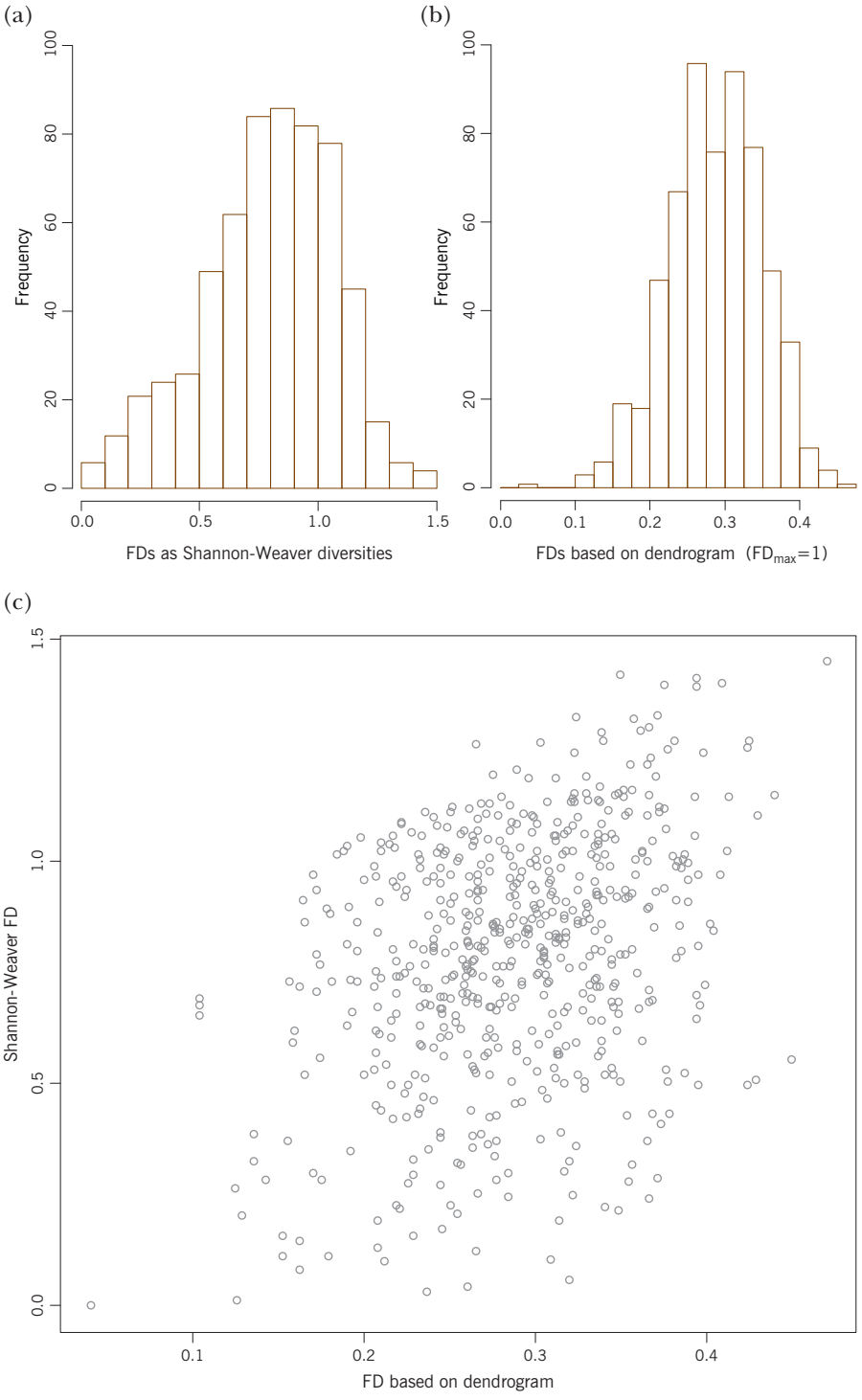
The other way is to measure diversity by summing branches on the dendrogram according to the mix of fish species found in the sample – in this case only presences of fish are used and not their abundances, although an abundance-weighted measure can be envisaged. First let us suppose that a sample contains all 62 fish, which would give the maximum diversity possible. The measure of diversity is obtained by summing all the vertical branches in Exhibit 20.3: since each of the $n-2$ nodes of the tree has two vertical branches below it, this is the sum of $2(n-1) = 122$ values in this example, equal here to 20.31. This value is called the *functional diversity* of the *species pool* (henceforward, we use the abbreviation FD for functional diversity).

Now for a general sample that contains only a subset of the fish species, the FD value is computed by summing all the branches linking this subset. Clearly, if the fish in a subset are "close" together in terms of trait distance, then the sum of the associated branches will be relatively low, while if there are fish species in the sample that are "far apart" with not so many common traits, then the sum of their linking branches will be relatively high. To normalize the FD measures, we shall express them relative to the maximum value of 20.31 for the species pool, so that FD will be between 0 and 1.

Using the same data as in Chapter 19, the FD values for each of the 600 sampling sites were computed in the two different ways, first as the diversity index H' taking into account the aggregated abundances, and second as the normalized value lying between 0 and 1 that only uses presences of the fish. Exhibit 20.5 shows the histograms of FD for each alternative, as well as a scatterplot of their paired values. The fairly low rank correlation of 0.3 suggests that these two measures reflect different information about the diversity.

Interestingly, it is feasible to make a permutation test on the species pool FD as an alternative test for overall clusteredness of the fish, different from testing for a particular number of groups. If the trait data are randomly permuted within each variable, e.g., within diet the three options are permuted together across the fish (and not separately), many alternative values of the species pool FD can be obtained, under a null hypothesis of no relationship between the traits. Exhibit 20.6 shows that the observed FD value is much lower than those obtained under the

Exhibit 20.5: *(a) Histogram of the group-based FDs defined as Shannon-Weaver diversities on the aggregated abundances in 600 samples for eight functional groups; (b) Histogram of the tree-based FDs using presences only and summing the branches in the dendrogram for the subset of observed species, normalized with respect to the FD of the species pool; (c) Scatterplot of the two functional diversity indices (Spearman rho correlation = 0.300)*



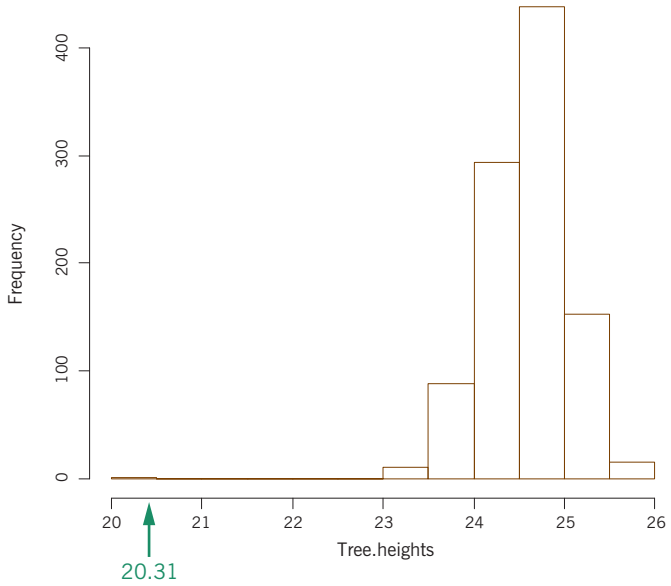


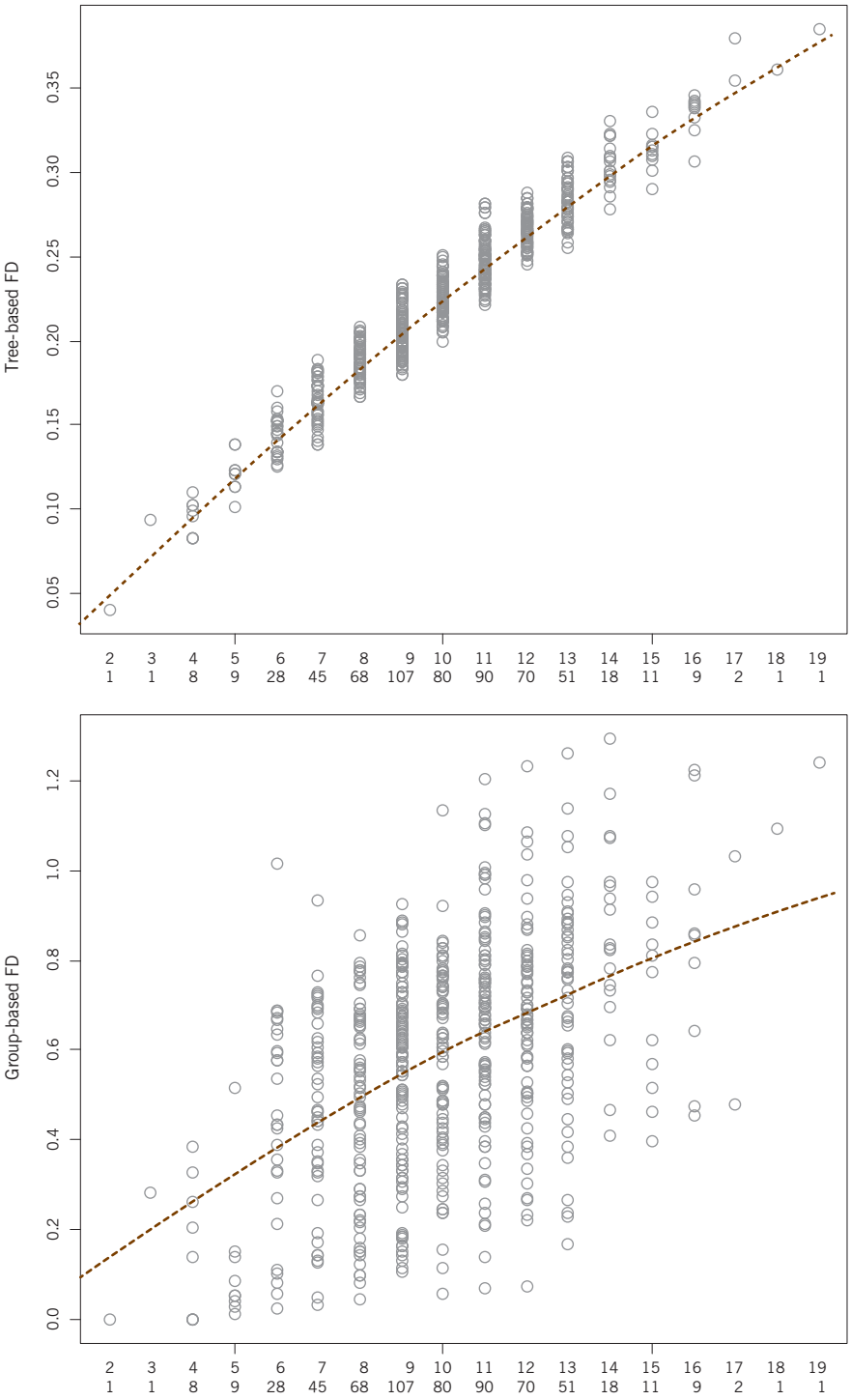
Exhibit 20.6: *Permutation distribution of the species pool FD, under the null hypothesis of no relationship between the traits. The observed value of 20.31 is the smallest and the associated* p*-value, based on 1,000 permutations, is thus* p = *0.001*

null hypothesis, and the estimated p-value is $p = 0.001$, showing that there are significant similarities between the fish across the traits.

Relating functional diversity to species richness

Both measures of functional diversity will be used and compared in the remainder of this chapter, since they appear to contain different information and are defined in different ways, the common feature being the dendrogram based on the trait distances. Exhibit 20.7 shows their relationships with species richness (SR), that is the number of species in each sample. Since the tree-based FD only takes presences into account and would clearly increase with increasing number of species, as more branch lengths are summed, it is no surprise that it follows species richness very closely (Exhibit 20.7(a)). Both relationships are slightly nonlinear, with concave curves, and so we would use a quadratic function, for example, as a model for the conditional means, shown in Exhibit 20.7 (in both cases the explanatory terms SR and SR^2 are highly significant in the regressions). The deviations of the functional diversity values from that expected by their relationship with species richness are used as a measure of so-called *functional dispersion*. Higher functional dispersions at a site are associated with greater ecosystem adaptability because the number of functions displayed by the species at this site is higher than expected given the number of species present – they possess more "tools" and are thus expected to be better prepared for environmental change. On the other hand, the impact on the FD due to the loss of a species would be proportionally larger at this site since each species contributes more to the FD as compared to a site with the same SR but a lower FD (i.e., lower functional dispersion).

Exhibit 20.7: *Scatterplots of the two FD measures versus species richness (SR, the number of species in sample), showing the modelled quadratic relationships. The horizontal axis is marked with the value of SR, and below the number of sites with the corresponding value)*



Relating functional diversity to space, time and environment

As a first bivariate view of associations between the two FD measures and the available covariates, Exhibit 20.8 shows the matrix of scatterplots, with Spearman rank correlations in the upper triangle and scatterplots and smooth relationhips in the lower triangle. Apart from the known features of the region, that depth is negatively correlated with longitude and temperature negatively correlated with latitude and longitude, the group-based FD is correlated with depth and the tree-based FD negatively with temperature and positively with latitude and longitude, although these last correlations are less than 0.30 in absolute value. As already seen in Chapter 19, the variable slope does not appear to have any association with any other, so we drop it from further consideration.

To show the spatial relationship latitude and longitude should be considered together along with their interaction. We can compare two ways of spatial modelling, by spatial fuzzy coding (Chapter 11) and by generalized additive modelling (GAM,

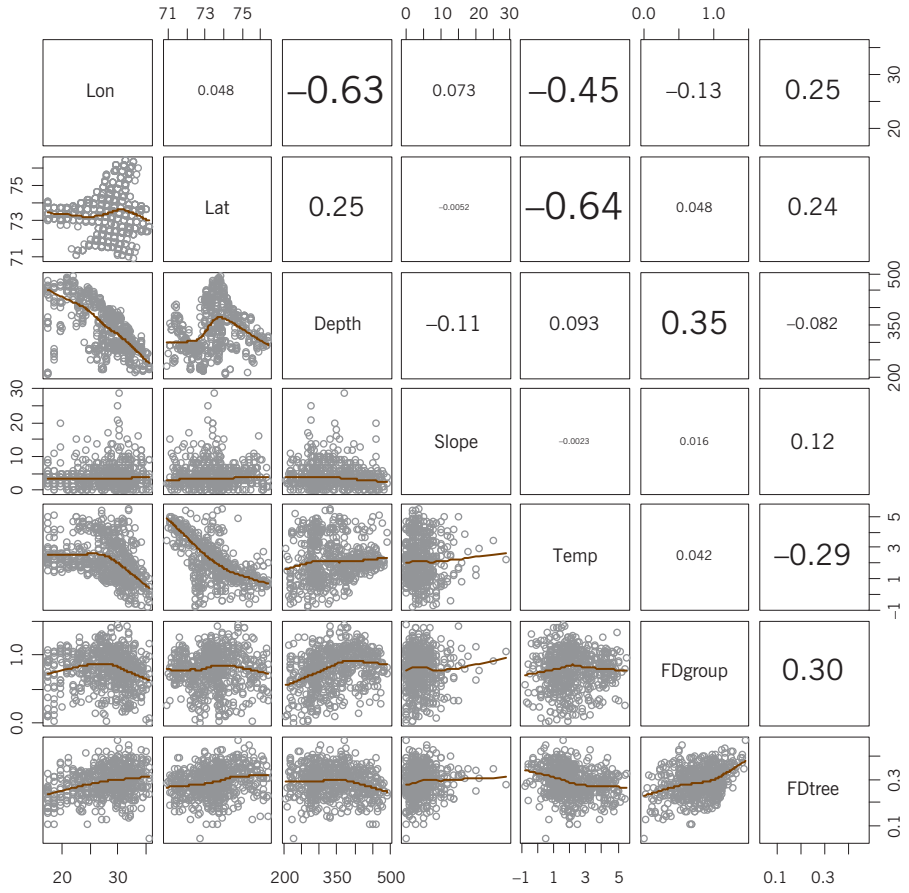


Exhibit 20.8: *Scatterplots of the variables depth, slope, temperature, longitude and latitude with one another as well as with the two measures of functional diversity, based on the functional groups (FDgroup) and on the dendrogram (FDtree). Spearman rank correlations are shown in the upper triangle, with font size proportional to their absolute values.*

Chapter 18). For both models we model each FD measure on latitude and longitude interactively, and the results are shown in Exhibit 20.9 in the form of contours of predicted FD values.

The results or the two types of FD are quite different, with the tree-based FD showing a west and south to north-east gradient, with higher FD in the north-east, whereas the group-based FD shows higher diversity in the central area, falling off to the west and the south-east. Remember that the group-based FD takes the abundance values into account and the water in the central areas is warm, and species from more southern areas (e.g., Norwegian Sea) migrate into these areas (often in schools), especially in warmer years, giving a more equal spread of relative abundance values in the functional groups. Concerning the tree-based FD, the GAM fit shows a ridge in the diversity values from south to north while the fuzzy

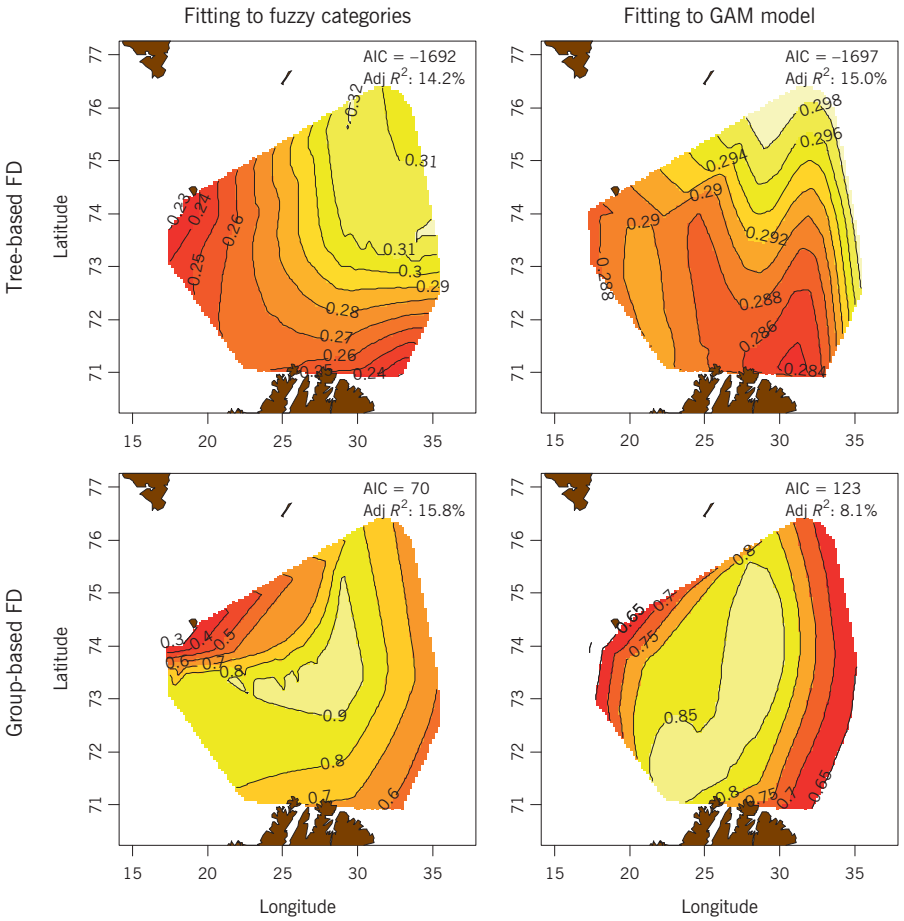


Exhibit 20.9: *Contour plots of the spatial component of functional diversity according to the two definitions (first row is the tree-based FD, second row is group-based FD) using two modelling methods (in columns, first column is using fuzzy spatial categories, second is using GAM modelling). The northern border of Norway with Russia and the southern tip of Svalbard situate the region of interest*

fit shows a wider ridge from south-west to north-east. For the group-based FD the results are similar between the two methodological approaches, but the fuzzy approach performs noticeably better according to AIC and the adjusted R^2. Another advantage of the approach using fuzzy-coded categories is that there is a p-value associated with every compass point's difference with the central category. So we can get results that for the group-based FD several sectors are significantly lower than the central (C) one: NW, E, SE and S (all with $p < 0.001$), NE ($p = 0.002$) and W ($p = 0.04$), whereas for the tree-based FD the following sectors are significantly lower than the central one: NW ($p < 0.001$), SE ($p = 0.002$) and S ($p = 0.005$).

Although the spatial variation is highly linked to the variation of environmental variables such as temperature and possibly also to temporal variation, we can study inter-year variation in the residuals from the above spatial models as well as any further relationships with the environmental variables temperature and depth. As an example, we consider the residuals of tree-based classification from the fuzzy spatial model (top left example in Exhibit 20.9), and model the residuals on year as a categorical variable, and temperature and depth either as regular continuous variables, or the four-category fuzzy versions used in Chapter 19, or as smooth functions using GAM. Both temperature and depth are found to be nonsignificant predictors of the residuals, irrespective of the coding. There is significant temporal variation, however, almost identical in all analyses, which can be plotted as in Exhibit 20.10. Remembering that these are the residuals from the spatial model, we can say that in 1999 and 2001 there were lower functional diversities compared to the spatial model (as measured by the dendrogram-based approach) and higher in 2003 and 2004. All effects are different from 0 (the mean of the residuals) and highly significant ($p < 0.0001$), apart from 2002 which is closer to 0 ($p = 0.025$).

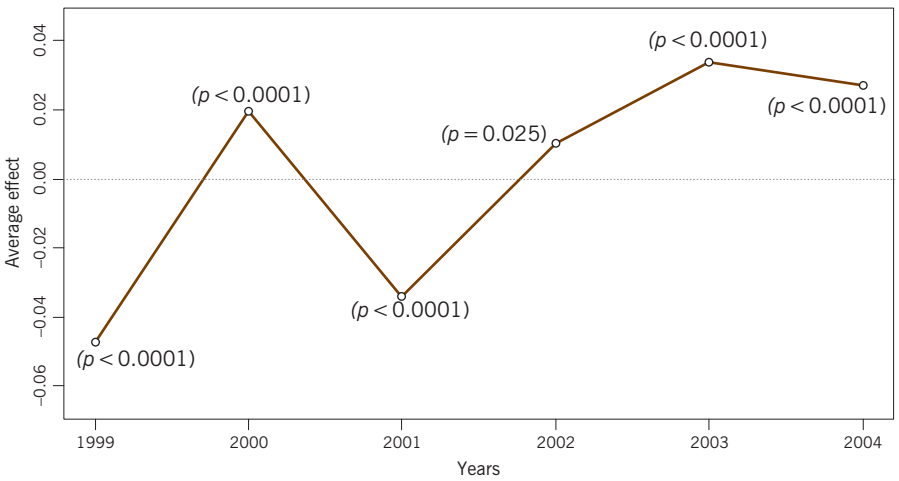
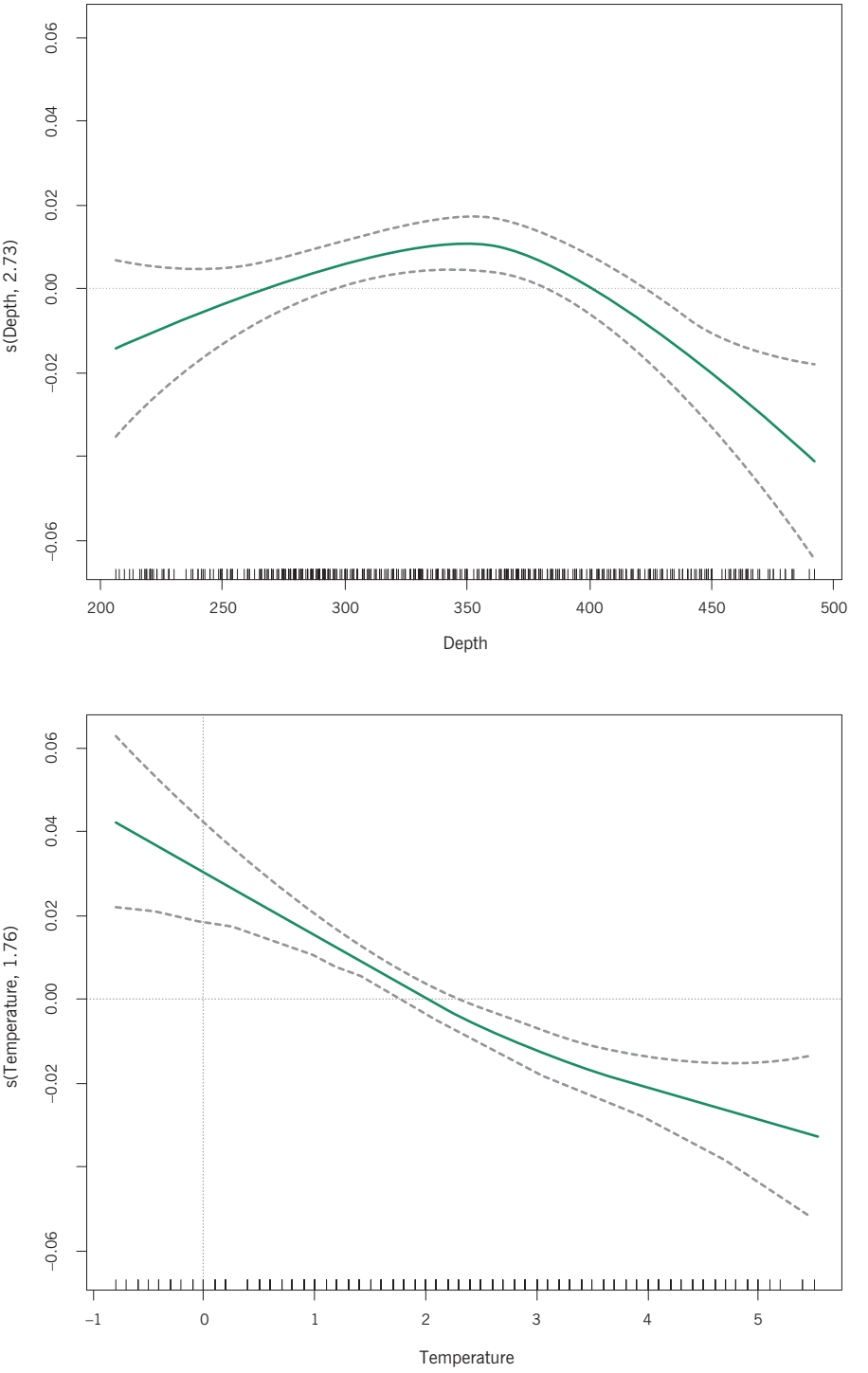


Exhibit 20.10: *Plot of regression coefficients for each year showing average estimated year effects for the residuals of (tree-based) functional diversities from the spatial model, with* p*-values for testing differences compared to the zero mean of the residuals (dashed line)*

Exhibit 20.11: *GAM of tree-based FD as a smooth function of depth* ($p = 0.0001$) *and temperature* ($p < 0.0001$). *To model these effects parametrically depth would be modelled as a quadratic and temperature linear*



From the above it is clear that the effects of spatial position and of the environmental variables temperature and depth are confounded and difficult to separate. If FD is first related to temperature and depth, ignoring the spatial component, highly significant relationships are found: for example, tree-based FD goes down with increasing temperature and we find the same quadratic relationship with depth as in Chapter 18 where the response was species diversity – see Exhibit 18.7 for the analysis of only 89 sites (where temperature was nonsignificant), and Exhibit 20.11 for the present example of 600 sites. More or less the same depth value, about 350 m, is found here for maximum FD as was found before in Exhibit 18.7 for maximum species diversity. Adding the year effects gives almost exactly the same pattern as in Exhibit 20.9, with 1999 and 2001 low and the other years high. Of the FD variance, 30.2% (adjusted R^2) is explained by depth, temperature and years. Residuals from this environmental and temporal relationship, accounting for about 70% of the FD variance, can then be modelled spatially: using a GAM model as in Exhibit 20.8 there is still a significant spatial component in the residuals, although the explained variance in these residuals is only 4.3%.

In summary, temperature and depth, both of which are related to spatial position in the Barents Sea, are found to be strongly associated with functional diversity, and there are also significant differences between the years. Residuals from a model of FD as a function of these environmental and temporal variables can be explained, although to a minor extent, by spatial position.

SUMMARY: Functional diversity of fish in the Barents Sea

1. Functional diversity measures diversity in the *functional traits* (feeding, motion, reproductive behaviour, habitat preferences, etc.) among species in an ecosystem.

2. Functional groups are groups of species that share the same functional traits.

3. To measure functional diversity two approaches are considered here, both based on a dendrogram obtained by hierarchical clustering of the species according to their functional traits. They are thus both dependent on the distance/dissimilarity function used as well as the type of clustering.

4. The first way is to use the hierarchical clustering to decide on the number of clusters that are sufficiently homogeneous internally to be considered separate groups. Functional diversity (FD) at a site can then be measured by any of the usual diversity measures, for example the Shannon-Weaver diversity, which is a function of relative abundances (or biomasses) of the functional groups. We call this *group-based FD*.

5. The second way is to add up the branches of the dendrogram of the particular mix of species at the site – this takes only presences of species into account, not their abundances. We call this *tree-based FD*.

6. These FD measures are found to have monotonically increasing, slightly concave, relationships with species richness (SR). Tree-based FD is very closely related to SR because both take only species presences into account.

7. Both FD measures can be related to spatial, temporal and environmental variables in the usual way using multiple regression. Spatial coordinates are interactively coded to explain the spatial relationship. Continuous explanatory variables can be coded in their original form, possibly transformed to account for nonlinear relationships, or coded as fuzzy variables.

8. An alternative modelling strategy is to use generalized additive modelling (GAM) which produces a smooth regression relationship with the two-dimensional spatial position and the continuous variables.

APPENDICES

Appendix A

Aspects of Theory

This appendix summarizes the theory described in this book. The treatment is definitely not exhaustive and the bibliography in Appendix B gives some pointers to additional reference material. We deal with the theory in more or less the same order as the corresponding methods appeared in the text, although some topics might be grouped slightly differently.

Contents

Transformations and standardization

The most common measurements scales are:

- *Continuous interval:* differences between values are measured and interpreted; variables on this scale can have negative values; we also say an *additive* scale. For example, time, temperature.
- *Continuous ratio:* ratios between values are measured and interpreted (i.e., percentage differences); variables on this scale have positive values; we also say a *multiplicative* scale. For example, heavy metal concentration, weight.
- *Categorical* (or *discrete*) *nominal:* only a few categories are possible and they have no particular order. For example, region, phylogenetic group.
- *Categorical* (or *discrete*) *ordinal:* only a few categories are possible and they do have an inherent ordering. For example, month, sediment class.

- *Count:* Variable that takes positive integer values, including 0; we also say a *frequency*. For example, abundance count, number of offspring.
- *Compositional:* this refers to a set of variables with the unit-sum constraint, proportions that add up to 1 (or 100% if percentages). For example, a set of fatty acid compositions, relative abundances of a set of species.

Standardization is often applied to put different variables on the same scale. For data $x_1, \ldots, x_n$ on an interval-scale variable the most common is to make them all mean 0, variance 1, by *centering* (i.e., subtracting) with respect to the mean $\bar{x}$ and *normalizing* (i.e., dividing) with respect to the standard deviation s.

$$z_i = \frac{x_i - \bar{x}}{s}, \quad i = 1, \ldots, n \tag{A.1}$$

Other forms of standardization might be deemed more suitable, such as centering with respect to the median and normalizing by the range.

For positive data $x_1, \ldots, x_n$ on a ratio-scale variable, a convenient transformation is the logarithm:

$$z_i = \log(x_i) \tag{A.2}$$

because it converts the ratio-scale variable x to an interval-scale variable z, which needs no further normalization.

Nominal and ordinal categorical data are often converted to *dummy variables*, which are as many variables as there are categories, taking the values 0 and 1.

Count data as well as compositional data are similar to ratio-scale variables and are usually logarithmically transformed, or root-transformed (square root, fourth root…). If there are zero count values, then they are often transformed as $\log(1 + x)$. In the case of compositional data, we prefer to replace zeros with small values equal to the detection limit in the context of the data.

The Box-Cox transformation is a general power transformation for ratio-scale, count and compositional data:

$$z = \frac{1}{\lambda}(x^{\lambda} - 1) \tag{A.3}$$

usually for powers λ less than 1, and where for zero values of x, $x^{\lambda} = 0$. As the power λ tends to 0 (we say as the root transformation gets stronger) the transformation gets closer and closer to the log-transformation $\log(x)$.

Measures of distance and dissimilarity

Differences between values of a single interval variable are computed simply by subtraction, while for a ratio variable or a count, multiplicative differences can be computed by taking ratios, or differences on the log-scale. For multivariate samples difference is measured by a distance or dissimilarity which combines differences across the variables. A *distance* has all the properties of a well-defined metric, including the triangular inequality property. A *dissimilarity* is an acceptable measure of inter-sample difference but does not obey the triangular inequality.

A general distance is the *weighted Euclidean distance*, computed between two samples $x_1,\ldots,x_p$ and $y_1,\ldots,y_p$ observed on p variables, with weights on the variables $w_1,\ldots,w_p$:

$$d_{x,y} = \sqrt{\sum_{j=1}^{p} w_j (x_j - y_j)^2} \qquad \text{(A.4)}$$

Well-known special cases are:

- Euclidean distance, when $w_j = 1$; applicable to set of interval variables all on the same scale that do not need normalization, or a set of ordinal variables, all on the same scale (e.g., five-point ordinal scales of plant coverage) for which the inter-category differences are accepted as interval measures.
- Standardized Euclidean distance, for a set of interval-scale variables: $w_j = 1/s_j^2$, the inverse of the variance of the j-th variable; this is the distance function computed by standardizing all the variables first and then applying the regular unweighted Euclidean distance.
- Chi-square distance, for abundance, relative abundance, and compositional data: $w_j = 1/c_j$, where c_j is the mean for variable j.

The Bray-Curtis (or Sørensen) dissimilarity (which is not a true distance function, since it does not obey the triangle inequality) is a popular choice for measuring differences between samples when the data are abundances, or other positive amounts such as biomasses:

$$b_{x,y} = \frac{\sum_{j=1}^{p} |x_j - y_j|}{\sum_{j=1}^{p} (x_j + y_j)} \qquad \text{(A.5)}$$

For one/zero data, for example presence/absence data, there are many possibilities and we only summarize the two presented in this book, the *matching coefficient* and *Jaccard dissimilarity*. For p variables observed on two samples, we define

a = number of variables matched with a 1 in both samples, d = number of matches of 0 in both samples, b = number of variables "mismatched" with 1s in the first sample, 0s in the second, c = number of mismatches with 0s in the first sample, 1s in the second, so $a + b + c + d = p$. Then:

- Matching: $(b + c)/p$ (actually, this is a measure of mismatching) (A.6)

- Jaccard: $(b + c)/(p - d)$ (A.7)

Both the above dissimilarities lie between 0 and 1, with 0 when there are no mismatches. For matching the maximum value of 1 is attained when $a + d = 0$ (no 1s or 0s matched), while for Jaccard, which ignores matching 0s, the maximum of 1 is reached when $a = 0$ (no 1s matched). Jaccard is preferable for presence/absence data when the co-occurrence of absences is not interesting, only the co-occurrence of presences.

For *mixed-scale* multivariate data, usually continuous and categorical mixed, some form of normalization or homogenization is required so that it makes sense to combine them into a measure of inter-sample difference. The *Gower index of dissimilarity* (not discussed in the book) involves applying a standardization on the continuous variables to make them comparable to the categorical ones that are dummy coded, after which Euclidean distance is applied. The alternative that is presented in this book is to *fuzzy code* the continuous variables into sets of fuzzy categories. Fuzzy categories corresponding to a continuous variable look like a set of dummy variables except that they have any values between 0 and 1, not exactly 0 or 1, and in this way preserve the exact value of the continuous variable in categorical form. With the categorical variables coded as dummy variables and the continuous variables coded as fuzzy categorical variables, Euclidean distance can be applied, possibly with weights to adjust the contributions of each variable to the measure of distance.

Cluster analysis

To define a method of cluster analysis one defines the algorithm used to implement the method. Two approaches are of interest, hierarchical and nonhierarchical clustering, both of which rely on a matrix of proximities (distances or dissimilarities) between pairs of *objects* to be clustered, where objects can be sampling units such as sites or variables such as species.

Hierarchical cluster analysis creates a *dendrogram*, or binary tree, in a stepwise fashion, successively aggregating objects, two at a time, and eventually aggregating groups of objects as well, according to their proximities. Assuming a decision about the measure of proximity has been made, the crucial decision is then how to measure proximity between groups of objects formed in the previous stage of the stepwise procedure. The main options in practice are: (1) complete linkage, where the

maximum distance or dissimilarity value between groups is used; (2) average linkage, where the average value is used; or (3) Ward clustering, a different ANOVA-like approach which maximizes the overall between-group variance at each step of the clustering, equivalently minimizing within-group variance. The final result is a dendrogram, which is then cut at a certain level to create a small number of groups of objects, designed to be internally homogeneous and distinct from one another.

Nonhierarchical cluster analysis is used when the number of objects is very large, say greater than100, when the dendrogram becomes unwieldy to interpret. The most popular example is *k-means clustering*, which has the same objective as Ward clustering, to maximize between-group variance while minimizing within-group variance. The number of groups k is pre-specified and the algorithm proceeds from a random start to create k groups iteratively, at each iteration assigning objects to the group with the closest mean. The solution is seldom globally optimum and several random starts are recommended, and the best final solution accepted.

Multidimensional scaling

While clustering results in a grouping of objects, multidimensional scaling (MDS) results in an ordination map of the objects. Given a matrix of inter-object proximities, MDS finds a configuration of the objects in a space of specified dimensionality, almost always a two-dimensional plane, such that the displayed inter-object distances are as close as possible to the given proximities. Different ways of measuring the fit between the displayed distances, gathered in a matrix $\mathbf{D}$, and the given proximities, gathered in a matrix Δ, lead to different MDS techniques.

Classical MDS, also called *principal coordinate analysis*, relies on the eigenvalue-eigenvector decomposition, called *eigen-decomposition*, of a square matrix of scalar products to obtain a solution. Initially, the elements of the given proximity matrix are squared – this matrix of squared distances or dissimilarities is denoted by $\Delta^{(2)}$. To give the most general form of classical MDS, we assume that there is a set of positive weights $w_1,\ldots,w_n$ assigned to the n objects, where $\sum_i w_i = 1$, so that the objective is to optimize a weighted fit where objects of higher weight are displayed more accurately in the solution. An operation of double-centering and multiplying by $-\frac{1}{2}$ is applied to $\Delta^{(2)}$ to obtain the scalar product matrix $\mathbf{S}$:

$$\mathbf{S} = -\tfrac{1}{2}(\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathsf{T}})\Delta^{(2)}(\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathsf{T}})^{\mathsf{T}} \qquad \text{(A.8)}$$

where $\mathbf{I}$ is the $n \times n$ identity matrix, $\mathbf{1}$ is the $n \times 1$ vector of 1s and $\mathbf{w}$ the $n \times 1$ vector of weights. Centering of the values in the columns of $\Delta^{(2)}$ is performed by premultiplying by the *centering matrix* $(\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathsf{T}})$, while post-multiplying by the transposed centering matrix centers the values in the rows. The eigen-decomposition is then obtained on a weighted form of $\mathbf{S}$, where $\mathbf{D}_w$ is the diagonal matrix of weights:

$$\mathbf{D}_w^{½}\,\mathbf{S}\,\mathbf{D}_w^{½} = \mathbf{U}\mathbf{D}_\lambda\mathbf{U}^{\mathsf{T}} \tag{A.9}$$

U contains the eigenvectors of **S** in its columns and $\mathbf{D}_\lambda$ is a diagonal matrix with the eigenvalues of **S** down the diagonal, in decreasing order. The *principal coordinates* of the objects are finally given by:

$$\mathbf{F} = \mathbf{D}_w^{-½}\mathbf{U}\mathbf{D}_\lambda^{½} \tag{A.10}$$

The rows of **F** refer to the objects and the columns to the principal axes of the solution in decreasing order of importance. For a two-dimensional display the first two columns provide the coordinate pairs (f_{i1}, f_{i2}) for displaying the i-th object. The sum of the eigenvalues are a measure of the total variance and each eigenvalue a measure of explained variance by a principal axis, hence the quality of display in a two-dimensional solution, which can be interpreted like an R^2 in regression, is $(\lambda_1 + \lambda_2)/\sum_k \lambda_k$.

When all the nonzero eigenvalues are positive, the given matrix of proximities is *Euclidean embeddable*, which means that it is possible to represent the objects in a Euclidean space, with dimensionality equal to the number of positive eigenvalues. When there are some negative eigenvalues, their absolute values quantify the part of variance that is impossible to represent in a Euclidean space. For example, a matrix of chi-square distances is always Euclidean embeddable, while a matrix of Bray-Curtis dissimilarities is not. This fact has led to practitioners preferring nonmetric MDS to display Bray-Curtis dissimilarities.

Nonmetric MDS relaxes the measure of fit between the displayed distances and the given proximities. A perfect fit in nonmetric MDS would be when the order of all the displayed distances is the same as the order of all the given proximities. Specifically, if the $½n(n-1)$ displayed distances are listed next to the $½n(n-1)$ given proximities, a perfect fit would give a Spearman rank correlation between the two lists of 1. Rather than measure quality of fit, nonmetric MDS measures error of fit using a quantity called *stress*, so a perfect fit would be a stress of 0. The measure of stress will always appear more optimistic than the measure of unexplained variance in classical MDS, but this does not imply that nonmetric MDS is an improvement – classical MDS has a stricter objective, and thus more error in achieving it.

Principal component analysis, correspondence analysis and log-ratio analysis

These three methods, abbreviated as PCA, CA and LRA, are variations of the same theme, so we treat them together. All three methods start with a rectangular data matrix, prepared according to the method for being decomposed by the singular-value decomposition (SVD). The SVD is similar to the eigen-decomposition but applicable to rectangular rather than square matrices. All three methods can be defined using eigen-decompositions as well, but the SVD approach is more

elegant and brings out clearly the features of the eventual joint display, for example, whether the display is a biplot or not.

The SVD is defined as follows, for a rectangular matrix $\mathbf{A}$ $(I \times J)$:

$$\mathbf{A} = \mathbf{U}\mathbf{D}_\sigma\mathbf{V}^\mathsf{T} \tag{A.11}$$

where $\mathbf{U}$ $(I \times R)$ and $\mathbf{V}$ $(I \times R)$ have orthonormal columns: $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$, and $\mathbf{D}_\sigma$ is a diagonal matrix of positive values in descending order: $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_R > 0$. R is the rank of $\mathbf{A}$. The columns of $\mathbf{U}$ and $\mathbf{V}$ are called *left* and *right singular vectors*, respectively, corresponding to the *singular values* σ_r. (A.11) can be written equivalently as the sum of R terms, each of which involves a singular value and associated pair of singular vectors:

$$\mathbf{A} = \sigma_1\mathbf{u}_1\mathbf{v}_1^\mathsf{T} + \sigma_2\mathbf{u}_2\mathbf{v}_2^\mathsf{T} + \cdots + \sigma_R\mathbf{u}_R\mathbf{v}_R^\mathsf{T} \tag{A.12}$$

Since each matrix $\mathbf{u}_r\mathbf{v}_r^\mathsf{T}$ has sum of squared elements equal to 1 and the singular values are in descending order, this already suggests that the first terms of (A.12) come close to reproducing the matrix $\mathbf{A}$. In fact, the famous Eckart-Young theorem states that the first R^* terms constitute a rank R^* least-squares matrix approximation of $\mathbf{A}$ – if we take the first two terms, for example, which is the most popular choice, then we have a rank 2 approximation of $\mathbf{A}$, and this will provide us with coordinates of points representing the rows and columns of $\mathbf{A}$ in a two-dimensional plot.

We need a slightly more general form of the SVD to take into account weights assigned to the rows and columns. Suppose $r_1,\ldots,r_I$ and $c_1,\ldots,c_J$ are, respectively, two such sets of weights, all positive and each set adding up to 1. Then, the weighted form of the SVD, which gives weighted least-squares approximations to $\mathbf{A}$, is obtained by first multiplying the elements a_{ij} of the matrix by the square roots of the weights, $(r_ic_j)^{1/2}$, then decomposing this reweighted matrix by the SVD, and finally "de-weighting" the final result. In matrix formulation these three steps are as follows, where $\mathbf{D}_r$ and $\mathbf{D}_c$ denote diagonal matrices of the row and column weights:

- Weight rows and columns: $\mathbf{D}_r^{1/2}\mathbf{A}\mathbf{D}_c^{1/2}$ (A.13)
- Compute SVD: $\mathbf{D}_r^{1/2}\mathbf{A}\mathbf{D}_c^{1/2} = \mathbf{U}\mathbf{D}_\sigma\mathbf{V}^\mathsf{T}$ (A.14)
- "De-weight" to get the solution: $\mathbf{A} = (\mathbf{D}_r^{-1/2}\mathbf{U})\mathbf{D}_\sigma(\mathbf{D}_c^{-1/2}\mathbf{V})^\mathsf{T}$ (A.15)

Solutions of PCA, CA and LRA can be found by specifying the input matrix $\mathbf{A}$ and the weights. In all cases there is some type of centering of the original data matrix

to obtain **A**. Centering of the columns, for example, as seen already in (A.8), is performed by pre-multiplying by $(\mathbf{I} - \mathbf{1r}^\mathsf{T})$, where **r** is the vector of row weights. Centering of the values in the rows involves post-multiplying by $(\mathbf{I} - \mathbf{1c}^\mathsf{T})^\mathsf{T}$, where **c** is the vector of column weights. Here follow the three variants:

PCA: The data matrix **Y** contains interval-scale data, cases in rows, variables in columns. Usually case and variable weights are equal, i.e. $\mathbf{r} = (1/I)\mathbf{1}$ and $\mathbf{c} = (1/J)\mathbf{1}$, where **1** denotes an appropriate vector of ones. The columns are centered and optionally standardized, for example in the unstandardized case, $\mathbf{A} = (\mathbf{I} - (1/I)\mathbf{1}\mathbf{1}^\mathsf{T})\mathbf{Y}$. For the standardized case, divide the values in each column of **Y** by their respective standard deviation.

CA: The data matrix **Y** contains nonnegative ratio-scale data, usually counts such as abundances, or biomasses or percentages. Suppose **P** equals **Y** divided by its grand total, so that the sum of all elements of **P** is 1. The row and column sums of **P** are **r** and **c**, the row and column weights. Compute the matrix of ratios $p_{ij}/(r_i c_j)$, i.e. $\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}$. Then **A** is the double-centered matrix of these ratios: $\mathbf{A} = (\mathbf{I} - \mathbf{1r}^\mathsf{T})\,\mathbf{D}_r^{-1}\mathbf{P}\,\mathbf{D}_c^{-1}(\mathbf{I} - \mathbf{1c}^\mathsf{T})^\mathsf{T}$.

LRA: The starting point of LRA is similar to CA, except that the data matrix **Y** must be strictly positive. Again the masses **r** and **c** are the row and column sums of **Y** relative to the grand total. Then **A** is the double-centered matrix of the logarithms of Y: $\mathbf{A} = (\mathbf{I} - \mathbf{1r}^\mathsf{T})\,\log(\mathbf{Y})\,(\mathbf{I} - \mathbf{1c}^\mathsf{T})^\mathsf{T}$.

After putting these options through steps (A.13)–(A.15), various coordinates can be computed:

Principal row coordinates: $\mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_\sigma$	*Principal column coordinates:* $\mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D}_\sigma$	(A.16)
Standard row coordinates: $\mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}$	*Standard column coordinates:* $\mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}$	(A.17)
Contribution row coordinates: **U**	*Contribution column coordinates:* **V**	(A.18)

In each method the total variance, customarily called *inertia* in CA, is the sum of squared singular values computed in (A.14). This is identical to the sum of squared elements of the weighted matrix in (A.13). The part of variance explained by the first R^* dimensions of the solution (e.g., $R^* = 2$) is the sum of the first R^* squared singular values. The squared singular values are, in fact, eigenvalues in the equivalent definitions in terms of eigen-decompositions.

Supplementary variables and points

In all three methods there is the concept of a *supplementary variable* and a *supplementary point*. A supplementary variable is an additional continuous variable that is related to the low-dimensional solution afterwards, using multiple regression.

When the supplementary variable is standardized and the row standard coordinates are used as explanatory variables, the regression coefficients reduce to the correlation coefficients of the variable with the dimensions, thanks to the dimensions being uncorrelated. Hence the supplementary variable can be represented by coordinates equal to their correlation coefficients. If the rows are displayed in standard coordinates then they have a biplot relationship with these supplementary variables: rows can be projected onto the supplementary variable direction to line up the rows on that variable, the origin being the average. Notice that if there are row weights, then the regression and correlation calculations have to be weighted.

A supplementary point is an additional row (or column) that one wants to add to the existing map. This point differs from a supplementary variable in that it is comparable in scale to the data matrix that was analysed (called the *active* data matrix – sometimes a supplementary point is referred to as *passive*). For example, in a CA of abundance data, there might be additional species, or groups of species, that one wants to situate in the ordination. These have profiles just like the active data, and they can be projected onto the solution just like the active profiles were projected. The only difference is that the supplementary points have not been used to construct the solution, as if they were data points with zero weight. Supplementary points are often used as an alternative way of representing a categorical variable in an ordination. For example, again in CA, suppose the data were fish abundances, with columns as fish and classified into two types, pelagic and demersal. Aggregating all the columns corresponding to pelagic fish and all those corresponding to demersal fish gives two new columns labelled *pelagic* and *demersal*. These aggregated abundances have well-defined profiles in the column space and can be displayed on the ordination – in fact, their positions will be at the respective weighted average positions of the set of pelagic and set of demersal fish. In a similar way, fuzzy categories can be displayed. For example, the rows (e.g., sites) may have fuzzy categories for temperature, so aggregation of abundances is now performed over the rows to get four fictitious sites representing the fuzzy categories. The aggregation must be fuzzy as well, in other words, the abundances are multiplied by the fuzzy value and summed.

Dimension reduction with constraints

The three methods defined above lend themselves in exactly the same way to include a second data matrix $\mathbf{X}$ $(I \times K)$ of K explanatory variables, continuous and/or categorical in the form of dummy variables, that serve to constrain the solution. The data matrix $\mathbf{Y}$ is then regarded as responses to these explanatory variables, or predictors. Suppose $\mathbf{X}$ is standardized, always taking into account the weights assigned to the rows, in other words the columns of $\mathbf{X}$ have weighted means zero and weighted variances 1. The matrix $\mathbf{A}$ is first projected onto the space of the explanatory variables:

$$\mathbf{A}_X = [\mathbf{X}(\mathbf{X}^T\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}_r]\mathbf{A} \quad (A.19)$$

and then the same three steps (A.13)–(A.15) are applied to $\mathbf{A}_X$ instead of $\mathbf{A}$, with the same options for the coordinates. This gives, respectively, redundancy analysis (PCA with constraints), canonical correspondence analysis (CCA, CA with constraints), and constrained log-ratio analysis. $\mathbf{A}_X$ is that part of the response data that is perfectly explained by the predictors. The matrix $\mathbf{A} - \mathbf{A}_X$ is the part of the response data that is uncorrelated with the predictors. If $\mathbf{X}$ includes variables that one wants to partial out, then $\mathbf{A} - \mathbf{A}_X$ is analysed using the same steps (A.13)–(A.15). In the case of CCA this is called *partial CCA*.

The total variance (or inertia) is now first partitioned into two parts, the part corresponding to the projected matrix $\mathbf{A}_X$, which is in the space of the predictors, and the part corresponding to $\mathbf{A} - \mathbf{A}_X$, which is in the space uncorrelated with the predictors. Otherwise, the computation of coordinates defined in (A.16)–(A.18) and the addition of supplementary variables and points follow in the same way.

Permutation testing and bootstrapping

The solutions obtained in all the multivariate analyses described in this book should be regarded as a complex *point estimate* – dendrograms and ordinations do not contain any information about the statistical significance of the results or whether the results would have been any different if the study were repeated in the same way. In order to perform hypothesis testing or to obtain intervals or regions of confidence, some standard multivariate tests exist for very special situations, which have quite restrictive assumptions, for example that data come from a multivariate normal distribution. We resort to computationally intensive methods to judge whether our solutions are nonrandom, reflecting some actual structure rather than random variation. In this book we have used permutation testing to obtain *p*-values associated with certain hypotheses, and bootstrapping to obtain measures of confidence, although this distinction is actually blurred (for example, one can do hypothesis testing using bootstrapping as well).

Permutation testing can be used for testing differences between groups. Under the null hypothesis that there is no inter-group difference, so that all the observations (e.g., sites) come from the same distribution, we can randomly assign the group labels to the observations and measure the inter-group difference by some reasonable statistic, such as the between-group sum of squares in multivariate space. Doing this a large number of times, obtaining a large number – say 9,999 – of values of the statistic, which defines its null distribution. Then, we see where the actual inter-group measure (in this case, the 10,000th) lies on this distribution and the estimated *p*-value is the proportion of all 10,000 values equal to or more

extreme than this value. The actual value is included in this proportion, so the smallest *p*-value obtainable would be 1/10,000 = 0.0001 in this case.

Permutation testing of inter-variable associations proceeds differently. In the case of a CCA, for example, there are two sets of variables, the response set and the explanatory set. We can measure how much inertia of the response data **Y** is explained by the explanatory data in **X** – this is the constrained inertia contained in the matrix $\mathbf{A}_X$ defined above. The null hypothesis is that there is no association, in which case every set of observations on the explanatory variables could be paired with any set of observations on the responses. So we randomize the order of one set of data, for example the rows of the explanatory data **X**, each time computing the amount of response inertia (or proportion) explained, doing this again thousands of times. The actual value of inertia explained is compared to the right tail of the null distribution to estimate the *p*-value.

Permutation testing can be used to give a guideline about the level at which a dendrogram should be cut to obtain significant clustering. Our approach has been to randomize the values within each column of data, that is shuffle them up randomly, assuming the columns contain the variables of the data, and recomputed the dendrogram each time. The node levels are stored for each dendrogram computed and this gives an idea of the null distribution of each level for data where there is no structure between the variables. The node levels of the actual dendrogram computed on the original data are then compared to these null distributions to obtain a *p*-value for each node. Here we are looking for values in the left tail of the respective null distributions, because significant clustering would be when the node levels are generally low in value. There can be several significant *p*-values in this case, and the final choice is based on these, substantive knowledge and the number of groups being sought.

Permutation testing can be similarly used for deciding on the dimensionality of the ordination solution. The columns of data are similarly randomized, each giving new parts of variance on the recomputed dimensions. This is done thousands of times, generating a null distribution of the parts of variance for the first dimension, second dimension, and so on. The original parts of inertia are compared to their corresponding null distributions to estimate a *p*-value for each dimension. In this case, *p*-values will generally increase for successive dimensions, and an obvious cut-off will appear, which usually coincides with the rule of thumb based on the *scree plot* of the eigenvalues.

To illustrate the use of bootstrapping for this last example, suppose we want a confidence region around the percentages of variance in a PCA, CA or LRA. The *I* rows of the data matrix are sampled, with replacement, until we have

a bootstrap sample, also of I rows. This means that some rows can be chosen more than once, others not at all – this differs from permutation testing where observations are simply re-arranged in a random order. For each bootstrap sample the multivariate method is recomputed and the percentages of inertia stored, and this is repeated thousands of times. This procedure results in an estimated distribution of percentages of inertia for each dimension, and a 95% confidence interval for each can be determined by cutting off 2.5% of the values on either tail of the distribution.

Statistical modelling

In the situation where we relate a single response to a set of explanatory variables, regular statistical modelling can be applied. Generalized linear modelling, generalized additive modelling and classification and regression trees, are alternative ways to model this relationship.

The most restrictive is generalized linear modelling (GLM), since it assumes that the effects of the explanatory variables are linear. But the way the linear effect translates to a change in the conditional mean of the response, called the *link function*, is different depending on the measurement scale of the response. The three most common types of responses are interval-scale continuous, ratio-scale count, and categorical binary:

Response variable	*Link function*	*Conditional distribution*	*Name of method*
Continuous:	Identity	Normal	Multiple linear regression
Count:	Logarithm	Poisson	Poisson regression
Categorical (binary):	Logit (log-odds)	Binomial	Logistic regression

The formulation of a generalized linear model is:

$$\eta(\overline{y}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots \tag{A.20}$$

with inverse transformation

$$\overline{y} = \eta^{-1}(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots) \tag{A.21}$$

where η is the link function and the conditional distribution of the response is the one corresponding to it, with mean given by (A.21). The inverse function η^{-1} is $\exp(\cdot)$ for $\eta = \log$ and $\exp(\cdot)/[1 + \exp(\cdot)]$ for $\eta = \text{logit}$.

Generalized additive modelling (GAM) is like GLM, but with a freer and more flexible range of possibilities for the shape of the relationship between the

response and each explanatory variable. The linear model on the right of (A.20) is replaced by a sum of smooth terms: $\alpha + s(x_1) + s(x_2) + \cdots$. Each smooth function $s(\cdot)$ is quite general, and involves tying together several cubic functions called a *smoothing spline*. These functions have estimated degrees of freedom and their form can either confirm approximate linearity of the relationship or suggest a transformation of the explanatory variables to accommodate a nonlinear relationship.

Both GLMs and GAMs can include interactions between the explanatory variables. Classification and regression trees (CART) form an alternative nonparametric approach that uses simple rules for predicting the response by cutting up the range of the predictors, but specifically looking for interactions in the form of combinations of intervals of the predictors which maximize the fit to the response. The result is a decision tree that allows every case to be run down it, according to the conditions at each node, to arrive at a terminal node that predicts the response, either the mean or median for a continuous response, or a set of probabilities for a categorical response that lead to the prediction of the most likely category.

Appendix B

Bibliography and Web Resources

This appendix lists various bibliographical resources, with short annotations, for further reading. In addition, some web resources are given for supporting information and material such as R software and tutorials.

Contents

Study design and data analysis

Anderson D.R., K.P. Burnham, W.R. Gould, and S. Cherry. "Concerns about finding effects that are actually spurious". *Wildlife Society Bulletin* 29 (2001): 311-316.

(The authors discuss the characteristics of research studies that are more exposed to the risk of finding spurious effects, and propose various ways to cope with the problem and avoid spurious results).

Beninger P.G., I. Boldina, and S. Katsakenevakis. "Strengthening statistical usage in marine ecology". *Journal of Experimental Marine Biology and Ecology* 426-427 (2012): 97-108.

(A review of common statistical fallacies in the ecological literature and how to avoid them).

Cottingham K.L., J.T. Lennon, and B.L. Brown. "Knowing when to draw the line: designing more informative ecological experiments". *Frontiers in Ecology and the Environment* 3 (2005): 145-152.

(Review of experimental design options for ANOVA and regression types of ecological studies).

Day R.W., and G.P. Quinn. "Comparisons of treatments after an analysis of variance in ecology". *Ecological Monographs* 59 (1989): 433-463.

(A review of approaches for comparisons of treatments following an ANOVA, including parametric and nonparametric tests, with discussion of pitfalls and solutions when dealing with hypothesis testing under unplanned multiple comparisons).

Graham M.H., and M.S. Edwards. "Statistical significance versus fit: estimating the importance of individual factors in ecological analysis of variance". *Oikos* 93 (2001): 503-515.

(The importance of effect size estimation and the available tools for variance decomposition in the context of complex ANOVA designs are presented clearly and succinctly).

Maindonald J. *The design of research studies – A statistical perspective. Part I: planning and reporting*, 2000, 120 p.

https://digitalcollections.anu.edu.au/bitstream/1885/41533/2/GS00_2.pdf

(A very informative introduction to the design of experimental and observational studies).

Nakagawa S., and I.C. Cuthill. "Effect size, confidence interval and statistical significance: a practical guide for biologists". *Biological Reviews* 82 (2007): 591-605.

(An indispensable review of effect size estimation. R code that allows to perform the analyses discussed in the paper is available and can be downloaded at:

http://www.bristol.ac.uk/biology/research/staff/cuthill.i).

Nakagawa S., and R.P. Freckleton. "Missing inaction: the dangers of ignoring missing data". *Trends in Ecology and Evolution* 23: 592-596.

(The authors warn against deletion of cases with missing observations due to the ensuing reduced statistical power and increased estimation bias, and provide a compact review of how to deal properly with missing data).

Parkhurst D.F. "Statistical significance tests: equivalence and reverse tests should reduce misinterpretation". *Bioscience* 51 (2001): 1051-1057.

(A gentle introduction to the concepts and methods of equivalence and reverse testing to help avoid pitfalls of results interpretation in classical null statistical hypothesis testing).

Quinn G., and M. Keough. *Experimental Design and Data Analysis for Biologists.* Cambridge, UK: Cambridge University Press, 2002.

(Introductory textbook to study design and data analysis. Popular in courses at undergraduate and graduate level in experimental study design and ecological statistics. Useful background material to refresh ideas while reading Multivariate Analysis of Ecological Data*).*

Regan H.M., M. Colivan, and M.A. Burgman. "A taxonomy and treatment of uncertainty for ecology and conservation biology". *Ecological Applications* 12 (2002): 618-628.

(A thorough discussion of sources of uncertainty in ecology and how to deal with them).

Scheiner S.M., and J. Gurevitch. *Design and Analysis of Ecological Experiments.* Oxford: Oxford University Press, 2001.

(A valuable collection of chapters by several authors dealing with study design, statistical modelling, spatial data analysis and meta-analysis).

Warton, D.I., and F.K.C. Hui. "The arcsine is asinine: the analysis of proportions in ecology". *Ecology* 92 (2011): 3-10.

(A brief review of useful transformations for proportions with some warnings against established traditions when dealing with this type of data).

Statistical modelling

Bolker B.M. *Ecological Models and Data in R*. Princeton, New Jersey: Princeton University Press, 2008.

(A gentle introduction to ecological modelling with clear and well structured coverage of maximum likelihood models and estimation).

Bolker B.M., M.E. Brooks, C.J. Clark, S.W. Geange, J.R. Poulsen, M.H. Stevens, and J.S. White. "Generalized linear mixed models: a practical guide for ecology and evolution". *Trends in Ecology and Evolution* 24 (2009): 127-135.

(The paper reviews how to deal with nonnormal data that include random effects with the help of generalized linear mixed models).

Clark J.S. *Models for Ecological Data: an Introduction.* Princeton, New Jersey: Princeton University Press, 2007.

(Rigorous and rich introduction to statistical modelling, including approaches to temporal and spatial data. A companion lab manual provides examples using R).

Grueber C.E., S. Nakagawa, R.L. Laws, and I.G. Jamieson. "Multimodel inference in ecology and evolution: challenges and solutions". *Journal of Evolutionary Biology* 24 (2011): 699-711.

(A comprehensive review of model selection and multimodel inference introducing basic concepts and approaches in a clear and balanced way).

Hilborn R., and C. Mangel. *The Ecological Detective: Confronting Models with Data.* Princeton, New Jersey: Princeton University Press, 1997.

(The book provides a very good introduction to theoretical and statistical modelling in ecology, explaining concepts, principles and protocols to the uninitiated).

Hobbs N.T., and R. Hilborn. "Alternatives to statistical hypothesis testing in ecology: a guide to self teaching". *Ecological Applications* 16 (2006): 5-19.

(A clear and concise introduction to statistical modelling, maximum likelihood estimation, model selection, Bayesian analysis and meta-analysis).

Stephens P.A., S.W. Buskirk, and C.M. del Rio. "Inference in ecology and evolution". *Trends in Ecology and Evolution* 22 (2007): 192-197.

(The authors discuss the limitations of traditional null hypothesis significance tests and suggest to rely on more useful approaches, briefly reviewed, such as effect size estimation and model selection).

Multivariate analysis

Anderson M.J. "Permutation tests for univariate or multivariate analysis of variance and regression". *Canadian Journal of Fisheries and Aquatic Sciences* 58 (2001): 626-639.

(An informative and concise review of rational and applications of permutation tests in experimental and observational studies with complex designs).

Anderson M.J., and T.J. Willis. "Canonical analysis of principal coordinates: a useful method of constrained ordination for ecology". *Ecology* 84 (2003): 511-525.

(A flexible method for constrained ordination capable of accommodating any distance or dissimilarity matrix).

Beals E.W. "Bray-Curtis ordination: an effective strategy for analysis of multivariate ecological data". *Advances in Ecological Research* 14 (1984): 1-55.

(A good introduction to Bray-Curtis (or polar) ordination, also covering other methods).

Borcard D., F. Gillet, and P. Legendre. *Numerical Ecology with R*. New York: Springer, 2011.

(Compact introduction to multivariate statistics, including multivariate analysis of spatial and temporal data. It is the R companion to Numerical Ecology by Legendre and Legendre, 2012).

Gauch H. G. Jr. *Multivariate Analysis in Community Ecology*. Cambridge, United Kingdom: Cambridge University Press, 1982.

(A thorough introduction to gradient analysis, it relates ecological theory and statistical methods clarifying the rational behind the approach).

Greenacre M.J. *Correspondence Analysis in Practice, 2nd Edition*. London: Chapman & Hall/CRC, 2007. Free download of the Spanish edition, published by the BBVA Foundation, 2008, at www.multivariatestatistics.org.

(Comprehensive introduction to correspondence analysis, multiple correspondence analysis, subset correspondence analysis and canonical correspondence analysis).

Greenacre M.J. 2010. *Biplots in Practice*. Madrid: BBVA Foundation, 2010. Free download from www.multivariatestatistics.org.

(A practical introduction to biplots, the concept underlying many multivariate methods that reduce dimensionality in large data sets, and visualize the results).

Greenacre M.J. "Correspondence analysis of raw data". *Ecology* 91 (2010): 958-963.

(Alternative approach to analysing abundance or biomass matrices where the data are not expressed relative to the row and column margins, in contrast to regular CA and CCA where relative amounts are analysed).

Greenacre M.J. "The contributions of rare objects in correspondence analysis". *Ecology* 94 (2013): 241-249.

(Shows that CA and CCA are not unduly affected by the presence of rare species in an ecological data set, contrary to a popular misconception that these analyses are over-sensitive to species that occur sparsely and in low abundance).

Greenacre M.J. "Fuzzy coding in constrained ordinations". *Ecology* 94 (2013): 280-286.

(The use of fuzzy coding for explanatory variables in the CCA context, demonstrating the benefits and also how to choose the number of fuzzy categories).

Greenacre M.J. "Contribution biplots". *Journal of Computational and Graphical Statistics* 22 (2013): 107-122.

(An alternative scaling of the results of ordination methods such as PCA, CA, LRA, CCA and RDA, where the variables that contribute most to the solution are immediately detectable in the ordination).

Greenacre, M.J., and P.J. Lewi. "Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements". *Journal of Classification* 26 (2009): 29-54.

(Demonstrates clearly the advantage of weighting of variables in log-ratio analysis, as is done in regular CA, as well as the ability of log-ratio biplots to diagnose multiplicative models when variables line up in the ordinations).

Jackson D.A. “Stopping rules in principal component analysis: a comparison of heuristical and statistical approaches”. *Ecology* 74 (1993): 2204-2214.

(A short introduction to some available options for evaluating the significance of principal components).

James F.C. “Multivariate analysis in ecology and systematics: panacea or pandora’s box?”. *Annual Review in Ecology and Systematics* 21 (1990): 129–166.

(An early review of multivariate statistical applications in ecology, at a time when increased computer and software availability made these methods available to all ecologists).

Johnson, R.A., and D.W. Wichern. *Applied Multivariate Statistical Analysis, 6th edition*. New Jersey: Prentice Hall, 2007.

(Widely read book reviewing applications of multivariate methods for biologists, physicists and sociologists).

Jongman R.H.G., C.J.F. Ter Braak, and O.F.R Van Tongeren. *Data Analysis in Community and Landscape Ecology*. Cambridge, United Kingdom: Cambridge University Press, 1995.

(A balanced treatment of ecological data analysis including extensive treatment of multivariate methods by a group of statisticians and ecologists with a strong quantitative background).

Legendre P., and M.J. Anderson. “Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments”. *Ecological Monographs* 69 (1999): 1-24.

(Introduction to the flexible tool of distance-based redundancy analysis by the authors that eventually developed and generalized further this useful numerical approach).

Legendre P., and E.D. Gallagher. “Ecologically meaningful transformations for ordination of species data”. *Oecologia* 129 (2001): 271-280.

(Review of effective transformation for species data that allow to extract relevant information when subject to ordination analysis).

Legendre P., and L. Legendre. *Numerical Ecology, 3rd English edition*. Amsterdam: Elsevier, 2012, 853 p.

(Classic introduction to statistics for ecologists with very good coverage of ecological data and multivariate methods, including an exhaustive treatment of distance and dissimilarity measures. It has a R companion [Borcard et al. 2011]).

Lepš J., and P. Šmilauer. *Multivariate Analysis of Ecological Data using Canoco*. Cambridge United Kingdom: Cambridge University Press, 2003.

(Much more than a handbook for Canoco applications, the book is an informative review of multivariate methods with many inspiring ecological examples).

Maindonald J., and J. Braun. *Data Analysis and Graphics with R. An Example Based Approach, 3rd edition.* Cambridge, United Kingdom: Cambridge University Press, 2011.

(The book provides a comprehensive overview of data analysis including parametric and nonparametric methods, statistical modelling and multivariate methods with R examples).

Manly B.F.J. *Multivariate Statistical Methods: a Primer, 3rd edition.* London: Chapman and Hall, 2004.

(A gentle introduction to multivariate methods blessed by the clear expository style of a distinguished and successful author).

Manly B.F.J. *Randomization, Bootstrap and Monte Carlo Methods in Biology, 3rd edition.* London: Chapman and Hall, 2007.

(The book provides a comprehensive overview of resampling and permutation methods with many relevant biological example applications).

McGarigal K., S. Cushman, and S. Stafford. *Multivariate Statistics for Wildlife and Ecology Research.* New York: Springer, 2000.

(Introduction to multivariate statics in ecology and wildlife management, focusing on practical applications).

Palmer M.W. "Putting things in even better order: the advantages of canonical correspondence analysis". *Ecology* 74 (1993): 2215-2230.

(A clear exposition of the advantages of canonical correspondence analysis (CCA) applied to ecological data).

Peres-Neto P.R. "How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test". *Oecologia* 129 (2001): 169-178.

(The use and value of Procrustean superimposition to compare (match) multivariate data sets).

Peres-Neto P.R., and D.A. Jackson. "The importance of scaling of multivariate analysis in ecological studies". *Ecoscience* 8 (2001): 522-526.

(A clear survey of the role of scaling in multivariate ecological data analysis, making use of intuitive graphical presentations to stress the important concepts and their implications).

Peres-Neto P.R., D.A. Jackson, and K.M. Somers. "Giving meaningful interpretation to ordination axes: assessing loading significance in principal component analysis". *Ecology* 84 (2003): 2347-2363.

(The authors compare a variety of approaches for assessing the significance of eigenvector coefficients in terms of type I error rates and power).

Pielou, E.C. *The Interpretation of Ecological Data: a Primer on Classification and Ordination.* New York: John Wiley & Sons, Inc., 1984.

(An early review of ecological data analysis linking ecological and statistical concepts to ease the interpretation of results).

Ter Braak, C.J.F. "Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis". *Ecology* 67 (1986): 1167-1179.

(Another citation classic for an author that has made important contributions to the field, ensuring a wide availability of the new methods via software development (Canoco, in collaboration with Šmilauer – see Lepš and Šmilauer 2003)).

Ter Braak, C.J.F., and I.C. Prentice. "A theory of gradient analysis". *Advances in Ecological Research* 18 (1988): 271-313.

(A classic introduction to gradient analysis theory and its ecological applications).

Wiedmann, M., M. Aschan, G. Certain, A. Dolgov, M. Greenacre, E. Johannesen, B. Planque, and R. Primicerio. "Functional diversity of the Barents Sea fish community". *Marine Ecology Progress Series,* in press, doi: 10.3354/meps10558.

(A more extensive analysis of functional traits of Barents Sea fish species, compared to our case study of Chapter 20, using more fish species and a more recent time series).

Yee T.W. "Constrained additive ordination". *Ecology* 87 (2006): 203-213.

(The paper introduces constrained additive ordination (CAO) models, described as "loosely speaking, […] generalized additive models fitted to a very small number of latent variables". The paper provides the R code to implement the CAO methodology with some clear example applications).

Zuur A.F., E.N. Ieno, and C.S. Elphick. "A protocol for data exploration to avoid common statistical problems". *Methods in Ecology and Evolution* 1 (2010): 3-14.

(The authors provide a protocol for data exploration, discussing "current tools to detect outliers, heterogeneity of variance, collinearity, dependence of observations, problems with interactions, double zeros in multivariate analysis, zero inflation in generalized linear modelling, and the correct type of relationships between dependent and independent variables; and []

provide advice on how to address these problems when they arise". The paper also provides R code to implement the protocol).

Zuur A.F., E.N. Ieno, N.J. Walker, A.A. Saveliev, and G.M. Smith. *Mixed Effects Models and Extensions in Ecology with R*. New York: Springer, 2009.

(A very popular introduction to mixed effects models rich with relevant ecological examples based on the kind of "messy" data ecologists need to cope with).

Web resources

http://www.multivariatestatistics.org

Greenacre M.J. *Multivariate books, data sets and R scripts.*

(This web site supports three books published by the BBVA Foundation: La Práctica del Análisis de Correspondencias *(Spanish translation of* Correspondence Analysis in Practice*, Second Edition),* Biplots in Practice *and the present book. Full text, data sets and R scripts available for free download).*

http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf

Oksanen J. *Multivariate analyses of ecological communities in R: vegan tutorial.* 43 p.

(This compact tutorial to the R package vegan is a very useful, brief introduction to the application of multivariate statistics in community ecology).

http://ordination.okstate.edu

Palmer M.W. *Ordination methods for ecologists*

(The web site is a very useful overview of ordination methods, providing links, references, and guides for self study, including a much needed glossary).

APPENDIX C

Computational Note

This appendix is a short summary of the software used for the analyses in this book, using packages from the R environment for statistical computing and graphics. Data sets and R code for reproducing the results are given online at the supporting website:

www.multivariatestatistics.org

As an introduction to the online code, we give here a list of some of the common R functions and packages used in the computations of this book.

Contents

Functions from the **base** package in R

The open-access R software has become the standard for statistical computing, especially for conducting research, thanks to its flexible programming environment. It is downloadable for free from the R project website

www.r-project.org

The simple installation process sets up R with what is called the **base** package, consisting of various functions that are commonly used (later we list more specialized packages that need to be downloaded and installed separately). Here we list some of the useful functions in the **base** package:

- `hist` – takes a set of data on a continuous or count variable and makes a histogram; user can choose the interval boundaries; see Exhibits 1.2 and 1.3, for example.

- `qqnorm` – takes a set of data on a continuous variable and plots the sample quantiles against the quantiles of a normal distribution; if the points follow the 45-degree diagonal line of the plot, the data can be regarded as normal, otherwise not; see Chapter 17.

- `shapiro.wilks` – takes a set of data on a continuous variable and performs the Shapiro-Wilks test for normality; if the *p*-value is small then normality is rejected; see Chapter 17.

- `pairs` – takes a rectangular data matrix as input and computes all bivariate scatterplots; see Exhibits 1.4 and 20.8.

- `boxplot` – takes a set of data on a continuous variable and makes a box-and-whisker plot, optionally with a categorical variable that makes boxplots for each category alongside one another, with a common scale; see Exhibits 1.5 and 1.8.

- `scale` – takes a set of data on a continuous variable and standardizes it by subtracting its mean (i.e., centering) and dividing by its standard deviation (i.e., normalization); centering or normalization can be switched off; see Chapter 3.

- `dist` – takes a rectangular data matrix as input and computes a distance matrix between the rows, with several choices of distance functions; for example, see Exhibit 4.5.

- `cor` – takes either two sets of data or a matrix of data with variables in columns and computes the single correlation in the former case, or the correlation matrix in the latter case; optionally computes Spearman rank correlations; see Exhibit 6.4.

- `table` – takes a single categorical and counts the frequencies in each category; if two categorical variables are given the function counts the frequencies in the cross-tabulation; see Exhibit 6.6.

- `sample` – takes a set of data and performs random sampling, without replacement (this re-arranges, or shuffles, the data set randomly) for permutation testing or with replacement for bootstrapping; see Chapter 18.

- `hclust` – takes a matrix of distance or dissimilarities (e.g., created by function `dist`) and performs hierarchical clustering; various clustering algorithms can be selected, including Ward clustering; see Chapters 7 and 8.

- `kmeans` – takes a rectangular matrix of data and the specified number of groups and performs *k*-means nonhierarchical clustering; see Chapter 8.

- `cmdscale` – takes a matrix of distances or dissimilarities (e.g., created by function `dist`) and performs classical multidimensional scaling; see Chapter 9.

- `lm` – takes data on a response variable and one or more explanatory variables (or predictors) and performs least-squares linear regression; weights can be specified for weighted least-squares regression; see Chapters 10–20.

- `glm` – takes data on a response variable and one or more explanatory variables (or predictors) and performs generalized linear modelling (GLM); several link functions and error distributions can be specified, giving linear regression, Poisson regression and logistic regression, for example; see Chapters 10 and 18.

- `prcomp` and `princomp` – alternative functions for computing a principal component analysis on a rectangular data matrix, where rows are assumed to be sampling units and columns to be variables; see Chapter 12.

- `kruskal.test` – takes a data set for a continuous variable and a grouping variable and performs the Kruskal-Wallis rank test of difference between groups (the nonparametric equivalent of a one-way ANOVA); see Chapter 17.

Package **ca**

The **ca** package performs correspondence analysis (function `ca`) and multiple correspondence analysis (function `mjca` – this generalization of CA to multivariate categorical data, more used in the social sciences, is not discussed in this book). Various graphical options are available using function `plot.ca`, including plotting with contribution coordinates and three-dimensional visualization of a CA solution with three principal axes, using function `plot3d.ca`, including interaction with 3d display such as rotation and zooming. The 3d graphics uses the R package **rgl**; see Chapter 13.

Package **vegan**

The **vegan** package performs a variety of multivariate analyses and includes most of the methods treated in this book, and aimed at biologists (specifically botanists, but the terminology can be equated to any biological application). Methods that are not included in R's base package described above are computation of Bray-Curtis and Gower dissimilarities (function `vegdist` with options `method="bray"` or `method="gower"` respectively), various diversity measures

(function `diversity`), nonmetric multidimensional scaling (function `metaMDS`), canonical correspondence analysis (function `cca`), redundancy analysis (function `rda` – like `cca` but for continuous response variables) and various permutation tests (e.g., function `permutest`); see Chapters 15 and 20.

Packages **maptools**, **mapdata** and **mapproj**

Packages **maptools** and **mapdata** provide functions that allow drawing of geographical maps, with **mapdata** containing the outlines of all the world's landmasses and several countries. The **mapproj** package performs a variety of map projections, using function `mapproject`, based on the latitude and longitude coordinates of a set of spatial locations. This is useful for obtaining coordinates on which Euclidean distances can be computed that approximate great circle distances; see Chapters 11 and 19.

Package **mgcv**

This package performs generalized additive modelling (GAM), using a function `gam` that functions very similarly to `glm` for generalized linear modelling. Explanatory variables can be defined as smooth functions using the function `s`, for example `s(x)` for predictor `x`; see Chapters 18, 19 and 20.

Packages **rpart** and **tree**

These packages are alternatives for classification and regression trees, also called *recursive partitioning* (hence **rpart**). They define tree models in the same style as functions `lm`, `glm` and `gam`, as a response variable ~ sum of explanatory variables. Plotting the result using `plot` gives the tree plot; see Chapter 18.

Additional functions in supporting material

Several additional functions that are used in our applications are given in the supporting material on www.multivariatestatistics.org.

- `fuzzy.tri` – takes a set of data on a continuous variable, with a specified number of categories, and transforms to fuzzy categories using triangular membership functions; hinges are by default defined as quantiles, but can be supplied by the user; see Exhibit 3.3.

- `chidist` – takes a rectangular matrix of same-scale nonnegative data as input and computes the matrix of chi-square distances between rows or between columns; see Exhibit 4.7.

- `jaccard` – takes a rectangular matrix of presence-absence data (ones and zeros) and computes the matrix of Jaccard dissimilarities between rows or between columns (this can also be achieved in the **vegan** package using function `vegdist` with `method="jaccard"`); see Chapter 5 and Exhibit 7.1.

List of Exhibits

Index

Note: E after the page number indicates a reference to an Exhibit.

About the Authors

Michael Greenacre, Professor of Statistics at the Pompeu Fabra University in Barcelona and research collaborator with the BBVA Foundation, obtained his master degree in his country of birth, South Africa, and then his doctorate in Paris at the Pierre et Marie Curie University (Paris VI). He specialized in the visualization of multivariate data, principally in the social and environmental sciences, and spent sabbatical research periods at Rothamsted Experimental Station (UK); Bell Laboratories, University of Rochester and Stanford University (USA); the École des Mines (France); and the Norwegian Polar Environmental Centre (now the FRAM centre) in Tromsø (Norway). Besides co-editing several books on data visualization, he has written four books on correspondence analysis and related methods: for example, *Biplots in Practice* was published by the BBVA Foundation in 2010. He has produced over 80 publications, in major journals such as *Applied Statistics* and *Ecology*, including invited contributions to several encyclopedias.

Raul Primicerio, Associate Professor of Ecology, Evolutionary Biology and Epidemiology at the University of Tromsø, obtained his master degree in his country of birth, Italy, and later his doctorate in Norway. His research and teaching focus on quantitative biology, and he has been training graduate students and professionals at several research institutions in scientific method, statistical inference and modelling. He has coordinated ecological modelling activities at the High North Research Centre for Climate and the Environment (FRAM, Tromsø), and has collaborated and helped to coordinate several research projects on global environmental change impact funded by the Norwegian and European research councils. He has produced over 50 papers on both basic and applied issues, such as harvesting and climate change impact, including publications in the multidisciplinary journals *Science* and *PNAS*.