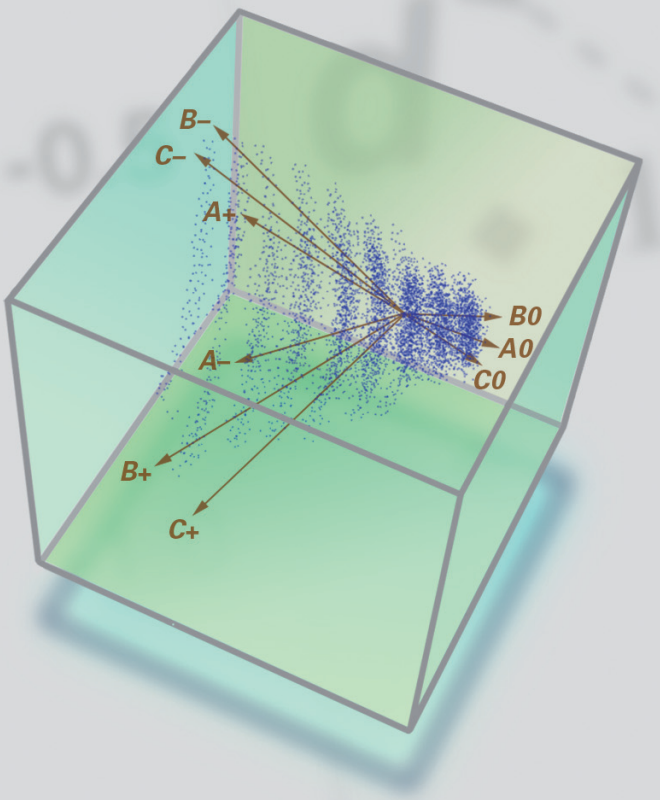
Michael Greenacre
Biplots in Practice



Fundación BBVA

Biplots in Practice

Biplots in Practice

Michael Greenacre

Fundación BBVA

First published September 2010

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

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

Edition and production: Rubes Editorial

ISBN: 978-84-923846-8-6
Legal deposit no: B-

Printed in Spain

Printed by S.A. de Litografía

This book is produced with paper in conformed with the environmental standards required by current European legislation.

In memory of Cas Troskie
(1936-2010)

Contents

Preface

Communicating and understanding data

In almost every area of research where numerical data are collected, databases and spreadsheets are being filled with tables of numbers and these data are being analyzed at various levels of statistical sophistication. Sometimes simple summary methods are used, such as calculating means and standard deviations of quantitative variables, or correlation coefficients between them, or counting category frequencies of discrete variables or frequencies in cross-tabulations. At the other end of the spectrum, advanced statistical modelling is performed, which depends on the researcher's preconceived ideas or hypotheses about the data, or often the analytical techniques that happen to be in the researcher's available software packages. These approaches, be they simple or advanced, generally convert the table of data into other numbers in an attempt to condense a lot of numerical data into a more palatable form, so that the substantive nature of the information can be understood and communicated. In the process, information is necessarily lost, but it is tacitly assumed that such information is of little or no relevance.

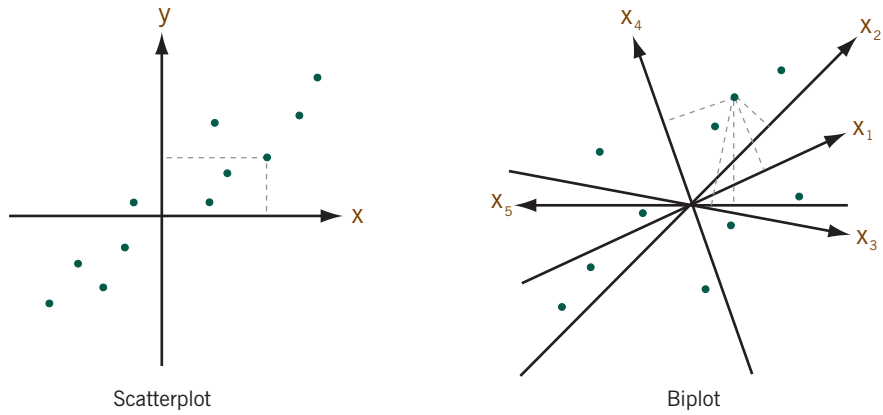
Graphical display of data

Graphical methods for understanding and interpreting data are another form of statistical data analysis; for example, a histogram of a quantitative variable or a bar chart of the categories of a discrete variable. These are usually much more informative than their corresponding numerical summaries—for a pair of quantitative variables, for example, a correlation is a very coarse summary of the data content, whereas a simple scatterplot of one variable against the other tells the whole story about the data. However, graphical representations appear to be limited in their ability to display all the data in large tables at the same time, where many variables are interacting with one another.

The biplot: a scatterplot of many variables

This book deals with an approach to statistical graphics for large tables of data which is intended to pack as much of the data content as possible into an easily digestible display. This methodology, called the *biplot*, is a generalization of the scatterplot of two variables to the case of many variables. While a simple scatterplot of two variables has two perpendicular axes, conventionally dubbed the horizontal x-axis and vertical y-axis, biplots have as many axes as there are variables, and these can take any orientation in the display (see Exhibit 0.1). In a scatterplot the cases, represented by points, can be projected perpendicularly onto the axes to read off their values on the two variables, and similarly in a biplot the cases

Exhibit 0.1:
A simple scatterplot of two variables, and a biplot of many variables. Green dots represent "cases" and axes represent "variables", labelled in brown



can be projected perpendicularly onto all of the axes to read off their values on all the variables, as shown in Exhibit 0.1. But, whereas in a scatterplot the values of the variables can be read off exactly, this is generally impossible in a biplot, where they will be represented only approximately. In fact, the biplot display is in a space of *reduced dimensionality*, usually two-dimensional, compared to the true dimensionality of the data. The biplot capitalizes on correlations between variables in reducing the dimensionality—for example, variables x and y in the scatterplot of Exhibit 0.1 appear to have high positive correlation and would be represented in a biplot in approximately the same orientation, like x_1 and x_2 in the biplot of Exhibit 0.1, where projections of the points onto these two axes would give a similar lining-up of the data values. On the other hand, the observation that variables x_3 and x_5 point in opposite directions probably indicates a high negative correlation. The analytical part of the biplot is then to find what the configuration of points and the orientations of the axes should be in this reduced space in order to approximate the data as closely as possible.

This book fills in all the details, both theoretical and practical, about this highly useful idea in data visualization and answers the following questions:

- How are the case points positioned in the display?
- How are the different directions of the variable axes determined?
- In what sense is the biplot an optimal representation of the data and what part (and how much) of the data is not displayed?
- How is the biplot interpreted?

Furthermore, these questions are answered for a variety of data types: quantitative data on interval and ratio scales, count data, frequency data, zero-one data and multi-category discrete data. It is the distinction between these various data types that defines most chapters of the book.

Summary of contents

As in my book *Correspondence Analysis in Practice* (2nd edition), this book is divided into short chapters for ease of self-learning or for teaching. It is written with a didactic purpose and is aimed at the widest possible audience in all research areas where data are collected in tabular form.

After an introduction to the basic idea of biplots in Chapter 1, Chapters 2 and 3 treat the situation when there is an existing scatterplot of cases on axes defined by a pair of "explanatory" variables, onto which other variables, regarded as "responses" are added based on regression or generalized linear models. Here readers will find what is perhaps the most illuminating property of biplots, namely that an arrow representing a variable is actually indicating a regression plane and points in the direction of steepest ascent on that plane, with its contours, or isolines, at right-angles to this direction.

Chapter 4 treats the positioning of points in a display by multidimensional scaling (MDS), introducing the concept of a space in which the points lie according to a measure of their interpoint distances. Variables are then added to the configuration of points using the regression step explained before.

Chapter 5 is a more technical chapter in which the fundamental result underlying the theory and computation of biplots is explained: the singular value decomposition, or SVD. This decomposition provides the coordinates of the points and vectors in a biplot with respect to dimensions that are ordered from the most to the least important, so that we can select the reduced-dimensional space of our choice (usually two- or three-dimensional) that retains the major part of the original data.

Chapter 6 explains and illustrates the simplest version of the biplot of a cases-by-variables matrix in the context of principal component analysis (PCA), where the variables are measured on what are considered to be interval scales. Here the issue of biplot scaling is treated for the first time.

Chapter 7 deals with the lesser-known topic of log-ratio analysis (LRA), which deserves much more attention by data analysts. The variables in this case are all measured on the same scale, which is a positive, multiplicative scale, also called a ratio scale. All pairs of ratios within rows and within columns are of interest, on a logarithmic scale. The LRA biplot shows the rows and the columns as points, but it is really the vectors connecting pairs of rows or pairs of columns that are interpreted, since these link vectors depict the log-ratios.

Chapter 8 treats biplots in correspondence analysis (CA), which competes with log-ratio analysis as a method for analyzing data on a common ratio scale, espe-

cially count data. In contrast to LRA, CA easily handles zero values, which are common in many research areas where count data are collected, for example linguistics, archaeology and ecology.

Chapters 9 and 10 consider biplots in the display of large sets of multivariate categorical data, typically from questionnaire surveys, using variations of CA. Here there are two approaches: one is to consider associations between two different sets of variables, which are cross-tabulated and then concatenated (Chapter 9), while the other considers associations within a single set of variables (Chapter 10) —the latter case, called multiple correspondence analysis (MCA), has become one of the standard tools for interpreting survey data in the social sciences.

Chapter 11 focuses on discriminant analysis (DA) of grouped data, where the biplot displays group differences rather than differences between individual cases. Each group is represented by its mean point, or centroid, and it is these centroids that are optimally displayed in the biplot, along with the variables that contribute to their separation.

Chapter 12 is a variant of the dimension-reduction theme where the optimal reduced space is obtained subject to constraints on the solution in terms of additional explanatory variables. This idea has applications in all the versions of the biplot treated in the book: MDS, PCA, LRA, CA, MCA and DA. In the PCA context this constrained form is known as redundancy analysis (RDA) and in the CA context as canonical correspondence analysis (CCA).

Throughout the book there are illustrations of biplots in many different research contexts. It then concludes with three more detailed case studies in the biomedical, social and environmental sciences respectively:

- Analysis of a large data set in cancer research based on gene-expression arrays —using the PCA biplot and the DA biplot (or centroid biplot) to distinguish between four cancer types (Chapter 13).
- Analysis of data on several thousand respondents in a questionnaire survey on attitudes to working women—using CA and MCA biplots, and also constrained biplots to study the effect of different response categories (Chapter 14).
- Analysis of morphological and diet data of a sample of fish, to identify possible components of the diet that are related to the fish's morphology—using the LRA biplot with constraints (Chapter 15).

Computations in the R language

The biplots reported and discussed in the book are all computed in the open-source R environment and the Computational Appendix explains several of these analyses by commenting on the R code used.

Finally, the book concludes with a bibliography for further reading and online resources, a glossary of terms, and an epilogue in which some of my personal opinions are expressed about this area of statistics.

Readers will find a supporting website for this book at:

http://www.fbbva.es
http://www.multivariatestatistics.org

containing additional material such as the glossary and summaries of the material in Spanish, and the complete script file of the R code.

Dedication to Cas Troskie

This book is appropriately dedicated to Prof. Cas Troskie, former head of the Department of Statistical Sciences at the University of Cape Town (UCT), South Africa, and a maestro of theoretical and applied multivariate analysis. Cas was one of the most influential people in my statistical career. In fact when he visited me in Barcelona on several occasions I always introduced him as the reason behind my decision to do statistics as a major in my initial Bachelor of Science studies at UCT. As early as 1969, aged 33 and the youngest department head on the UCT campus, he was encouraging students like myself to write computer programs and put decks of punched cards into card readers linked to the university computer and wait expectantly for printouts to emerge with the results. He had a singular faith in principal components of a data set, which prepared me for my subsequent studies in France on correspondence analysis. I am not alone in being affected by his dynamic personality and sharp intelligence, since he inspired dozens of Masters and PhD theses, leaving a huge legacy to the statistical community, not only in South Africa but worldwide. One of his theoretical papers, co-authored with one of his PhD students, has been cited often in the electrical engineering literature and has made a significant impact in the design of MIMO (multiple input multiple output) wireless communications systems, which will form the cornerstone of most future wireless technologies.

Acknowledgements

This book owes its publishing to the BBVA Foundation and its Director, Prof. Rafael Pardo. One of the visions of the Foundation is to disseminate advanced educational material in a form that is easily accessible to students and researchers worldwide; hence this series of manuals on applicable research, attractively produced, distributed online for free and complemented by a supporting website with additional online material. For an academic it is like a dream come true to have such an outlet and I express my gratitude and appreciation to Prof. Pardo for including me in this wonderful project. Thanks are also due to the Foundation's publications director, Cathrin Scupin, for her continuing co-operation and support throughout the publishing process. Then there is the fantastic production team at Rubes Edi-

torial in Barcelona, Jaume Estruch, Núria Gibert and Imma Rullo, to whom I am equally grateful—they are responsible for the physical aspects of this book, expert copy-editing of the manuscript, and the design of the supporting website. Thanks are due to the Pompeu Fabra University in Barcelona, and for partial funding by the Spanish Ministry of Science and Technology grants MTM2008-00642 and MTM2009-09063. Finally, there are many friends who have supported me in this project—too many to list individually, but they know who they are!

So, if you have this book in your hands or are seeing this online, I wish you good reading, good learning and especially good biplotting!

Michael Greenacre
Barcelona, July 2010

CHAPTER 1

Biplots—the Basic Idea

The basic idea of the biplot is very simple, and like all simple solutions to complex problems it is both powerful and very useful. The biplot makes information in a table of data become transparent, revealing the main structures in the data in a methodical way, for example patterns of correlations between variables or similarities between the observations. Rectangular data matrices are the "raw material" in many research areas, existing in spreadsheets or databases. The rows of a data matrix are usually observed sampling units such as individuals, countries, demographic groups, locations, cases,..., and the columns are variables describing the rows, such as responses in a questionnaire, economic indicators, products purchased, environmental parameters, genetic markers, ... Throughout this book several data matrices from different areas of research will be used as illustrations of the power of the biplot to reveal the inherent structure in the data. In this initial chapter, we describe the basic geometric concepts that form both the foundation of the biplot's definition as well as its practical interpretation.

Contents

Scatterplots

A biplot is the generalization of the well-known *scatterplot* of observations on two variables. Exhibit 1.1 shows data for 12 European countries in 2008 on the following three variables: X_1 = purchasing power per capita (expressed in euros), X_2 = gross domestic product (GDP) per capita (indexed at 100 for all 27 countries in the European Union for 2008) and X_3 = inflation rate (percentage). Exhibit 1.2 shows two scatterplots, of X_1 versus X_2 and X_3 versus X_2 respectively. At

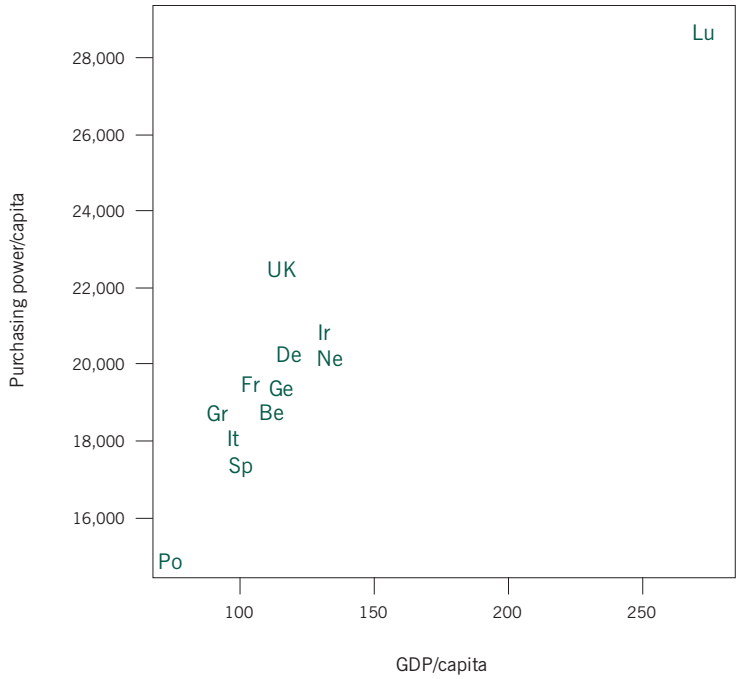
Exhibit 1.1: *Economic data for 12 European countries in 2008. X_1 = purchasing power per capita (expressed in euros), X_2 = gross domestic product (GDP) per capita (indexed at 100 for all 27 countries in the European Union for 2008) and X_3 = inflation rate (percentage)*

	Country	X_1	X_2	X_3
Be	Belgium	19,200	115.2	4.5
De	Denmark	20,400	120.1	3.6
Ge	Germany	19,500	115.6	2.8
Gr	Greece	18,800	94.3	4.2
Sp	Spain	17,600	102.6	4.1
Fr	France	19,600	108.0	3.2
Ir	Ireland	20,800	135.4	3.1
It	Italy	18,200	101.8	3.5
Lu	Luxembourg	28,800	276.4	4.1
Ne	Netherlands	20,400	134.0	2.2
Po	Portugal	15,000	76.0	2.7
UK	United Kingdom	22,600	116.2	3.6

a glance we can see in the first scatterplot that X_1 and X_2 are strongly correlated, as one would expect for these variables, whereas the second scatterplot shows the correlation to be weaker between X_3 and X_2. In this book we will be interested in visualizing the relationships between many variables, but let us consider for a start how we could see the inter-relationships between all three of these variables. One way is to use three-dimensional graphics, but we still have to resort to a planar representation of the points, as shown inside a cube in Exhibit 1.3. Two different views of the points are shown and the first view shows the countries lining up more or less in a linear spread, as indicated by the dashed line. We can capitalize on this by turning the cloud of points around so we look at it from a viewpoint perpendicular to this line of spread, as shown in the second view. The countries are now spread out more, and all that we miss in the second view is the small deviation of each country from the dashed line. In the second view of Exhibit 1.3 directions have been added, starting more or less at the middle of the cloud of country points and pointing in the directions of the variables as given by the sides of the cube. It is no coincidence that the two variables which were seen to be correlated turn out pointing in similar directions, since they were seen in the left hand scatterplot of Exhibit 1.2 to be lining up the countries in more or less the same order. What we have done in this informal example is reduce the three-dimensional example to a flat two-dimensional display, trying to lose as little of the spread of the countries in their original three-dimensional space as possible. The rest of this book formalizes this idea and extends it to showing clouds of points in high-dimensional spaces in a subspace of reduced dimensionality.

A simple example: a matrix expressed as the product of two other matrices

To introduce the biplot in a very simple way, consider the following equality between a 5×4 matrix on the left-hand side and the product of two matrices, 5×2 and 2×4 respectively:



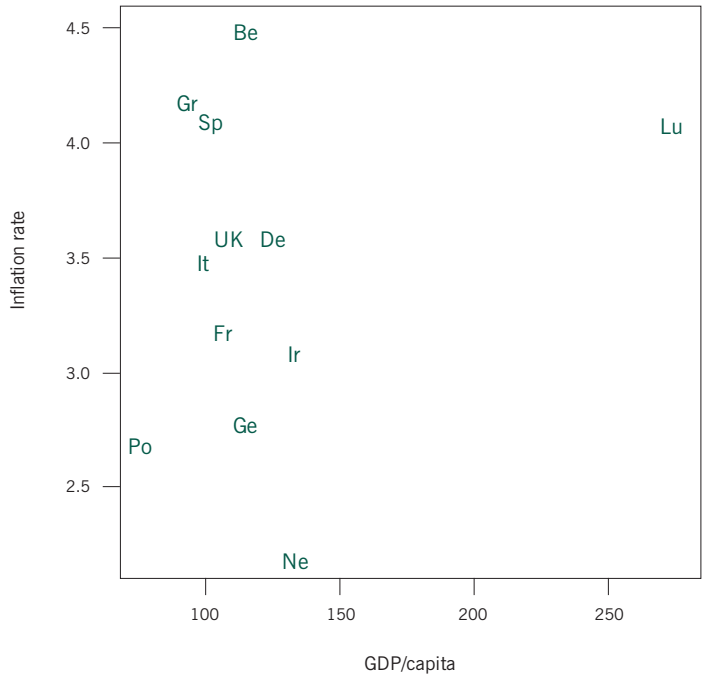
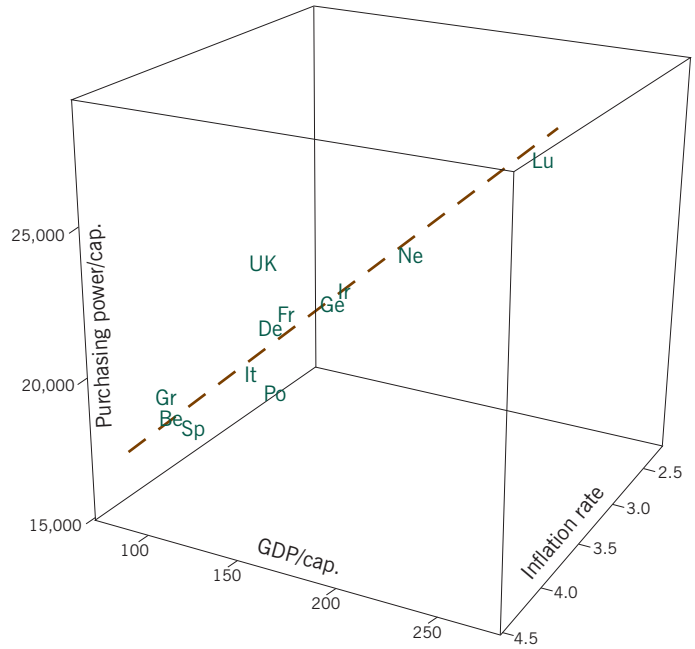
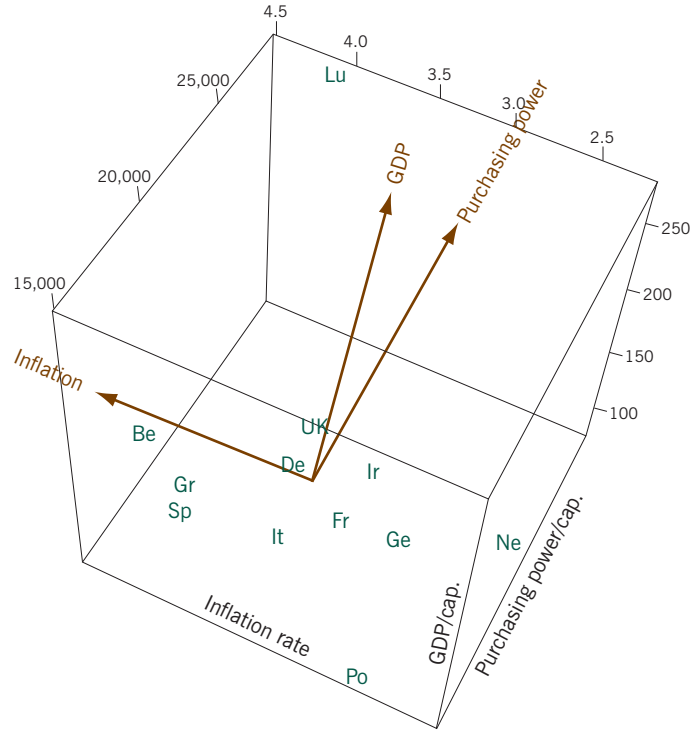


Exhibit 1.2:
Two scatterplots constructed from the three variables in Exhibit 1.1

Exhibit 1.3:
A three-dimensional scatterplot of all three variables in Exhibit 1.1, seen from two points of view. The second one is more informative about the relative positions of the countries in the three-dimensional space and axes are shown parallel to their respective sides of the cube





$$\begin{pmatrix} 8 & 2 & 2 & -6 \\ 5 & 0 & 3 & -4 \\ -2 & -3 & 3 & 1 \\ 2 & 3 & -3 & -1 \\ 4 & 6 & -6 & -2 \end{pmatrix} = \begin{pmatrix} 2 & 2 \\ 1 & 2 \\ -1 & 1 \\ 1 & -1 \\ 2 & -2 \end{pmatrix} \begin{pmatrix} 3 & 2 & -1 & -2 \\ 1 & -1 & 2 & -1 \end{pmatrix} \tag{1.1}$$

To give some general names to these matrices, let us call the first matrix the *target matrix*, and the two matrices of the product the *left matrix* and the *right matrix* respectively, so that expressions of the form (1.1) can be written as:

$$\textit{target matrix} = \textit{left matrix} \cdot \textit{right matrix} \tag{1.2}$$

Notice that for the matrix product to be valid, the number of columns of the left matrix must be equal to the number of rows of the right matrix. The rules of matrix multiplication are that the elements of a row of the left matrix are multiplied by the corresponding elements of a column of the right matrix, and then added, producing that particular row–column element of the target matrix. For example, the first value of the target matrix, 8, is equal to $2 \times 3 + 2 \times 1$; and the value –6 in row 1, column 4, is calculated from the first row of the left matrix and the fourth column of the right matrix, as $2 \times (-2) + 2 \times (-1)$. Such "sums of cross-products" are called *scalar products*, and are the basis of the geometry of the biplot.

Scalar product

For any two vectors $\mathbf{a}^\top = [a_1\ a_2\ \ldots\ a_m]$ and $\mathbf{b}^\top = [b_1\ b_2\ \ldots\ b_m]$, with m elements each, the *scalar product* between $\mathbf{a}$ and $\mathbf{b}$ is

$$\mathbf{a}^\top\mathbf{b} = a_1b_1 + a_2b_2 + \cdots + a_mb_m \tag{1.3}$$

The notation $^\top$ stands for "transpose of", turning rows into columns and vice versa. Notice that the convention is to write vectors as columns, so that a row vector is defined as the transpose of a column vector.

In the case of (1.1) each row vector of the left matrix has two elements and each column vector of the right matrix has two elements, so each of these row–column pairs has a scalar product equal to the corresponding element of the target matrix. We write (1.1) as

$$\mathbf{S} = \mathbf{X}\mathbf{Y}^\top = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \mathbf{x}_3^\top \\ \mathbf{x}_4^\top \\ \mathbf{x}_5^\top \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 & \mathbf{y}_2 & \mathbf{y}_3 & \mathbf{y}_4 \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top\mathbf{y}_1 & \mathbf{x}_1^\top\mathbf{y}_2 & \mathbf{x}_1^\top\mathbf{y}_3 & \mathbf{x}_1^\top\mathbf{y}_4 \\ \mathbf{x}_2^\top\mathbf{y}_1 & \mathbf{x}_2^\top\mathbf{y}_2 & \mathbf{x}_2^\top\mathbf{y}_3 & \mathbf{x}_2^\top\mathbf{y}_4 \\ \mathbf{x}_3^\top\mathbf{y}_1 & \mathbf{x}_3^\top\mathbf{y}_2 & \mathbf{x}_3^\top\mathbf{y}_3 & \mathbf{x}_3^\top\mathbf{y}_4 \\ \mathbf{x}_4^\top\mathbf{y}_1 & \mathbf{x}_4^\top\mathbf{y}_2 & \mathbf{x}_4^\top\mathbf{y}_3 & \mathbf{x}_4^\top\mathbf{y}_4 \\ \mathbf{x}_5^\top\mathbf{y}_1 & \mathbf{x}_5^\top\mathbf{y}_2 & \mathbf{x}_5^\top\mathbf{y}_3 & \mathbf{x}_5^\top\mathbf{y}_4 \end{pmatrix} \tag{1.4}$$

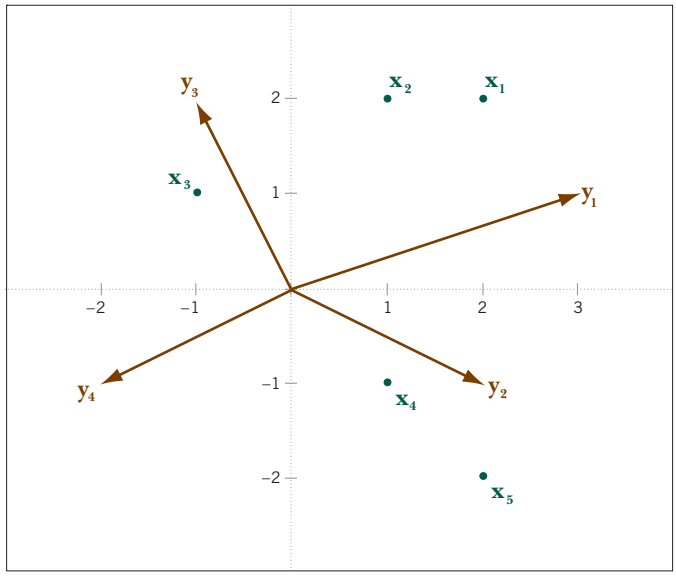
where $\mathbf{X}$ consists of a set of row vectors $\mathbf{x}_i^{\mathsf{T}}$, $i = 1,\ldots,5$ (each with two elements), and $\mathbf{Y}^{\mathsf{T}}$ consists of a set of column vectors $\mathbf{y}_j$, $j = 1,\ldots,4$ (also with two elements each). The right matrix is written as a transpose so that we have two matrices, $\mathbf{X}$ and $\mathbf{Y}$, whose *rows* both contain the vectors used in the scalar product calculations:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \mathbf{x}_3^{\mathsf{T}} \\ \mathbf{x}_4^{\mathsf{T}} \\ \mathbf{x}_5^{\mathsf{T}} \end{pmatrix} \qquad \mathbf{Y} = \begin{pmatrix} \mathbf{y}_1^{\mathsf{T}} \\ \mathbf{y}_2^{\mathsf{T}} \\ \mathbf{y}_3^{\mathsf{T}} \\ \mathbf{y}_4^{\mathsf{T}} \end{pmatrix} \tag{1.5}$$

Geometric interpretation of scalar product

There is a specific geometric interpretation of the scalar product which gives its usefulness as a method of data visualization. Using the simple example above, let us plot the rows of $\mathbf{X}$ and the rows of $\mathbf{Y}$ as points in a Euclidean space, which is two-dimensional in this case:

Exhibit 1.4:
The five points $\mathbf{x}_i$ of the left matrix and four points $\mathbf{y}_j$ of the right matrix in decomposition (1.1) (the latter points are shown as vectors connected to the origin). The scalar product between the i-th row point and the j-th column point gives the (i,j)-th value s_{ij} of the target matrix in (1.1)



$\mathbf{x}_1 = [\,2 \quad 2\,]^{\mathsf{T}}$
$\mathbf{x}_2 = [\,1 \quad 2\,]^{\mathsf{T}}$
$\mathbf{x}_3 = [-1 \quad 1\,]^{\mathsf{T}}$
$\mathbf{x}_4 = [\,1 \quad -1\,]^{\mathsf{T}}$
$\mathbf{x}_5 = [\,2 \quad -2\,]^{\mathsf{T}}$

$\mathbf{y}_1 = [\,3 \quad 1\,]^{\mathsf{T}}$
$\mathbf{y}_2 = [\,2 \quad -1\,]^{\mathsf{T}}$
$\mathbf{y}_3 = [-1 \quad 2\,]^{\mathsf{T}}$
$\mathbf{y}_4 = [-2 \quad -1\,]^{\mathsf{T}}$

In Exhibit 1.4 one of the sets of points, in this case the four points of the right matrix, are drawn as vectors connected to the origin. Each of these vectors defines a *biplot axis* onto which the other set of points can be projected. By projection of a point onto a vector we mean dropping the point perpendicularly onto the vector as shown in Exhibit 1.5. For example, the projection of $\mathbf{x}_1$ onto the vector $\mathbf{y}_1$ can be calculated as having length 2.530 (using simple trigonometry, the length of $\mathbf{x}_1$

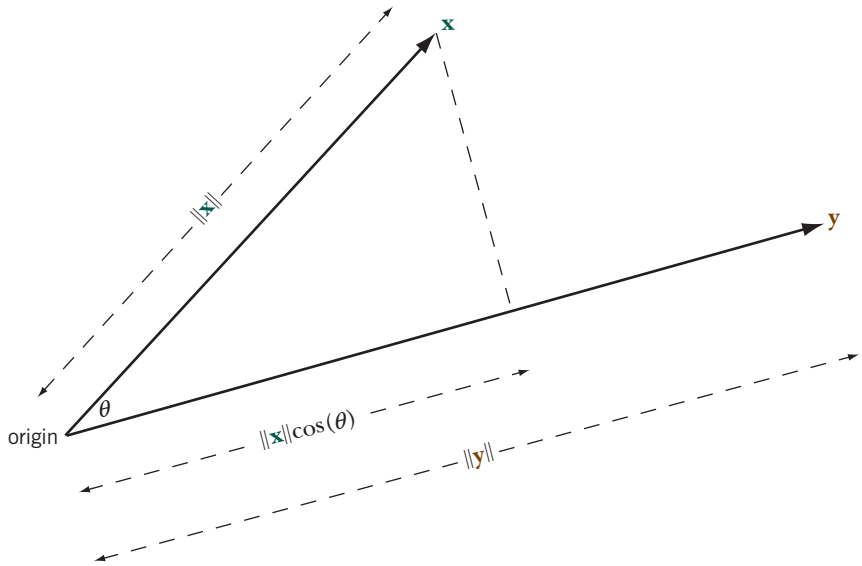


Exhibit 1.5:
Example of two points $\mathbf{x}$ *and* $\mathbf{y}$ *whose vectors subtend an angle of* θ *with respect to the origin. The scalar product between the points is the length of the projection of* $\mathbf{x}$ *onto* $\mathbf{y}$, $||\mathbf{x}|| \cos(\theta)$, *multiplied by the length of* $\mathbf{y}$, $||\mathbf{y}||$. *The result is identical if* $\mathbf{y}$ *is projected onto* $\mathbf{x}$ *and the projected length,* $||\mathbf{y}|| \cos(\theta)$, *is multiplied by the length of* $\mathbf{x}$, $||\mathbf{x}||$. *If* θ *is an obtuse angle (>90°), then* $\cos(\theta)$ *is negative and the projection has a negative sign, hence the scalar product is negative*

is 2.828, the square root of 8, and the angle between $\mathbf{x}_1$ and $\mathbf{y}_1$ is 26.57°, so the projection of $\mathbf{x}_1$ onto $\mathbf{y}_1$ has length $2.828 \times \cos(26.57°) = 2.828 \times 0.8944 = 2.530$). Now if this projected length is multiplied by the length of $\mathbf{y}_1$ (equal to 3.162, the square root of 10), the result is $2.530 \times 3.162 = 8.00$. This result is nothing else but the scalar product between $\mathbf{x}_1$ and $\mathbf{y}_1$: $(2 \times 3) + (2 \times 1) = 8$. So we have illustrated the result, shown in Exhibit 1.5, that the scalar product between two vectors is the length of the projection of the first vector onto the second multiplied by the length of the second one (or vice versa):

$$\mathbf{x}^T\mathbf{y} = ||\mathbf{x}|| \cdot ||\mathbf{y}|| \cdot \cos(\theta) \tag{1.6}$$

All biplots have the above property, in fact the property is inherent to the definition of the biplot. Two sets of points are graphed in a joint plot, according to coordinates provided by the left and right matrices of the decomposition of a target matrix. Either set can be chosen as a set of biplot axes, and often this set is connected by vectors to the origin of the biplot to indicate the choice. Let us call these the *biplot vectors* as opposed to the other set of *biplot points*. The biplot points can then be projected onto the biplot axes, and their projections, multiplied by the length of the corresponding biplot vectors, give the scalar products which are equal to the elements of the target matrix.

Calibration of biplot axes

The practical value of the biplot property of representing data by scalar products is only truly apparent if one thinks of the biplot axes as being *calibrated*. This is the

placing of tic marks on the biplot axes indicating a scale for reading off the values in the target matrix by simply projecting the biplot points onto the biplot axes. For example, consider again the biplot axis in Exhibit 1.4 defined by the biplot vector $\mathbf{y}_1$, and now consider the projections of all the biplot points $\mathbf{x}_1, \ldots, \mathbf{x}_5$ onto this axis. These five projections will all be multiplied by the same value, the length of $\mathbf{y}_1$, to obtain the five values in the first column of the target matrix. This means that the five projected values are directly proportional to the target values, so we can calibrate the biplot axis with a scale that depends on the length of $\mathbf{y}_1$, after which we no longer have to perform the multiplication by the length of $\mathbf{y}_1$. To know what the length of one unit is along the biplot axis, consider again the way we obtain a particular target value in row i from the corresponding scalar product:

$$\text{target value in row } i = \begin{pmatrix}\text{length of projection}\\ \text{of } i\text{-th biplot point}\end{pmatrix} \times \begin{pmatrix}\text{length of biplot}\\ \text{vector}\end{pmatrix}$$

In order to know what one unit is on the biplot axis, we need to invert this formula for a target value of 1:

$$\text{length of projection of one unit} = 1 \,/\, \text{length of biplot vector} \qquad (1.7)$$

This means that the inverse of the lengths of the biplot vectors give us the lengths of one unit on the biplot axis (see Exhibit 1.6)—if a biplot vector is short, the intervals between units on that biplot axis are large (so values change slowly), and if the biplot vector is long, the intervals between units on that biplot axis are short (so values change fast). For example, the length of $\mathbf{y}_1$ is the square root of 10, 3.162. Hence, unit tic marks on the biplot axis through $\mathbf{y}_1$ are a distance of $1/3.162 = 0.3162$ apart.

Each vector $\mathbf{y}_j$ defines a biplot axis that can be calibrated in the same way. Although we will generally not show calibrations along the axes, the fact that axes can be calibrated gives the *raison d'être* of the biplot's interpretation. If the columns of the target matrix are variables and the rows are cases, the biplot representation will mean that variables can be depicted as axes pointing in a direction such that the values for the cases on that variable are obtained by projecting the cases onto the variables. The actual values of that variable are not as important as being able to see how the cases line up along that variable. And if two biplot axes lie in the same orientation we shall be able to deduce that the cases have the same relative positions along both variables, which translates into high inter-variable correlation.

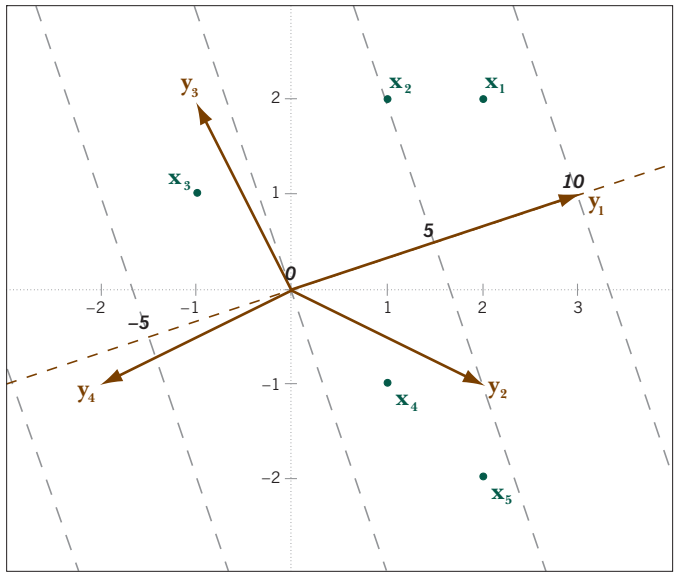


Exhibit 1.6:
Calibrating a biplot axis through vector $\mathbf{y}_1$*, shown as dashed line. The distance between units on this axis is the inverse of the length of* $\mathbf{y}_1$*, 0.3162, and allows placing values on the axis (shown in black). Points projected perpendicularly onto the biplot axis give values on the calibrated scale equal to the values in the first column of the target matrix (corresponding to* $\mathbf{y}_1$*, the first biplot vector). Thus we can read off the target values of 8, 5, –2, 2 and 4 for points* $\mathbf{x}_1$*, . . .,* $\mathbf{x}_5$*, respectively —see first column of target matrix in (1.1)*

Moving into higher-dimensional spaces

In this introductory example of how scalar products and projections give values of a target matrix, the left and right matrices of the decomposition provide two-dimensional coordinates, easily visualized in the planar display of Exhibits 1.4 and 1.6. If these matrices were three-dimensional (**X** and **Y** with three columns), we could still see what was happening if we could view the points and vectors in three-dimensional space. Calibration and projection are performed in exactly the same way. Of course, for a given data matrix in practice, we will not be able to find a decomposition into a product of matrices with only two or three dimensions: in fact, the number of dimensions needed is equal to the *rank* of the target matrix. For these higher-dimensional situations we shall need methods for reducing the dimensionality of the target matrix—this theme will be introduced in Chapter 5 and used extensively in the rest of the book.

SUMMARY: Biplots—the Basic Idea

1. *Scatterplots* typically plot observations on two variables with respect to rectangular coordinate axes. Three-dimensional scatterplots showing observations on three variables are possible using special software and a two-dimensional view of these three variables can be optimized to show a maximum amount of the variance (i.e., dispersion) of the plotted points. The biplot generalizes this idea to many variables being observed and viewed simultaneously in an optimal fashion.
2. *Biplots* are defined as the decomposition of a *target matrix* into the product of two matrices, called *left and right matrices*: $\mathbf{S} = \mathbf{X}\mathbf{Y}^{\top}$. Elements in the target ma-

trix **S** are equal to *scalar products* between corresponding pairs of vectors in the rows of **X** and **Y** respectively.

3. The geometry of scalar products provides the following rule for interpreting the biplot graphically. The vectors in the left and right matrices provide two sets of points, one of which can be considered as a set of *biplot vectors* defining *biplot axes*, and the other as a set of *biplot points*. Points can be projected perpendicularly onto biplot axes to recover the values in the target matrix, since the lengths of these projections multiplied by the lengths of the corresponding biplot vectors are equal to the scalar products, and thus in turn equal to the target values.
4. *Calibration* of the biplot axes is possible, which means that values of the target matrix can be read off directly from the projections, just as in scatterplots where points are projected onto the axes to read their values.
5. The "bi" in biplot refers to the fact that two sets of points (i.e., the rows and columns of the target matrix) are visualized by scalar products, not the fact that the display is usually two-dimensional. The biplot and its geometry hold for spaces of any dimensionality, but we shall need *dimension-reducing techniques* in practice when data matrices have high inherent dimensionality and a representation is required with respect to a low number of dimensions, usually two or three.

CHAPTER 2

Regression Biplots

Biplots rely on the decomposition of a target matrix into the product of two matrices. A common situation in statistics where we have such a decomposition is in regression analysis: $\hat{\mathbf{Y}} = \mathbf{XB}$, where $\mathbf{X}$ is a set of explanatory variables, $\mathbf{B}$ contains estimated regression coefficients and the values in $\hat{\mathbf{Y}}$ are the estimated values of one or more response variables. Thus $\hat{\mathbf{Y}}$ serves as the target matrix and $\mathbf{X}$ and $\mathbf{B}$ serve as the left and right matrices (actually $\mathbf{B}^{\mathsf{T}}$, since the right matrix of the decomposition is written in transposed form—see (1.4)). The coefficients in $\mathbf{B}$ are estimated to minimize the sum-of-squared errors between the original response variables in $\mathbf{Y}$ and the estimated values in $\hat{\mathbf{Y}}$. This context provides an excellent introduction to biplots as an approximation to higher-dimensional data.

Contents

Data set "bioenv"

Throughout this book we shall be using a small data set which serves as an excellent example of several biplot methods. The context is in marine biology and the data consist of two sets of variables observed at the same locations on the sea-bed: the first is a set of biological variables, the counts of five groups of species, and the second is a set of four environmental variables. The data set, called "bioenv", is shown in Exhibit 2.1. The species groups are abbreviated as "a" to "e". The environmental variables are "pollution", a composite index of pollution combining measurements of heavy metal concentrations and hydrocarbons; "depth", the depth in metres of the sea-bed where the sample was taken; "temperature", the temperature of the water at the sampling point; and "sediment", a classification

Exhibit 2.1: *Typical set of multivariate biological and environmental data: the species data are counts, while the environmental data are continuous measurements, each variable on a different scale; the last variable is a categorical variable classifying the substrate as mainly* C *(=clay/silt),* S *(=sand) or* G *(=gravel/stone)*

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

of the substrate of the sample into one of three sediment categories. Initially, we are going to consider the biological variables and only two of the environmental variables, "pollution" and "depth".

Simple linear regression on two explanatory variables

To start off let us consider species "d" as a response variable, denoted by d, being modelled as a linear function of two explanatory variables, "pollution" and "depth", denoting these two variables by y and x respectively. Simple linear regression leads to the following estimates of the response variable:

$$\hat{d} = 6.135 - 1.388\,y + 0.148\,x \qquad R^2 = 0.442 \qquad (2.1)$$

From a statistical inference point of view, both y and x are significant at the 5% level—their p-values based on the classical t-tests are 0.008 and 0.035 respectively.[1] The regression coefficients on the explanatory variables have the following interpretation: for every unit increase of pollution (variable y), abundance of species d decreases by 1.388 on average; while for every unit increase of depth (variable x), abundance of d increases by 0.148 on average. The amount of variance in d explained by the two variables is 44.2%, which means that the sum of squared errors across the 30 observations, $\sum_i (d_i - \hat{d}_i)^2$, which is minimized by the linear regression, is 55.8% of the total variance of d.

Standardized regression coefficients

The estimated regression coefficients in (2.1), i.e. the "slope" coefficients –1.388 and 0.148, have scales that depend on the scale of d and the scale of the two explanatory variables y and x, and so are difficult to compare with each other. To remove the effect of scale, all variables should be expressed in a comparable scale-free way. The most common way of doing this is to standardize all variables by centring them with respect to their respective means and dividing them by their respective standard deviations. We denote the standardized values of the three variables (their "z-scores") as d^*, y^* and x^* respectively, each having mean zero and variance 1 thanks to the standardization. The estimated regression relationship (2.1), including what are called the *standardized regression coefficients*, then becomes:[2]

$$\hat{d}^* = -0.446\,y^* + 0.347\,x^* \qquad R^2 = 0.442 \qquad (2.2)$$

Notice that there is no intercept, since all variables have mean zero. The regression coefficients now quantify the change in the standardized value of the response variable estimated from an increase of one standardized unit (i.e., one standard deviation) of each explanatory variable. The two coefficients can be compared and it seems that pollution has a bigger (negative) effect on species d than the (positive) effect of depth. Exhibit 2.2 shows schematically the difference between the regression plane for the unstandardized and standardized variables respectively.

Gradient of the regression plane

The standardized regression coefficients in (2.2) have an interesting geometric interpretation. They are the partial derivatives of $\hat{d}^*$ with respect to the two variables y^* and x^*, which are written mathematically as:

1. All calculations can be followed and repeated using the R code given in the Computational Appendix.
2. Since some regression programs do not automatically give the standardized regression coefficients, these can be easily calculated from the original ones as follows: standardized coefficient = original coefficient × (standard deviation of explanatory variable / standard deviation of response variable). See the Computational Appendix for examples of this calculation.

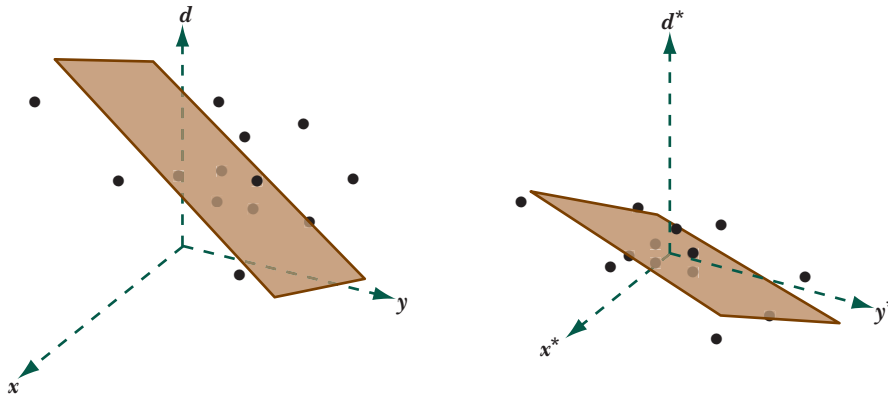


Exhibit 2.2: *Schematic geometric representation in three-dimensions of the regression plane for the original data* (on left) *and standardized data* (on right)*, where the response variable is the vertical dimension, and the two explanatory variables are on the "floor", as it were (imagine that we are looking down towards the corner of a room). The plane on the right goes through the origin of the three axes*

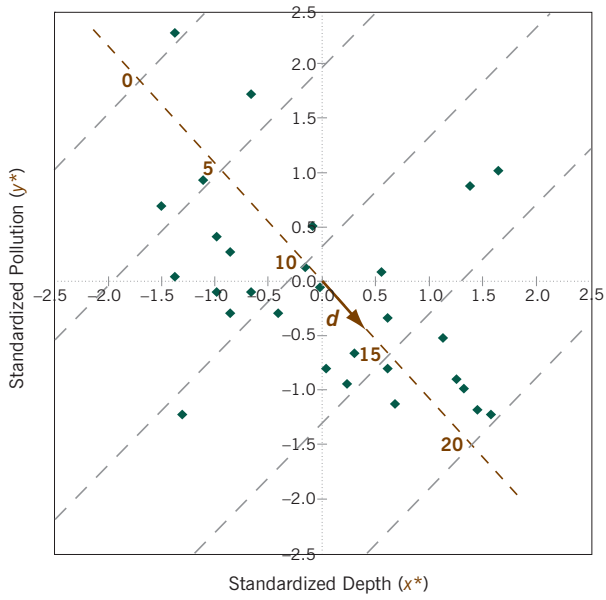
$$\frac{\partial \hat{d}^*}{\partial y^*} = -0.446 \qquad \frac{\partial \hat{d}^*}{\partial x^*} = 0.347$$

As a vector $[-0.446 \;\; 0.347]^T$ these two numbers indicate the *gradient* of the plane (2.2), that is the direction of steepest ascent up the plane in the right-hand image of Exhibit 2.2.

The complete geometry of the regression can then be drawn in the two-dimensional space of the explanatory variables, as shown in Exhibit 2.3. This is the scatterplot of the two variables y^* and x^*, whose values are given in the table alongside the figure. The brown arrow is the gradient vector, with coordinates -0.446 on the y^*-axis and 0.347 on the x^*-axis. So here we are looking at the plane on the right of Exhibit 2.2 from the top, down onto the y^*–x^* plane. The arrow indicates the gradient vector, pointing in the direction of steepest ascent of the plane, which is all we need to know to understand the orientation of the plane.

Contours of the regression plane

Now the contours of the regression plane, i.e., the lines of constant "height", by which we mean constant values of $\hat{d}^*$, form lines perpendicular to the gradient vector, just like in Exhibit 2.3. From the steepness of the plane (which we know from the gradient vector) it seems intuitively obvious that we can work out the heights of these contours on the standardized scale of d and then transform these back to d's original scale—in Exhibit 2.3 we show the contours for 0, 5, 10, 15 and 20. That is, we can calibrate the biplot axis for species d (we will explain exactly how to calibrate this axis below). Hence, to obtain the estimates of d for any given point y^* and x^* we simply need to see which contour line it is on, that is project it perpendicularly onto biplot axis d.



y*	x*
0.132	−0.156
−0.802	0.036
0.413	−0.988
1.720	−0.668
−0.288	−0.860
−0.895	1.253
0.039	−1.373
0.272	−0.860
−0.288	−0.412
2.561	−0.348
0.926	−1.116
−0.335	0.613
2.281	−1.373
0.086	0.549
1.020	1.637
−0.802	0.613
0.880	1.381
−0.054	−0.028
−0.662	0.292
0.506	−0.092
−0.101	−0.988
−1.222	−1.309
−0.989	1.317
−0.101	−0.668
−1.175	1.445
−0.942	0.228
−1.129	0.677
−0.522	1.125
0.693	−1.501
−1.222	1.573

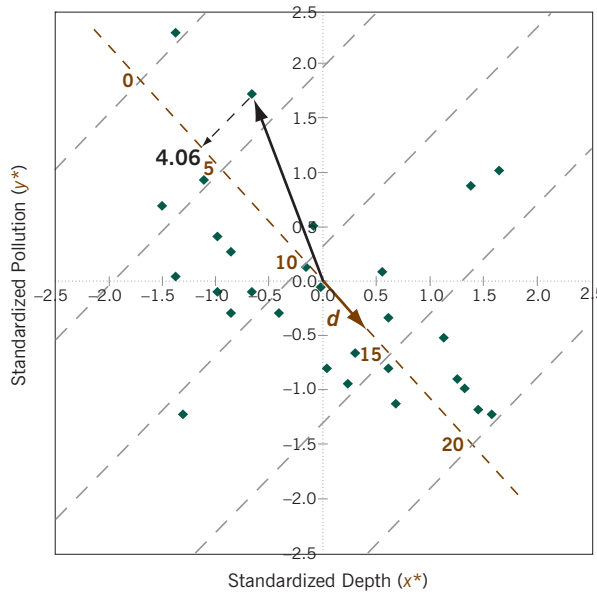
Exhibit 2.3:
The regression plane for species ***d*** *is shown by its gradient vector in the* x^*–y^* *space of the explanatory variables. Contour lines (or isolines) are drawn at selected heights*

Seeing which contour line corresponds to any given sample point is thus equivalent to projecting the point perpendicularly onto the biplot axis and reading off the value. Exhibit 2.4 shows an example of estimating the value of d for the fourth sample, giving a value of 4.06. This is not equal to the observed value of 3 (see original data on the left of Exhibit 2.4), but we do not expect it to be, since the variance explained by the regression of d on y and x is 44.2%. If we projected all the sample points onto the d-axis, and recovered the original values exactly, this would mean the regression plane passes exactly through the data points and the explained variance would be 100%. We are not in this situation, unfortunately, since our estimated values explain only 44.2% of the variance of d, in other words 55.8% of the variance is due to differences between the estimates and the observed values.

All this sounds exactly like what we described in Chapter 1, particularly concerning Exhibit 1.3, and indeed it is, because the regression equation (2.2) is nothing else but the scalar product between the gradient vector (indicating the biplot axis) and a general point in the x^*–y^* plane. Hence, this equation for estimating the response, given values of the two predictors as a biplot point, can be converted into the scalar product between this biplot point and the biplot gradient vector (the regression coefficients). This is equivalent to projecting the biplot point onto the biplot axis defined by the gradient vector, which is calibrated in units of the response variable.

Exhibit 2.4:
Projection of sample 4 onto the biplot axis, showing sample 4's original values in the table on the left and standardized values of the predictors on the right. The predicted value is 4.06, compared to the observed value of 3, hence an error of 1.06. The sum of squared errors for the 30 samples accounts for 55.8% of the variance of ***d****, while the explained variance* (R^2) *is 44.2%*

d	*y*	*x*
14	4.8	72
11	2.8	75
8	5.4	59
3	8.2	64
10	3.9	61
16	2.6	94
11	4.6	53
0	5.1	61
14	3.9	68
9	10.0	69
6	6.5	57
15	3.8	84
0	9.4	53
9	4.7	83
12	6.7	100
3	2.8	84
0	6.4	96
20	4.4	74
16	3.1	79
9	5.6	73
4	4.3	59
9	1.9	54
17	2.4	95
7	4.3	64
23	2.0	97
10	2.5	78
25	2.1	85
20	3.4	92
8	6.0	51
18	1.9	99



*y**	*x**
0.132	−0.156
−0.802	0.036
0.413	−0.988
1.720	−0.668
−0.288	−0.860
−0.895	1.253
0.039	−1.373
0.272	−0.860
−0.288	−0.412
2.561	−0.348
0.926	−1.116
−0.335	0.613
2.281	−1.373
0.086	0.549
1.020	1.637
−0.802	0.613
0.880	1.381
−0.054	−0.028
−0.662	0.292
0.506	−0.092
−0.101	−0.988
−1.222	−1.309
−0.989	1.317
−0.101	−0.668
−1.175	1.445
−0.942	0.228
−1.129	0.677
−0.522	1.125
0.693	−1.501
−1.222	1.573

Calibrating a regression biplot axis

In Chapter 1 we showed how to calibrate a biplot axis—one unit is inversely proportional to the length of the corresponding biplot vector (see equation (1.7)). In regression biplots the situation is the same, except the unit is a standardized unit and we prefer to calibrate according to the original scale of the variable. To express the regression coefficients on the original scale of d in the present case, we would simply multiply them by the standard deviation of d, which is 6.67, making them 6.67×-0.446 and 6.67×0.347 respectively. Then the calculation is as before, using the rescaled regression coefficients:

$$\begin{aligned}
\text{one unit of } d &= 1 \,/\, \text{length of biplot vector} \\
&= 1 \,/\, \sqrt{(6.67\times -0.446)^2 + (6.67\times 0.347)^2} \\
&= 1 \,/\, \left(6.67\times \sqrt{(-0.446)^2 + (0.347)^2}\right) \\
&= 1 \,/\, \left(6.67\times 0.565\right) \\
&= 0.265
\end{aligned}$$

In general, the calculation is:

$$\text{one unit of variable} = 1 \,/\, \left(\begin{array}{c}\text{standard deviation}\\ \text{of variable}\end{array} \times \begin{array}{c}\text{length of biplot}\\ \text{vector}\end{array}\right) \qquad (2.3)$$

that is, the unit length of the standardized variable divided by the (unstandardized) variable's standard deviation. As far as the centre of the biplot is concerned, the variable's average is at the origin—in the present example the origin should be at the value 10.9, the average of *d*. We know that values are increasing in the direction of the biplot vector (towards bottom right in Exhibits 2.3 and 2.4), and also have computed the length of one unit on the biplot axis, so we have all we need to calibrate the axis. In the exhibits we calibrated at every 5 units, so the distance interval along the axis between consecutive values is $5 \times 0.265 = 1.325$.

Regression biplots with several responses

Each of the five species in Exhibit 2.1 can be linearly regressed on the two predictors "pollution" (*y*) and "depth" (*x*), using standardized scales for all variables, to obtain standardized regression coefficients that can be used as biplot vectors. Exhibit 2.5 shows the five regression analyses in one biplot. Each biplot vector points in the direction of steepest ascent of the regression plane. The larger the regression coefficients, the longer are the arrows and thus the steeper is the regression plane. If two biplot vectors are pointing in the same direction (for example, *b* and *d*) their relationships with the explanatory variables are similar. Species *c* clearly has an opposite relationship to the others, in that its regression

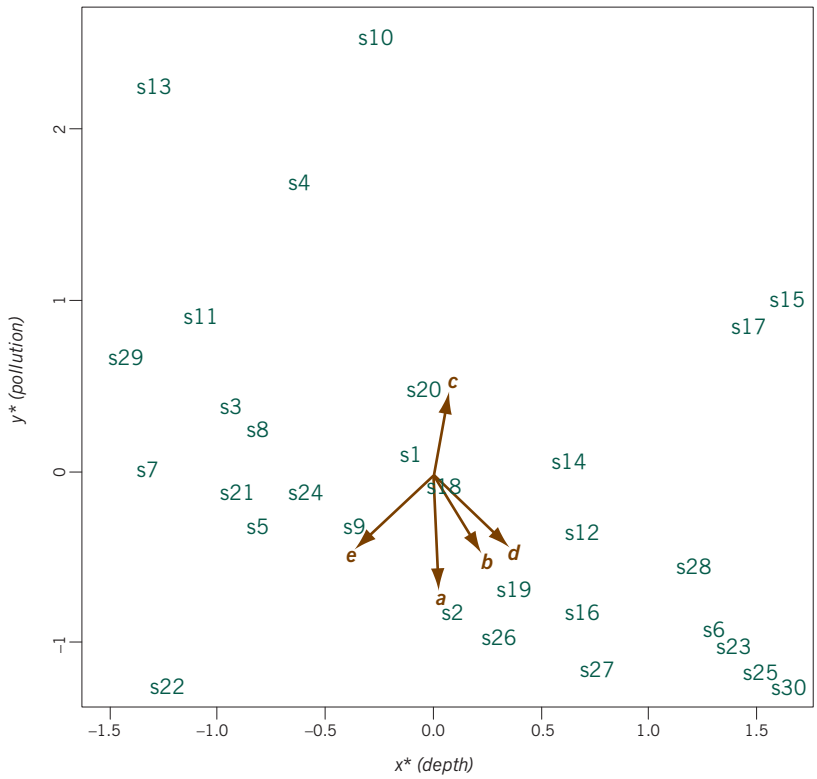


Exhibit 2.5:
Regression biplot of five response variables, species ***a*** *to* ***e****, in the space of the two standardized explanatory variables. The overall explained variance for the five regressions is 41.5%, which is the measure of fit of the biplot*

coefficient with pollution is positive while all the others are negative. The biplot axes through the biplot vectors can each be calibrated in the same way as explained above for response variable *d*, and the projections of the 30 samples (numbered in Exhibit 2.5) onto a particular biplot axis give the estimated values for that response. How close these estimated values are to the observed ones is measured by the R^2 for each regression: the percentages of explained variance, respectively for *a* to *e*, are 52.9%, 39.1%, 21.8%, 44.2% and 23.5%, with an overall R^2 of 41.5%. This overall value measures the quality of the regression biplot to explain all five response variables.

Finally, to show how regression biplots fit the general definition of the biplot given in Chapter 1, we write the estimation equation (2.4) for all five response variables in a matrix decomposition formulation as follows:

$$\begin{bmatrix} \hat{\mathbf{a}}^* & \hat{\mathbf{b}}^* & \hat{\mathbf{c}}^* & \hat{\mathbf{d}}^* & \hat{\mathbf{e}}^* \end{bmatrix} = \begin{bmatrix} \mathbf{y}^* & \mathbf{x}^* \end{bmatrix} \begin{bmatrix} -0.717 & -0.499 & 0.491 & -0.446 & -0.475 \\ 0.025 & 0.229 & 0.074 & 0.347 & -0.400 \end{bmatrix} \quad (2.4)$$

that is, $$\hat{\mathbf{S}} = \mathbf{U}\mathbf{B}^{\mathsf{T}}$$

where the target matrix is the 30 × 5 matrix of estimated response values (standardized), the left matrix of the decomposition is the 30 × 2 matrix of standardized explanatory variables and the right matrix contains the standardized regression coefficients. The target matrix is an estimation $\hat{\mathbf{S}}$ of the observed (standardized) responses $\mathbf{S} = [\mathbf{a}^* \ \mathbf{b}^* \ \mathbf{c}^* \ \mathbf{d}^* \ \mathbf{e}^*]$, which can be written as: $\mathbf{S} \approx \hat{\mathbf{S}}$, which reads "**S** is approximated by $\hat{\mathbf{S}}$". In this case the sense of the approximation is that of least-squares regression, where $\mathbf{U} = [\mathbf{y}^* \ \mathbf{x}^*]$ is the fixed matrix of explanatory variables and the regression coefficients $\mathbf{B}^{\mathsf{T}}$ are calculated in the usual way by least squares as follows:

$$\mathbf{B}^{\mathsf{T}} = (\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{S} \quad (2.5)$$

The complete process of the regression biplot can thus be summarized theoretically as follows:

$$\mathbf{S} \approx \hat{\mathbf{S}} = \mathbf{U}(\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{S} \quad (2.6)$$

The matrix $\mathbf{U}(\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}}$ is called the *projection matrix* of **S** onto the explanatory variables in **U**. In fact, we can write **S** as the following sum:

$$\begin{aligned} \mathbf{S} &= \hat{\mathbf{S}} + (\mathbf{S} - \hat{\mathbf{S}}) \\ \mathbf{S} &= (\mathbf{U}(\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}})\mathbf{S} + (\mathbf{I} - \mathbf{U}(\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}})\mathbf{S} \end{aligned} \quad (2.7)$$

where the first term is the projection of **S** onto the space of the explanatory variables, and the second term is the projection of **S** onto the space orthogonal to (uncorrelated with) the explanatory variables. (**I** denotes the identity matrix, a diagonal matrix with 1's down the diagonal.) The part of **S** that is explicable by the explanatory variables **U** can be biplotted according to (2.6), as we have done in Exhibit 2.5, using **U** as the left matrix and the standardized regression coefficients in $(\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}\mathbf{U}^{\mathsf{T}}\mathbf{S}$ as the right matrix.

Notice that in these initial chapters we consider the case of only two explanatory variables, which conveniently gives a biplot in two dimensions. If we used the three variables "pollution", "depth" and "temperature" we would need three dimensions to show the variables and the regression relationships. We should also stress again that the "bi" in biplot does not refer to the bidimensional nature of the figures, but the fact that we depict rows and columns together. The case of three or more explanatory variables will be dealt with in later chapters (from Chapter 5 onwards).

SUMMARY: Regression Biplots

1. A *regression biplot* shows the cases (usually rows) and a set of response variables (usually columns) of a data matrix in the same joint representation, which is constructed using a set of explanatory variables, or predictors. Cases are shown as biplot points with respect to standardized values of the predictors and variables are shown as biplot vectors, each according to the standardized regression coefficients of its regression on the predictors.
2. The biplot vectors represent the separate linear regressions and define biplot axes onto which the case points can be projected. The axes can be calibrated so that predicted values from the regressions can be read off.
3. The quality of the regression biplot is measured by the percentages of variance explained by the individual regressions that build the biplot.

CHAPTER 3

Generalized Linear Model Biplots

Generalized linear models are a wide class of regression-type models for different types of data and analytical objectives. Linear regression is the simplest form of a generalized linear model, where the mean of the response variable, given values of the predictors, is a linear function of the predictors and the distribution of the data around the mean is normal. The regression coefficients are estimated by fitting lines or planes (or "hyperplanes" for more than two explanatory variables) by least squares to the data points. Then, as we saw in Chapter 2, estimated mean values of response variables can be obtained by projecting case points onto variable vectors. This idea is developed and extended in two ways to generalized linear models: first, the mean of the response can be transformed nonlinearly, and it is this transformed mean that is modelled as a linear function of the predictors; and second, the probability distribution of the response variable around the mean function can be different from the normal distribution. In this chapter we first discuss data transformations and then give two examples of generalized linear model biplots: Poisson regression biplots and logistic regression biplots.

Contents

Data transformations

As an intermediate step towards considering *generalized linear models* (GLM), let us suppose that we wished to transform the response variables and then perform the regression analysis. There could be many reasons for this, such as making the data more symmetrically distributed, or reducing the effect of outliers. Power transformations are commonly used on data such as the species counts in Exhibit 2.1—these can be either a square root or double square root (i.e., fourth root) transformation, or the Box-Cox family of power transformations which includes the logarithmic transformation. For example, considering species *d* again, let us con-

Exhibit 3.1: *The regression coefficients for the five regressions where in each case the response variable has been fourth root transformed. Overall variance explained is 33.9%*

	Std devn	*Constant*	y^*	x^*	R^2
$a^{1/4}$	0.905	1.492	–0.672	0.073	60.5%
$b^{1/4}$	0.845	1.301	–0.506	0.006	36.2%
$c^{1/4}$	0.907	1.211	0.387	0.086	15.9%
$d^{1/4}$	0.602	1.639	–0.288	0.060	27.6%
$e^{1/4}$	0.755	0.815	–0.375	–0.255	22.8%

sider the fourth root transformation $d_0 = d^{1/4}$. Fitting this transformed response to the two standardized predictors y^* ("pollution") and x^* ("depth") as before leads to the following equation for predicting d_0:

$$\hat{d}_0 = 1.642 - 0.288\,y^* + 0.060\,x^* \qquad R^2 = 0.276 \tag{3.1}$$

Notice first that we have not centred the transformed data, hence the presence of the constant in the regression—the constant is equal to the mean of d_0 because the predictor variables are centred. Also, because the power transformation tends to homogenize the variances (i.e., make them more similar in value), we have not standardized d_0 either. If some transformed variables have more variance, then we would like to preserve this fact.[3]

The complete set of results for all of the transformed responses is given in Exhibit 3.1.

The constants give the predicted value for mean values of y and x, when $y^* = x^* = 0$; for example, the average of $d^{1/4}$ is 1.639, which transforms back to a value of d of $1.639^4 = 7.216$.

Biplot with nonlinear calibration

The regression coefficients can again be used as biplot vectors, shown in Exhibit 3.2. The positions of the sample points are identical to the previous Exhibits 2.3 and 2.4. The difference between this biplot and the previous one for untransformed data (Exhibit 2.4) is that the regression surfaces (in the third dimension, "above" the biplot space) indicated by the biplot vectors are linear planes for the transformed variables, and thus nonlinear in terms of the original ones. So the calibration of the biplot axes in terms of the original variables is more complicated because the intervals between scale units are not constant.

3. In this example the standard deviations of the original frequencies varied from 3.96 for species *e* to 12.6 for species *a*, while in the double square root transforms the range is from 0.602 for species *d* to 0.907 for species *c*. Notice that the ordering of the standard deviations is not necessarily preserved by the power transformation of the data, even though the transformation is monotonic.

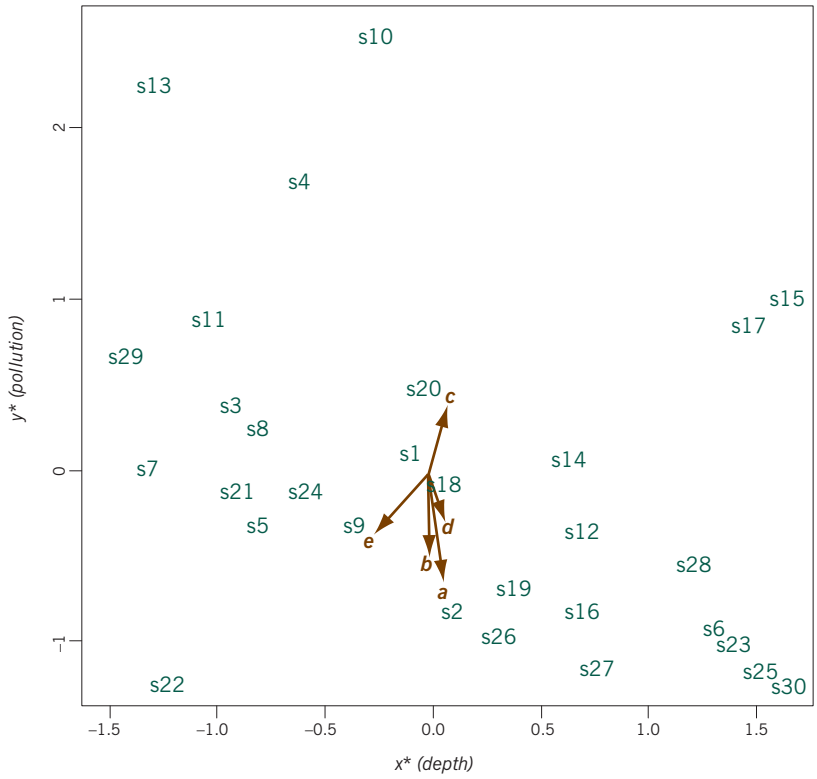
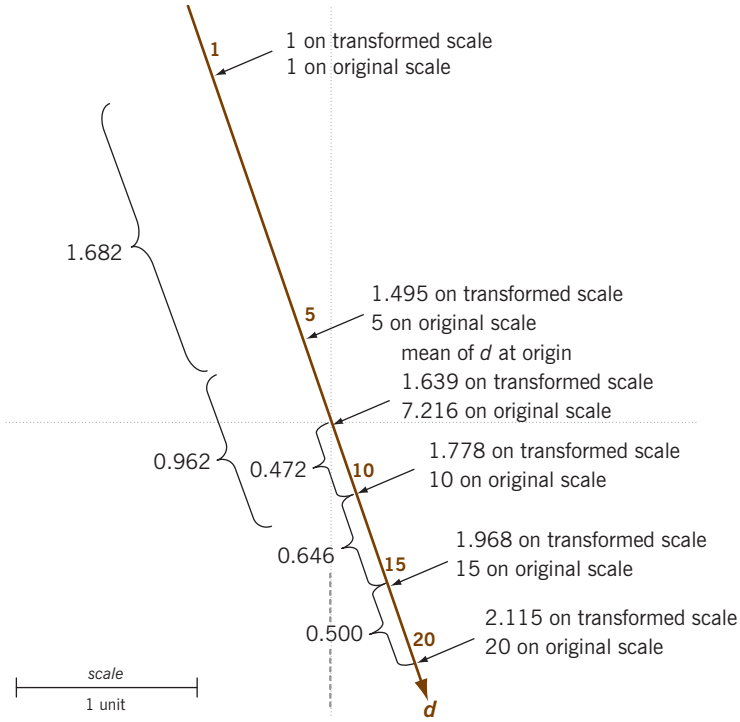


Exhibit 3.2:
Biplot of the fourth root transformed species data, showing biplot vectors given by regression coefficients in Exhibit 3.1, i.e., the directions of planes corresponding to regressions of the transformed species variables on standardized "pollution" (y) and standardized "depth" (x*)*

For example, let us consider variable d again, and its transformed version $d_0 = d^{1/4}$. Exhibit 3.3 illustrates the calculations which we describe now. The centre of the map corresponds to the mean of the transformed variable, 1.639, which we already calculated above to correspond to 7.216 of the original variable, so we know that the value of d at the origin is 7.216 and that it is increasing in the direction of the arrow, i.e., downwards in the biplot.

The first "tic mark" we would put on the biplot axis through d is at the value 10, which on the transformed scale is $10^{1/4} = 1.778$. The difference between this value and the mean value of 1.639 is 0.139. A unit length on the axis is, as before, 1 divided by the length of the biplot vector: $1 / \sqrt{0.288^2 + 0.060^2} = 3.399$, as shown in Exhibit 3.3. Hence, the distance along the biplot axis to put the tic mark is $0.139 \times 3.399 = 0.472$. The next tic mark at $d = 15$ corresponds to $15^{1/4} = 1.968$, a difference of $1.968 - 1.778 = 0.190$ from the position of $d = 10$, hence at a distance of $0.190 \times 3.399 = 0.646$ from the tic mark for 10. Going in the other direction, to put a tic mark for $d = 5$, the transformed value is $5^{1/4} = 1.495$, a difference relative to the position of $d = 10$ of $1.778 - 1.495 = 0.283$, or $0.283 \times 3.399 = 0.962$ units away from the tic mark for 10 in the negative direction for d (i.e., upwards

Exhibit 3.3:
Nonlinear calibration of biplot axis through reponse variable ***d****. Because* ***d*** *has been fourth root transformed, the calibrations are not at regular intervals*



in Exhibit 3.2). The tic mark for 1 (same on original and transformed scales) is 0.495 transformed units away from 5, and so $0.495 \times 3.399 = 1.682$ units away from 5 on the biplot axis. The nonlinearity of the calibration on the biplot axis is clear in Exhibit 3.3, with the tic marks getting closer together as d increases. The contours are still perpendicular to the biplot axis, so the interpretation is still based on projecting the biplot points onto the biplot axes, bearing in mind the contracting scale due to the transformation.

The fourth root transformation of the response variable is a *monotonically increasing* function, hence the calibration of the biplot axis shows values increasing in the direction of the biplot vector, albeit increasing in a nonlinear fashion. Although seldom done, a non-monotonic transformation, for example a quadratic transformation which rises and then falls, could also be applied to the response variable. The effect would be that the calibrations on the biplot axis would increase and then decrease again.

Poisson regression biplots

The regression biplots in Chapter 2 and those described above with transformed responses use regression coefficients for the biplot vectors that have been obtained using least-squares fitting of the response variable, with or with-

	Constant	y^*	x^*	*Error*
$\log(\bar{a})$	2.179	–1.125	–0.067	0.388
$\log(\bar{b})$	1.853	–0.812	0.183	0.540
$\log(\bar{c})$	2.041	0.417	0.053	0.831
$\log(\bar{d})$	2.296	–0.337	0.199	0.614
$\log(\bar{e})$	0.828	–0.823	–0.568	0.714

Exhibit 3.4:
The regression coefficients for the five Poisson regressions of the species responses on the predictors "pollution" y^* and "depth" x^*. Rather than variance explained, the "error" of the model fit is reported as the deviance of the solution relative to the null deviance when there no predictors (so low values mean good fit)

out transformation, to the explanatory variables. This idea originates in the assumption that, conditional on the explanatory variables, the distribution of the response variable in the population is normal, with the conditional mean of the response equal to a linear function of the explanatory variables. In generalized linear modelling this idea is extended to different distributions, but in every case some transformation of the mean of the response, called the *link function*, is modelled as a linear function of the explanatory variables. Linear regression is the simplest example of a *generalized linear model* (GLM) where there is no transformation of the mean (i.e., the link function is the identity function) and the conditional distribution is normal. The coefficients of a GLM are estimated using the principle of maximum likelihood, which has as a special case the least-squares procedure when the assumed distribution is normal.

The first example of a "non-normal" GLM that we consider is *Poisson regression*. Since the species variables *a* to *e* are counts, a more appropriate distribution would be the Poisson distribution. In Poisson regression the link function is logarithmic, so the model postulates that the logarithm of the response mean is a linear function of the explanatory variables, and the assumed conditional distribution of the response is Poisson. Fitting this model for each of the five responses is just as easy in R as fitting regular regression, using the `glm` function (see the Computational Appendix) and the estimated coefficients are given in Exhibit 3.4. Notice that the way the success of the model fit is measured is the opposite here, in the sense that for good fit the "error" should be low. In the simple regression case, subtracting the "error" from 1 would give R^2.

Notice the difference between the GLM and what we did before: we have not log-transformed the original data, which would have been a problem since there are zeros in the data, but have rather modelled the logarithm of the (conditional) mean as a linear function of the explanatory variables. For example, in the case of species *d* the model estimates are given by:

$$\log(\bar{d}) = 2.296 - 0.337y^* + 0.199x^* \qquad (3.2)$$

Exponentiating both sides, this gives the equation:

$$\bar{d} = \exp(2.296) \cdot \exp(-0.337y^*) \cdot \exp(0.199x^*) \tag{3.3}$$

so that the exponentials of the coefficients −0.337 and 0.199 model the multiplicative effect on the estimated mean response: a one (standard deviation) unit increase in y^* multiplies $\bar{d}$ by $\exp(-0.337) = 0.714$ (a 28.6% decrease), while a one unit increase in x^* multiplies $\bar{d}$ by $\exp(0.199) = 1.221$ (a 22.1% decrease). Again, the coefficients define a biplot vector in the space of the explanatory variables. To calibrate the vector, the value at the origin corresponds to the value of the mean response at the means of the explanatory variables $y^* = x^* = 0$, that is $\bar{d} = \exp(2.296) = 9.934$. To place tic marks on the biplot axis we would again calculate what a unit length on the axis is: $1/\sqrt{0.337^2 + 0.199^2} = 2.556$, which corresponds to a unit on the logarithmic scale. Using this we can work out where tic marks should be placed for values of $\bar{d}$ such as 0, 5, 10, 15, etc.—this will be a logarithmic scale on the biplot axis, with intervals between tic marks contracting as the response variable increases. We do not show the Poisson biplot here, but it can be computed using the script in the Computational Appendix.

Logistic regression biplots

Let us suppose, as is indeed common in ecological research, that we are interested more in the presence/absence of species than their actual abundances; that is, we replace all positive counts by 1 and leave the zeros as 0. The mean of 0/1 data is the probability p of a presence (i.e., a 1), so we write p_a, p_b, …, p_e, for the probabilities of the five species presence/absence variables. Logistic regression can be used to predict the dichotomous presence/absence response variables, given the explanatory variables. This is another GLM where the assumed distribution of the 0/1 data is binomial and the link function is the *logit*, or *log-odds*, function. The logit function is $\log(p/(1-p))$, abbreviated as $\text{logit}(p)$. Again, the fitting of this GLM is a simple option of the R `glm` function (see the Computational Appendix) and the estimated coefficients are listed in Exhibit 3.5.

Using species d once more as an example, the estimating equation is:

$$\text{logit}(p_d) = \log\left(\frac{p_d}{1-p_d}\right) = 2.712 - 1.177y^* - 0.137x^* \tag{3.4}$$

and the coefficients −1.177 and −0.137 estimate the changes in the log-odds of the probability of species d. Using the coefficients we can again make a biplot of the five species in the space of y^* and x^*, shown in Exhibit 3.6. This could be calibrated in units of odds, $p_d/(1-p_d)$, or transformed back to units of p_d as follows, thanks to the inverse transformation:

	Constant	y^*	x^*	*Error*
$\text{logit}(p_a)$	2.384	–2.889	0.863	0.464
$\text{logit}(p_b)$	1.273	–1.418	–0.143	0.756
$\text{logit}(p_c)$	0.831	0.973	0.315	0.911
$\text{logit}(p_d)$	2.712	–1.177	–0.137	0.798
$\text{logit}(p_e)$	0.253	–1.280	–0.786	0.832

Exhibit 3.5:
The regression coefficients for the five logistic regressions of the species responses on the predictors "pollution" y^* and "depth" x^*, showing their error deviances

$$p_d = \frac{\exp(2.712 - 1.177y^* - 0.137x^*)}{1 + \exp(2.712 - 1.177y^* - 0.137x^*)} \tag{3.5}$$

So for $y^* = x^* = 0$, $\exp(2.712) = 15.06$ and the estimated probability of d is $15.06/16.06 = 0.938$ (from Exhibit 2.1 species d occurs at 27 out of 30 sites, so its probability of a presence is high, but comes down mainly when y^* increases). So the origin of the map corresponds to an estimated p_d of 0.938. Where would the

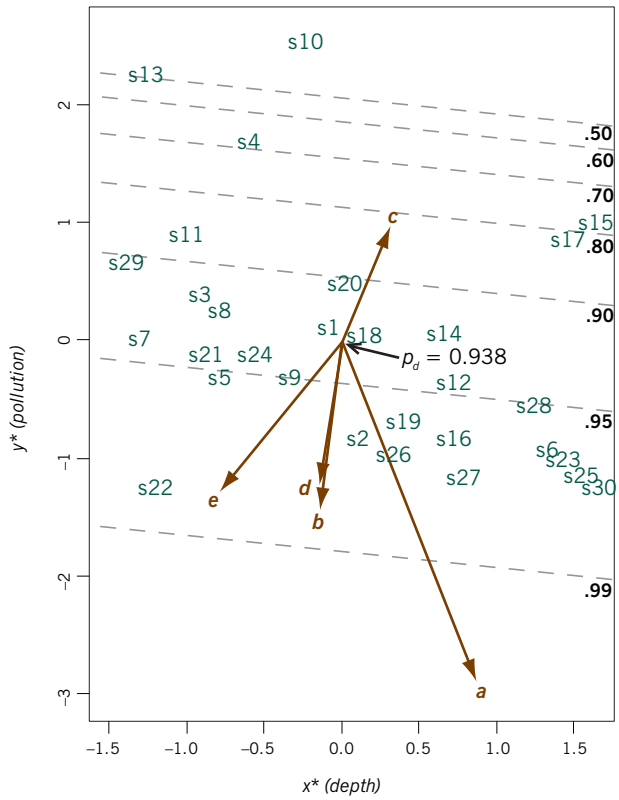


Exhibit 3.6:
Logistic regression biplot of the presence/absence data of the five species. The calibration for species ***d*** *is shown in the form of contours in units of predicted probability of presence. The scale is linear on the logit scale but non-linear on the probability scale, as shown*

tic mark be placed for 0.95? The corresponding logit is $\log(0.95/0.05) = 2.944$, which is $2.944 - 2.712 = 0.232$ units higher on the logit scale from the origin. The unit length is once more the inverse of the length of the biplot vector $1/\sqrt{1.177^2 + 0.137^2} = 0.844$, so the tic mark for 0.95 is at a distance $0.232 \times 0.844 = 0.196$ from the origin in the positive direction of *d*. Exhibit 3.6 shows the logistic regression biplot with the contours of the probability function for species *d*.

In a similar way the logistic regression surfaces could be indicated for each of the other species as a sequence of probability contour lines at right angles to the biplot vectors in Exhibit 3.6, where the origin corresponds to the probability for the means of the explanatory variables and the probability contours increase in the direction of the respective biplot vectors.

SUMMARY: Generalized Linear Model Biplots

1. A regression biplot can still be made if a nonlinear transformation of the response variable is performed: the effect is that the tic marks on the biplot axes are not at equal intervals, that is, the calibration is nonlinear.
2. *Generalized linear models* generalize linear regression to include different relationships between the conditional mean of the response variable and the explanatory variables as well as different distributions for the response variable. In each generalized linear model the conditional mean, transformed by the *link function*, is modelled as a linear function of the explanatory variables.
3. Examples of generalized linear models are *Poisson regression* (for count data), where the link function is the logarithm and the assumed distribution is Poisson; and *logistic regression* (for discrete responses), where the link function is the logit and the assumed distribution is binomial.

CHAPTER 4

Multidimensional Scaling Biplots

Multidimensional scaling is the graphical representation of a set of objects based on their interpoint distances. The method originated in psychological experiments where people were asked to judge the similarities of the objects, examples of which were a set of paintings, a set of writings, a set of countries or a set of products. This approach is often called *perceptual mapping* because it results in a spatial map of the respondents' perceptions of the set of objects. These maps are multidimensional, although they are usually represented in their best two-dimensional aspects, appearing like a scatterplot of the objects. In this map the horizontal and vertical axes are known as *principal axes*, which are artificially created to provide the support on which the objects are represented. If in addition there are variables characterizing the set of objects, then these variables can be added as biplot axes to the multidimensional scaling map.

Contents

Multidimensional scaling—data set "countries"

In the first class that I give in my postgraduate course "Methods of Marketing Research", I ask the students, who usually come from a number of different countries, to evaluate the similarities and differences between these countries on a scale from 1 (most similar) to 9 (most different). Exhibit 4.1 is an example of a table given by one of the students, with initials MT, for the 13 countries represented in that particular class. A low value given to a pair of countries, for example a 2 between Italy and Spain, means that MT perceives these countries as being very similar to one another, whereas a high value, for example a 9 between Russia and Spain, means that he perceives them to be very different. The idea in *multidimensional scaling* (MDS) is to represent the countries in a spatial map such that the physical

Exhibit 4.1: *Student MT's ratings of the similarities/differences between 13 countries, on a scale 1 = most similar to 9 = most different. The column labels are the international codes for the countries used in the MDS maps*

Countries	I	E	HR	BR	RU	D	TR	MA	PE	NG	F	MX	ZA
Italy	0	2	5	5	8	7	3	5	6	8	4	5	7
Spain	2	0	5	4	9	7	4	7	3	8	4	4	6
Croatia	5	5	0	7	4	3	6	7	7	8	4	6	7
Brazil	5	4	7	0	9	8	3	6	2	7	4	3	5
Russia	8	9	4	9	0	4	7	7	8	8	7	7	7
Germany	7	7	3	8	4	0	7	8	8	8	4	8	8
Turkey	3	4	6	3	7	7	0	5	4	5	6	4	5
Morocco	5	7	7	6	7	8	5	0	7	4	6	6	4
Peru	6	3	7	2	8	8	4	7	0	6	7	2	4
Nigeria	8	8	8	7	8	8	5	4	6	0	6	3	3
France	4	4	4	4	7	4	6	6	7	6	0	8	7
Mexico	5	4	6	3	7	8	4	6	2	3	8	0	4
South Africa	7	6	7	5	7	8	5	4	4	3	7	4	0

distances in the map approximate as closely as possible the values in the matrix. The way this approximation is measured and optimized distinguishes the different methods of MDS, but we do not enter into those details specifically here (see the Bibliographical Appendix for some recommended literature).

Classical scaling

The result of one approach, called classical MDS (function `cmdscale` in R—see Computational Appendix) is given in Exhibit 4.2. The countries are depicted as points and the distances between pairs of countries are approximations of the numbers in Exhibit 4.1. In this map we can see that Russia and Spain indeed turn out to be the furthest apart, while Italy and Spain appear close together, so at a first glance it seems like we have a good representation. We can approximately measure the interpoint distances in Exhibit 4.2 according to the scale shown on the sides, then the distances are always less than those in the table of ratings: for example, the distance between Italy and Spain in Exhibit 4.2 is about 1 unit whereas the given rating is 2. This is because classical scaling approximates the distances "from below"—the country points actually reside in a higher-dimensional space and have been projected onto a two-dimensional plane within this space. So all distances become shortened by this projection.

To measure how good the map is, a quality of approximation is measured in a similar way as it is done in regression analysis. In Exhibit 4.2 56.7% of the variance is accounted for. If we added a third dimension to the solution, depicting the countries in a three-dimensional map, a further 12.9% of the variance would be visualized, bringing the overall quality to 69.6%. In the Web Appendix a three-dimensional rotation of these country points is shown to illustrate the additional benefit of viewing the results in a three-dimensional space. For

our present purpose, however, we shall use the two-dimensional map. In Chapter 5 the topic of dimension reduction is explained more fully, with some technical details.

Principal axes

Exhibit 4.2 differs from the maps up to now (for example, Exhibits 2.5, 3.2 and 3.6) in one important respect: previously these maps were drawn using two observed variables, the (standardized) pollution and depth variables, whereas in MDS the axes on which the plot is constructed are so-called *principal axes*. These are not observed, but derived from the data with the objective of explaining the most variance possible: alternative names for the principal axes are *latent variables* or *factors*. As mentioned above, Exhibit 4.2 is the best view of the country points that can be achieved by projecting them onto a plane—in this plane the two axes are defined in order to be able to orientate the representation. These principal axes have the property that they are uncorrelated and the variance of the country points along each axis is equal to that part of the variance accounted for by that axis. The principal axes are also *nested*, which means that the first principal axis gives the best one-dimensional solution, explaining 33.3% of the variance in

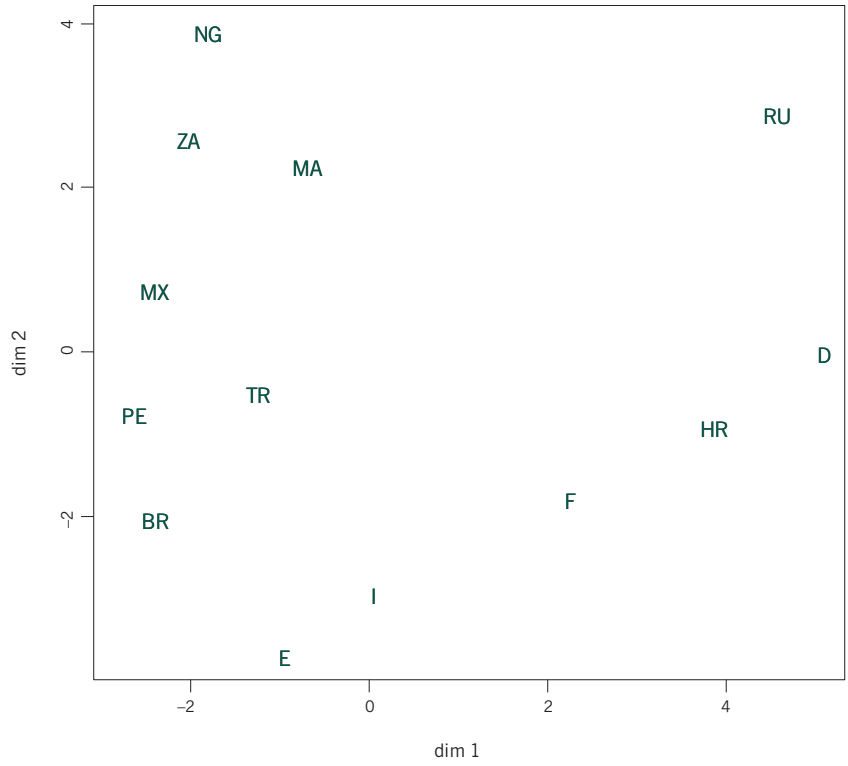


Exhibit 4.2: *MDS map of the 13 countries according to the ratings in Exhibit 4.1. The percentage of variance accounted for is 56.7%, with 33.3% on the first dimension, and 23.4% on the second*

Exhibit 4.3: *Student MT's ratings of the 13 countries on six attributes:* standard of living *(1 = low,…,9 = high);* climate *(1 = terrible,…, 9 = excellent);* food *(1 = awful,…, 9 = delicious);* security *(1 = dangerous,…, 9 = safe);* hospitality *(1 = friendly,…, 9 = unfriendly);* infrastructure *(1 = poor,…,9 = excellent). On the right are the coordinates of the country points in the MDS map of Exhibit 4.2*

Countries	*living*	*climate*	*food*	*security*	*hospitality*	*infrastructure*	*dim1*	*dim2*
Italy	7	8	9	5	3	7	0.01	–2.94
Spain	7	9	9	5	2	8	–1.02	–3.68
Croatia	5	6	6	6	5	6	3.70	–0.88
Brazil	5	8	7	3	2	3	–2.56	–2.01
Russia	6	2	2	3	7	6	4.41	2.91
Germany	8	3	2	8	7	9	5.01	0.00
Turkey	5	8	9	3	1	3	–1.38	–0.48
Morocco	4	7	8	2	1	2	–0.87	2.27
Peru	5	6	6	3	4	4	–2.77	–0.74
Nigeria	2	4	4	2	3	2	–1.97	3.91
France	8	4	7	7	9	8	2.18	–1.76
Mexico	2	5	5	2	3	3	–2.58	0.77
South Africa	4	4	5	3	3	3	–2.16	2.62

this example, the space defined by the first two principal axes gives the best two-dimensional solution, explaining 33.3 + 23.4 = 56.7% of the variance, and so on. Each principal axis simply builds on the previous ones to explain additional variance, but in decreasing amounts. This is identical to the situation in stepwise regression when all the explanatory variables are uncorrelated.

Multidimensional scaling biplot—data set "attributes"

Suppose now that we had additional variables about the 13 countries, which could be economic or social indicators or even further ratings by the same student. In fact, each student had to supply, in addition to the inter-country ratings, a set of ratings on six attributes, listed in Exhibit 4.3. The idea is now to relate these ratings to the MDS map in exactly the same way as we did before, and represent each of these attributes as a biplot vector. This will give us some idea of how these attributes relate to the general perception summarized in Exhibit 4.2. Each of the variables in Exhibit 4.3 is linearly regressed on the two dimensions of Exhibit 4.2 (the country coordinates used as predictors are given in Exhibit 4.3 as well), giving the regression coefficients in Exhibit 4.4.

The regression coefficients for the two dimensions again define biplot vectors which can be overlaid on the MDS plot—see Exhibit 4.5. Since the dimensions are centred in the MDS, the constants are the means for each attribute, situated at the origin of Exhibit 4.5. Each of the biplot axes through the biplot vectors could then be calibrated by working out what one unit is on its axis, as before. A unit will be inversely proportional to the length of the biplot vector, so the tic marks for "infrastructure", one of the longest vectors, will be closer together than those for "security", a shorter vector. Thus, even though both of

	Constant	*dim1*	*dim2*	R^2
Living	5.231	0.423	–0.513	0.754
Climate	5.692	–0.395	–0.618	0.693
Food	6.077	–0.399	-0.645	0.610
Security	4.000	0.502	–0.444	0.781
Hospitality	3.846	0.660	0.010	0.569
Infrastructure	4.923	0.627	–0.591	0.818

Exhibit 4.4:
The regression coefficients for the regressions of the six attributes on the two dimensions of the MDS solution in Exhibit 4.2, as well as the measure of fit (R^2) in each case

these vectors point in exactly the same direction, there will be more variance in the projections of the countries onto "infrastructure" than onto "security". Notice that "hospitality" is worded negatively, so that the biplot vector is pointing to the "unfriendly" end of the scale: "friendly" would point to the left. It seems that the perception of the student in separating the South American countries on the left is due to their friendly hospitality, and that Brazil is not only hospitable but has a good climate and food as well.

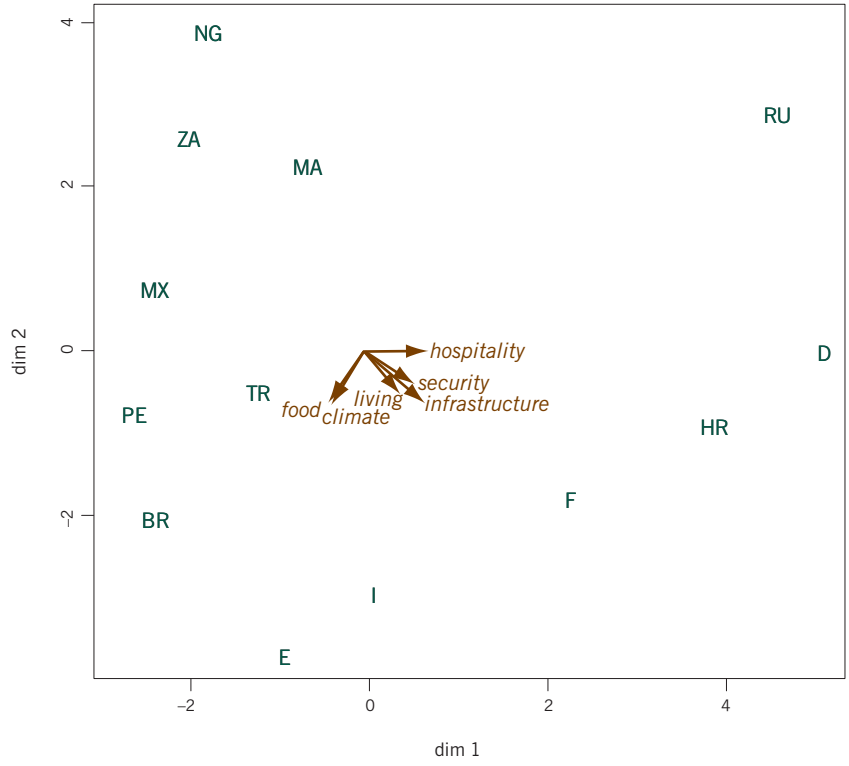


Exhibit 4.5:
MDS biplot, showing the countries according to the data of Exhibit 4.1 (i.e., the map of Exhibit 4.2), with the six attributes added as biplot vectors. Each biplot vector can be calibrated, as before, in its units from 1 to 9

Remember that the countries are positioned according to the student's overall perception of between-country differences, while the biplot vectors indicate how the ratings on specific attributes relate to this perception. The parts of variance explained (R^2) in Exhibit 4.4 show how well the attribute variables are represented. We can check some features of the map against the data in Exhibit 4.3 to get a qualitative idea of how well, or how badly, the biplot is performing. The three attributes "standard of living", "security" and "infrastructure" all point towards the European countries Germany, Croatia, France, Italy and Spain—these all fall above the averages for these three variables, but in decreasing value as one can see if one projects them onto the three corresponding biplot axes. To see if this agrees with the data, we average the three ratings for these variables, obtaining 8.3, 5.7, 7.7, 6.3 and 6.7 respectively, which are all above the average (which is equal to 4.7 in this case—Russia has an average of 5.0, which agrees with its position in the map, while all the other countries are below average and on the negative side of these three attributes). The value of 5.7 for Croatia is most out of line with the data—the general perception of Croatia, based on inter-country differences, places Croatia between Germany and France, but according to the attribute data Croatia should be lower down on the direction defined by these three correlating attributes. Hence the lack of fit, or unexplained variance, in the attributes in Exhibit 4.5 includes this "error" in their display with respect to Croatia. But, of course, Exhibit 4.5 is not designed to show the attributes optimally—these have been superimposed *a posteriori* on the map. In Chapter 6 we shall perform an analysis specifically of the attribute data, in which we will see the attributes represented with much higher R^2, and showing Croatia in a position more in accordance with them.

Chi-square distance biplot

We can anticipate the chapter on correspondence analysis (Chapter 8) by reconsidering data set "bioenv" of Exhibit 2.1. Previously we performed regressions of the five species variables on two of the concomitant variables "pollution" and "depth" and showed the results in the space of these two environmental variables. We now take the MDS biplot approach, performing an MDS of the 30 stations in terms of their species information and then show how the explanatory variables relate to the MDS map. The only decision we need to make is how to measure distance between the 30 stations—in contrast to the "countries" example above, the distances are not the original data but need to be calculated from the species data. Here the *chi-square distance* will be used, the distance function that is the basis of correspondence analysis. This distance is based on the relative frequencies of the species at each station and also adjusts the contribution of each species according to its average across the 30 stations. This will be explained more in Chapter 8, but to give one example, consider the distance between stations `s1` and `s2` (see Exhibit 2.1). The relative frequencies of the species at these two stations are, respectively, [0 0.074 0.333 0.519 0.074] and [0.481 0.074 0.241 0.204 0]—for

example, for station s2, a count of 26 for species *a* is 0.481 of the total of 54 individuals counted at that station (the second row total). The totals for each species (column totals) have relative frequencies [0.303 0.196 0.189 0.245 0.067], showing that species *a* is the most common and species *e* the most rare. The chi-square distance between the two stations is computed as:

$$\sqrt{\frac{(0-0.481)^2}{0.303}+\frac{(0.074-0.074)^2}{0.196}+\frac{(0.333-0.241)^2}{0.189}+\frac{(0.519-0.204)^2}{0.245}+\frac{(0.074-0)^2}{0.067}}=1.139$$

The division of each squared difference by the overall species proportion is a form of standardization of the relative frequencies. There are bigger differences between the more common species and smaller differences between rare species, and the standardization serves to compensate for this natural variability found in frequency data. Having computed the 30 × 30 chi-square distance matrix between the 30 stations, the MDS procedure leads to a visualization of these distances, shown in Exhibit 4.6. Then, by performing the regressions of the species variables

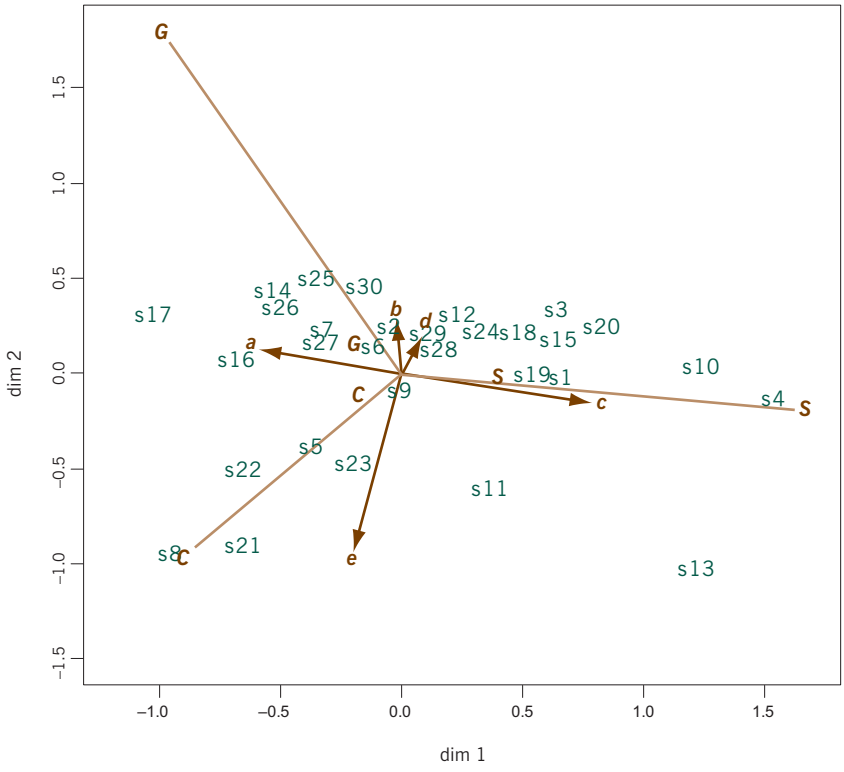


Exhibit 4.6:
MDS biplot, showing approximate chi-square distances between sites, upon which are added the biplot vectors of the five species (using linear regression), biplot vectors of the three sediment types (using logistic regression) and the averages of the stations according to the three sediment types

on the two dimensions of the map, these can be depicted on the map (here we use the relative frequencies, divided by the square roots of their respective averages, to be consistent with the way the chi-square distances were calculated). In addition, the three categories of the variable sediment can be shown, either as the averages of the site points in for each category or using the logistic regression biplot. Notice once more that there are two levels of error in this biplot: first, the chi-square distances between sites are not perfectly displayed (74.4% explained in the map—52.4% on dimension 1, 22.0% on dimension 2—i.e., 25.6% error); and second, the two dimensions only explain parts of variance of each species (*a*: 76.7%, *b*: 18.6%, *c*: 92.1%, *d*: 14.6%, *e*: 95.8%) and of each category of sediment (expressed as percentages of deviance explained in the logistic regressions, *C*: 7.3%, *G*: 11.8%, *S*: 15.8%).

SUMMARY: Multidimensional Scaling Biplots

1. Multidimensional scaling (MDS) is a method that represents a set of objects as a set of points in a map of chosen dimensionality, usually two-dimensional, based on their given interpoint distances. The objective is to maximize the agreement between the displayed interpoint distances and the given ones.
2. Any variable observed on the same set of objects can be superimposed on such a map using the regression coefficients obtained from the regression of the variable (or its standardized equivalent) on the dimensions of the MDS. The resultant joint plot is a biplot: the objects can be projected onto the biplot vectors of the variables to approximate the values of the variables. The optional standardization of the variable only changes the lengths of the biplot vectors, not their directions.
3. There are two different levels of error in the display. First, there is the error incurred in the MDS, because the distances are not perfectly displayed. Second, there are the errors in the regressions of the variables on the dimensions.

CHAPTER 5

Reduced-Dimension Biplots

In the previous chapter, multidimensional scaling (MDS) involved *reduction of dimensionality* in order to visualize a high-dimensional configuration of points in a low-dimensional representation. In some high-dimensional space distances between points are exact (or as near exact as possible for non-Euclidean dissimilarity measures), while they are approximated in some optimal sense in the low-dimensional version. In this chapter we look at the theory and practice of dimensionality reduction, and how a data matrix of a certain dimensionality can be optimally approximated by a matrix of lower, or reduced, dimensionality. Algebraically, the geometric concept of dimensionality is equivalent to the *rank* of a matrix, hence this chapter could also be called *reduced-rank* biplots. This topic is concentrated almost entirely on one of the most useful results in matrix algebra, the *singular value decomposition* (SVD). Not only does this result provide us with a solution to the optimal reduced-rank approximation of a matrix, but it also gives the coordinate values of the points in the corresponding biplot display.

Contents

Matrix approximation

Data matrices usually have many rows (cases) and many columns (variables), such as the 13 × 6 matrix of Exhibit 4.3. The *rank* of a matrix is the minimum number of row or column vectors needed to generate the rows or columns of the matrix exactly through linear combinations. Geometrically, this algebraic concept is equivalent to the *dimensionality* of the matrix—if we were lucky enough to have a data matrix of rank 2, then we could represent the rows or columns in a two-dimensional plot. In practice, however, no large matrix is of low rank, but we can

approximate it optimally by a matrix of low rank and then view this approximate matrix in a low-dimensional space.

Suppose that $\mathbf{Y}$ is an $n \times m$ matrix with rank r (in most examples in practice, r will be n or m, whichever is the smaller). Then the idea of matrix approximation is to find another $n \times m$ matrix $\hat{\mathbf{Y}}$ of lower rank $p < r$ that resembles $\mathbf{Y}$ as closely as possible. Closeness can be measured in any reasonable way, but least-squares approximation makes the solution of the problem particularly simple. Hence we want to find a matrix $\hat{\mathbf{Y}}$ that minimizes the following objective function over all possible rank p matrices:

$$\underset{\hat{\mathbf{Y}} \text{ of rank } p}{\text{minimize}} \ \text{trace}[(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^{\mathsf{T}}] = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\hat{y}_{ij})^2 \tag{5.1}$$

The notation trace[…] signifies the sum of the diagonal elements of a square matrix, and the square matrix $(\mathbf{Y}-\hat{\mathbf{Y}})(\mathbf{Y}-\hat{\mathbf{Y}})^{\mathsf{T}}$ has exactly all the squared differences between the corresponding elements of $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ down its diagonal. Thanks to our choosing a least-squares objective function, this minimization problem is very simple to solve using a famous result in matrix algebra.

Singular value decomposition (SVD)

The singular value decomposition, or SVD for short, is one of the most useful results in matrix theory. Not only will it provide us with the solution of the matrix approximation problem described above, but it also provides the solution in exactly the form that is required for the biplot. The basic result is as follows: any rectangular $n \times m$ matrix $\mathbf{Y}$, of rank r, can be expressed as the product of three matrices:

$$\mathbf{Y} = \mathbf{U}\,\mathbf{D}_{\alpha}\mathbf{V}^{\mathsf{T}} \tag{5.2}$$

where $\mathbf{U}$ is $n \times r$, $\mathbf{V}$ is $m \times r$ and $\mathbf{D}_{\alpha}$ is a $r \times r$ diagonal matrix with positive numbers $\alpha_1, \alpha_2, \ldots, \alpha_r$, on the diagonal in descending order: $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_r > 0$. Furthermore, the columns of $\mathbf{U}$ and of $\mathbf{V}$ are *orthonormal*, by which we mean that they have unit length (sum of squares of their elements = 1) and are orthogonal, or perpendicular to one another (i.e., scalar products between columns = 0, that is they are geometrically perpendicular to one another); this property can be written as $\mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$ (where $\mathbf{I}$ denotes the identity matrix, a diagonal matrix with 1's down the diagonal). The columns $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r$, and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r$, of $\mathbf{U}$ and $\mathbf{V}$ are called *left* and *right singular vectors* respectively, and the values $\alpha_1, \alpha_2, \ldots, \alpha_r$ the *singular values* of $\mathbf{Y}$.

If the rank of $\mathbf{Y}$ happened to be low, say 2 or 3, then (5.2) would give us immediately the form "target matrix = left matrix · right matrix" of the biplot we need

(see (1.2)) for a two- or three-dimensional display, the only decision being how to distribute the matrix $\mathbf{D}_\alpha$ of singular values to the left and the right in order to define the biplot's left matrix and right matrix (we shall discuss this critical decision at length soon). In the more usual case that the rank of $\mathbf{Y}$ is much higher than 2 or 3, then the SVD provides us immediately with any low-rank matrix approximation we need, as follows. Define $\hat{\mathbf{Y}}$ as in (5.2) but use only the first p columns of $\mathbf{U}$, the upper left $p \times p$ part of $\mathbf{D}_\alpha$ and the first p columns of $\mathbf{V}$, in other words the first p components of the SVD: $\hat{\mathbf{Y}} = \mathbf{U}_{[p]}\,\mathbf{D}_{\alpha\,[p]}\mathbf{V}_{[p]}^{\mathsf{T}}$, where the subindex $_{[p]}$ means the "first p components". $\hat{\mathbf{Y}}$ is of rank p and is exactly the solution to the least-squares matrix approximation problem. And, once more, the decomposition provided by the SVD is exactly in the form that we need for the biplot.

The singular values provide us with quantifications of the closeness of the approximation of $\hat{\mathbf{Y}}$ to $\mathbf{Y}$. The sum-of-squares of the singular values is equal to the sum-of-squares of the matrix $\mathbf{Y}$: $\text{trace}(\mathbf{Y}\mathbf{Y}^{\mathsf{T}}) = \sum_i \sum_j y_{ij}^2 = \alpha_1^2 + \alpha_2^2 + \ldots + \alpha_r^2$, and the sum-of-squares of the matrix $\hat{\mathbf{Y}}$ is $\text{trace}(\hat{\mathbf{Y}}\hat{\mathbf{Y}}^{\mathsf{T}}) = \alpha_1^2 + \alpha_2^2 + \ldots + \alpha_p^2$, the sum of the first p squared singular values. The latter is expressed as a fraction (or percentage) of the first to quantify the quality of the approximation, while the remainder from 1 quantifies the error (5.1) of the approximation.

Some numerical examples

Consider again the target matrix on the left-hand side of (1.1) and let us pretend we do not know that it decomposes as shown there. The SVD of this matrix calculated in R using the built-in function `svd` in the second command below:

```
> Y<-matrix(c(8,5,-2,2,4,2,0,-3,3,6,2,3,3,-3,-6,-6,-4,1,-1,-2),
  nrow=5)
> svd(Y)
$d
[1]    1.412505e+01  9.822577e+00  6.351831e-16  3.592426e-33

$u
             [,1]         [,2]          [,3]           [,4]
[1,]   -0.6634255   -0.4574027   -0.59215653    2.640623e-35
[2,]   -0.3641420   -0.4939878    0.78954203    2.167265e-34
[3,]    0.2668543   -0.3018716   -0.06579517   -9.128709e-01
[4,]   -0.2668543    0.3018716    0.06579517   -1.825742e-01
[5,]   -0.5337085    0.6037432    0.13159034   -3.651484e-01

$v
             [,1]         [,2]         [,3]         [,4]
[1,]   -0.7313508   -0.2551980   -0.6276102   -0.0781372
[2,]   -0.4339970    0.4600507    0.2264451    0.7407581
[3,]    0.1687853   -0.7971898    0.0556340    0.5769791
[4,]    0.4982812    0.2961685   -0.7427873    0.3350628
```

The `svd` function returns the three parts of the decomposition: the singular values in `$d`, the left singular vectors in `$u` and the right singular vectors in `$v`. It is clear from the singular values that only the first two are nonzero, so the matrix is of rank 2 and can be written as (showing values to 4 decimal places):

$$\begin{pmatrix} 8 & 2 & 2 & -6 \\ 5 & 0 & 3 & -4 \\ -2 & -3 & 3 & 1 \\ 2 & 3 & -3 & -1 \\ 4 & 6 & -6 & -2 \end{pmatrix} = \begin{pmatrix} -0.6634 & -0.4574 \\ -0.3641 & -0.4940 \\ 0.2669 & -0.3019 \\ -0.2669 & 0.3019 \\ -0.5337 & 0.6037 \end{pmatrix} \begin{pmatrix} 14.1251 & 0 \\ 0 & 9.8226 \end{pmatrix} \begin{pmatrix} -0.7314 & -0.4340 & 0.1688 & 0.4983 \\ -0.2552 & 0.4601 & -0.7972 & 0.2962 \end{pmatrix}$$

To define a left and right matrix for the biplot, we can—for example—split the singular values in the middle equally between the left and right singular vectors. That is, multiply the two left singular vectors (two columns above) and the two right singular vectors (two rows above) by the square roots $\sqrt{14.1251} = 3.7583$ and $\sqrt{9.8226} = 3.1341$ respectively. (This way of splitting the singular values equally between the left and right vectors leads to the so-called "symmetric biplot"). This gives the following biplot solution and corresponding plot in Exhibit 5.1, which is:

$$\begin{pmatrix} 8 & 2 & 2 & -6 \\ 5 & 0 & 3 & -4 \\ -2 & -3 & 3 & 1 \\ 2 & 3 & -3 & -1 \\ 4 & 6 & -6 & -2 \end{pmatrix} = \begin{pmatrix} -2.4934 & -1.4335 \\ -1.3686 & -1.5482 \\ 1.0029 & -0.9461 \\ -1.0029 & 0.9461 \\ -2.0059 & 1.8922 \end{pmatrix} \begin{pmatrix} -2.7487 & -1.6311 & 0.6344 & 1.8727 \\ -0.7998 & 1.4418 & -2.4985 & 0.9282 \end{pmatrix}$$

Generalized matrix approximation and SVD

In (5.1) the problem of minimizing fit to a given matrix by another of lower rank was formulated. The idea can be generalized to include a system of weighting on both the rows and columns of the table, the objective being to give them differential importance in the fitting process. For example, in survey analysis the rows are respondents that are often not representative of the population from which they are sampled. If there are proportionally too many women, say, in the sample, then giving lower weights to the individual female respondents can restore the representativeness in the sample. The same is true for the column variables: there are many reasons why some variables may need to be downweighted, for example their variance is by their very nature too high, or there are several variables that basically measure the same trait in the population. The idea of weighting can be carried to the limit of giving zero weight to some respondents or variables—this is the idea behind supplementary points, which will be explained in future chapters.

Suppose then that we have a set of positive weights $w_1, w_2, \ldots, w_n$ for the rows of a matrix and a set of positive weights $q_1, q_2, \ldots, q_m$ for the columns. We can as-

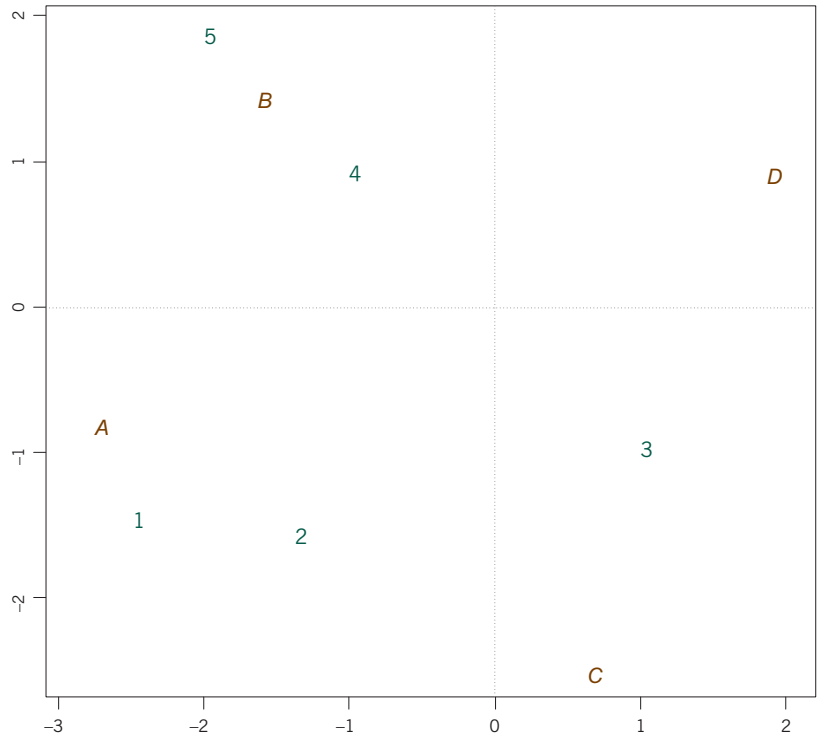


Exhibit 5.1:
Symmetric biplot of the rank 2 example, rows labelled 1 to 5, columns A to D. The square roots of the singular values are assigned to both left and right singular vectors to establish the left and right coordinate matrices. The row–column scalar products perfectly reproduce the original target matrix

sume that the weights add up 1 in each case. Then, rather than the objective (5.1), the weighted (or generalized) matrix approximation is formulated as follows:

$$\underset{\hat{\mathbf{Y}} \text{ of rank } p}{\text{minimize}} \ \text{trace}[\mathbf{D}_w(\mathbf{Y}-\hat{\mathbf{Y}})\mathbf{D}_q(\mathbf{Y}-\hat{\mathbf{Y}})^{\mathsf{T}}] = \sum_{i=1}^{n}\sum_{j=1}^{m} w_i q_j (y_{ij}-\hat{y}_{ij})^2 \qquad (5.3)$$

This solution involves a weighted (or generalized) SVD, which can be solved using the usual (unweighted) SVD as follows: (1) pre-transform the matrix $\mathbf{Y}$ by multiplying its rows and columns by the square roots of the weights, (2) perform the SVD on this transformed matrix as in (5.2), and (3) "untransform" the left and right singular vectors by the inverse square roots of the respective row and column weights. These three steps can be expressed as follows:

$$(1)\ \mathbf{S} = \mathbf{D}_w^{1/2}\mathbf{Y}\mathbf{D}_q^{1/2} \qquad (5.4)$$

$$(2)\ \mathbf{S} = \mathbf{U}\mathbf{D}_\beta\mathbf{V}^{\mathsf{T}} \qquad (5.5)$$

$$(3)\ \tilde{\mathbf{U}} = \mathbf{D}_w^{-1/2}\mathbf{U}, \text{ and } \tilde{\mathbf{V}} = \mathbf{D}_q^{-1/2}\mathbf{V} \qquad (5.6)$$

The best-fitting matrix of rank p, which minimizes (5.3), is calculated as before, but using $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$:

$$\hat{\mathbf{Y}} = \tilde{\mathbf{U}}_{[p]}\mathbf{D}_{\beta[p]}\tilde{\mathbf{V}}_{[p]}^{\mathsf{T}} \tag{5.7}$$

Notice that $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ have columns (the generalized singular vectors) that are of unit length and orthogonal in terms of weighted sum of squares and weighted sum of cross-products: $\tilde{\mathbf{U}}^{\mathsf{T}}\mathbf{D}_w\tilde{\mathbf{U}} = \tilde{\mathbf{V}}^{\mathsf{T}}\mathbf{D}_q\tilde{\mathbf{V}} = \mathbf{I}$. The singular values $\beta_1, \beta_2, \ldots, \beta_p$ in (5.7) are then split between the left and right singular vectors to obtain the left and right matrices in the biplot, either by assigning their square roots to the left and right, or by assigning the singular values completely to either the left or the right.

Generalized principal component analysis

The introduction of weights into the matrix approximation broadens the class of methods that can be defined in terms of the SVD. All the biplots of interest turn out to be special cases, depending on the definition of the matrix $\mathbf{Y}$ to be approximated, and the weights $\mathbf{w}$ and $\mathbf{q}$ assigned to the rows and columns. A useful general method, which we call generalized *principal component analysis* (PCA), includes almost all the techniques to be described in the rest of this book. In this definition we think of the matrix either as a set of rows or as a set of columns—we shall assume that we think of the rows as points in a multidimensional space.

Suppose that the rows of $\mathbf{X}$ ($n \times m$) define n points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, in m-dimensional space—notice that vectors are always denoted as column vectors, so that

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix}$$

The points have weights in the $n \times 1$ vector $\mathbf{w}$, where the weights are positive and sum to 1: $\mathbf{1}^{\mathsf{T}}\mathbf{w} = 1$ ($\mathbf{1}$ is an appropriate vector of ones in each case). Distances in the m-dimensional space are defined by a weighted metric where the dimensions are weighted by the positive elements of the $m \times 1$ vector $\mathbf{q}$: for example, the square of the distance between the i-th and i'-th rows $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ is $(\mathbf{x}_i - \mathbf{x}_{i'})^{\mathsf{T}}\mathbf{D}_q(\mathbf{x}_i - \mathbf{x}_{i'})$. The objective is to find a low-dimensional version of the rows of $\mathbf{X}$ which are the closest to the original ones in terms of weighted least-squared distances.

There is a side result which proves that the low-dimensional solution necessarily includes the centroid (weighted average) of the points, so we can centre all the points beforehand. This is easily proved by assuming that the low-dimensional so-

lution does not include the centroid and then arrive at a contradictory conclusion. The centroid can, in fact, be thought of as the closest "zero-dimensional subspace" (i.e., a point) to the n points. This means that we first centre the row points by subtracting their centroid $\mathbf{w}^{\mathsf{T}}\mathbf{X}$:

$$\mathbf{Y} = \mathbf{X} - \mathbf{1}\mathbf{w}^{\mathsf{T}}\mathbf{X} = (\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathsf{T}})\mathbf{X} \tag{5.8}$$

The matrix $(\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathsf{T}})$ is called the *centring matrix*: the rows of $\mathbf{Y}$ now have a centroid of $\mathbf{0}$: $\mathbf{w}^{\mathsf{T}}\mathbf{Y} = \mathbf{0}$.

To find an approximating matrix $\hat{\mathbf{Y}}$ of rank $p < m$, the rows of which come closest to the rows of $\mathbf{Y}$ in terms of weighted sum of squared distances, we need to solve the following:

$$\underset{\hat{\mathbf{Y}} \text{ of rank } p}{\text{minimize}} \ \sum_{i=1}^{n} w_i(\mathbf{y}_i - \hat{\mathbf{y}}_i)^{\mathsf{T}}\mathbf{D}_q(\mathbf{y}_i - \hat{\mathbf{y}}_i) \tag{5.9}$$

which is identical to (5.3). Hence the solution is contained in the generalized SVD described above, with the matrix approximation given by (5.7). The coordinates of the row points in the low-dimensional display are given by

$$\mathbf{F} = \tilde{\mathbf{U}}_{[p]}\mathbf{D}_{\beta[p]} \tag{5.10}$$

called the *principal coordinates* (of the rows in this case), which thus form the left matrix of the biplot. The right matrix representing the columns is then $\tilde{\mathbf{V}}_{[p]}$, defining biplot axes. The singular values are thus assigned totally to the singular vectors corresponding to the rows in this case. Most of the subsequent chapters will deal with applications of this theory.

Classical multidimensional scaling with weighted points

The MDS problem of Chapter 4 can be formulated as a SVD problem as well, in fact the matrix decomposed is square symmetric and the SVD reduces to its special case, the eigenvalue/eigenvector decomposition, or *eigendecomposition*. The general formulation for the case when points are weighted is as follows:

- Suppose that the matrix of squared distances between n objects is denoted by $\mathbf{D}^{(2)}$ and that the objects are weighted by the n positive elements in the vector $\mathbf{w}$.
- Double-centre $\mathbf{D}^{(2)}$ using the weights in $\mathbf{w}$ (the centring matrix is $\mathbf{I} - \mathbf{1}\mathbf{w}^{\mathsf{T}}$, pre-multiplied to centre the rows, or transposed and post-multiplied to centre the columns), weight the points by pre- and post-multiplying by $\mathbf{D}_w^{½}$, and finally multiply the result by $-½$ before calculating the eigendecomposition

$$\mathbf{S} = -\tfrac{1}{2}\,\mathbf{D}_w^{\frac{1}{2}}(\mathbf{I} - \mathbf{1}\mathbf{w}^\top)\mathbf{D}^{(2)}(\mathbf{I} - \mathbf{1}\mathbf{w}^\top)^\top\mathbf{D}_w^{\frac{1}{2}} = \mathbf{U}\mathbf{D}_\lambda\mathbf{U}^\top \tag{5.11}$$

– Calculate the coordinates of the points: $\mathbf{F} = \mathbf{D}_w^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_\lambda^{\frac{1}{2}}$ (5.12)

If we start off with a matrix $\mathbf{X}$ as in the previous section, where squared distances between rows are calculated in the metric $\mathbf{D}_q$, with points weighted by $\mathbf{w}$, then the above algorithm gives the same coordinates as the principal coordinates in (5.10), and the eigenvalues here are the squared singular values in the generalized PCA: $\lambda_k = \beta_k^2$.

SUMMARY: Reduced-Dimension Biplots

1. Reduced-dimension biplots rely on approximating a matrix of high dimensionality by a matrix of lower dimensionality. The matrix of low dimensionality is then the target matrix for the biplot.
2. If approximation of a matrix is performed using a least-squares objective, then the singular value decomposition (SVD) provides the solution in a very convenient form as the product of three matrices: the left and right matrices of the biplot are then provided by distributing the second matrix of the SVD (the diagonal matrix of singular values) to the first and third matrices (of singular vectors).
3. The objective of least-squares matrix approximation can be generalized to include weights for the rows and the columns. This leads to a simple modification of the SVD, called the generalized SVD, involving pre-transformation of the matrix to be approximated and post-transformation of the singular vectors.
4. Generalized principal component analysis (PCA) is a geometric version of matrix approximation, where a set of n vectors in m-dimensional space is projected onto a subspace of lower dimensionality. The resultant reduced-dimension biplot depicts the approximate positions of the n points along with m directions showing the biplot axes.
5. Multidimensional scaling (MDS), including the general case where points have any positive weights, can also be formulated as an eigenvalue/eigenvector special case of the SVD problem, because the matrix decomposed is square and symmetric. The resultant coordinates are identical to those found in generalized PCA, if the interpoint distances are defined using the same metric.

CHAPTER 6

Principal Component Analysis Biplots

Principal component analysis (PCA) is one of the most popular multivariate methods in a wide variety of research areas, ranging from physics to genomics and marketing. The origins of PCA can be traced back to early 20th century literature in biometrics (Karl Pearson) and psychometrics (Harold Hotelling). The method is inextricably linked to the singular value decomposition (SVD)—this powerful result in matrix theory provides the solution to the classical PCA problem and, conveniently for us, the solution is in a format leading directly to the biplot display. In this section we shall consider various applications of PCA and interpret the associated biplot displays. We will also introduce the notion of the contribution biplot, which is a variation of the biplot that will be especially useful when the rows and/or columns have different weights.

Contents

PCA of data set "attributes"

The last section of Chapter 5 defined a generalized form of PCA where rows and columns were weighted. If we consider the 13×6 data matrix of Exhibit 4.3, there is no need to differentially weight the rows or the columns: on the one hand, the countries should be treated equally, while on the other hand, the variables are all on the same 1 to 9 scale, and so there is no need to up- or downweight any variable with respect to the others (if variables were on different scales, the usual way to equalize out their roles in the analysis is to standardize them). So in this

example all rows and all columns obtain the same weight, i.e. $\mathbf{w} = (1/13)\mathbf{1}$ and $\mathbf{q} = (1/6)\mathbf{1}$, where $\mathbf{1}$ is an appropriate vector of ones in each case. Referring to (5.9), the matrix $\mathbf{D}_q$ defines the metric between the row points (i.e., the countries in this example), so that distances between countries are the average squared differences between the six variables.

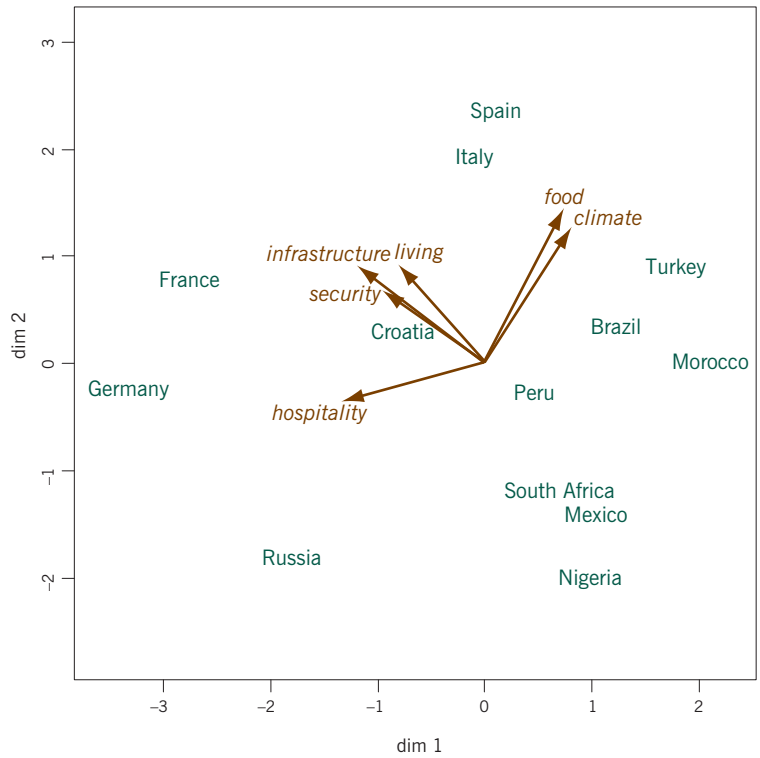
The computational steps are then the following, as laid out in the last section of Chapter 5. Here we use the general notation of a data matrix $\mathbf{X}$ with I rows and J columns, so that for the present example $I = 13$ and $J = 6$.

– Centring (cf. (5.8)): $\mathbf{Y} = [\mathbf{I} - (1/I)\mathbf{1}\mathbf{1}^{\mathsf{T}}]\mathbf{X}$ (6.1)

– Weighted SVD (cf. (5.4) and (5.5)):
$$\mathbf{S} = (1/I)^{1/2}\mathbf{Y}(1/J)^{1/2} = (1/IJ)^{1/2}\mathbf{Y} = \mathbf{U}\mathbf{D}_\beta\mathbf{V}^{\mathsf{T}} \quad (6.2)$$

– Calculation of coordinates; i.e., the left and right cf. (5.6) and (5.10)):
$$\mathbf{F} = I^{1/2}\mathbf{U}\mathbf{D}_\beta \text{ and } \mathbf{\Gamma} = J^{1/2}\mathbf{V} \quad (6.3)$$

Exhibit 6.1:
PCA biplot of the data in Exhibit 4.3, with the rows in principal coordinates, and the columns in standard coordinates, as given in (6.3). This is the row-metric-preserving biplot, or form biplot (explained on following page). Remember that the question about hospitality was worded negatively, so that the pole "friendly" is in the opposite direction to the vector "hospitality"—see Exhibit 4.3



We use the term "weighted SVD" above even though there is no differential weighting of the rows and columns: the whole of the centred matrix $\mathbf{Y}$ is simply multiplied by a constant, $(1/IJ)^{½}$. The resultant biplot is given in Exhibit 6.1.

Principal and standard coordinates

When the singular values are assigned totally to the left or to the right, the resultant coordinates are called *principal coordinates*. The other matrix, to which no part of the singular values is assigned, contains the so-called *standard coordinates*. The choice made in (6.3) is thus to plot the rows in principal coordinates and the columns in standard coordinates. From (6.3) it can be easily shown that the principal coordinates on a particular dimension have average sum of squared coordinates equal to the square of the corresponding singular value; for example, for $\mathbf{F}$ defined in (6.3):

$$\mathbf{F}^{\mathsf{T}}\mathbf{D}_w\mathbf{F} = (1/I)\ (I^{½}\mathbf{U}\mathbf{D}_\beta)^{\mathsf{T}}(I^{½}\mathbf{U}\mathbf{D}_\beta) = \mathbf{D}_\beta^{\mathsf{T}}\mathbf{U}^{\mathsf{T}}\mathbf{U}\mathbf{D}_\beta = \mathbf{D}_\beta^2 \tag{6.4}$$

By contrast, the standard coordinates on a particular dimension have average sum of squared coordinates equal to 1 (hence the term "standard"); for example, for $\mathbf{\Gamma}$ defined in (6.3):

$$\mathbf{\Gamma}^{\mathsf{T}}\mathbf{D}_q\mathbf{\Gamma} = (1/J)\,(J^{½}\mathbf{V})^{\mathsf{T}}(J^{½}\mathbf{V}) = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I} \tag{6.5}$$

In this example the first two squared singular values are $\beta_1^2 = 2.752$ and $\beta_2^2 = 1.665$. If $\mathbf{F}$ has elements f_{ik}, then the normalization of the row principal coordinates on the first two dimensions is $(1/13)\sum_i f_{i1}^2 = 2.752$ and $(1/13)\sum_i f_{i2}^2 = 1.665$. For the column standard coordinates γ_{jk} the corresponding normalizations are $(1/6)\sum_j \gamma_{j1}^2 = 1$ and $(1/6)\sum_j \gamma_{j2}^2 = 1$.

Form biplot

There are various names in the literature for this type of biplot. It can be called the *row-metric-preserving biplot*, since the configuration of the row points approximates the interpoint distances between the rows of the data matrix. It is also called the *form biplot*, because the row configuration is an approximation of the *form matrix*, composed of all the scalar products $\mathbf{Y}\mathbf{D}_q\mathbf{Y}^{\mathsf{T}}$ of the rows of $\mathbf{Y}$:

$$\mathbf{Y}\mathbf{D}_q\mathbf{Y}^{\mathsf{T}} = \mathbf{F}\mathbf{\Gamma}^{\mathsf{T}}\mathbf{D}_q\mathbf{\Gamma}\mathbf{F}^{\mathsf{T}} = \mathbf{F}\mathbf{F}^{\mathsf{T}} \tag{6.6}$$

In fact, it is the form matrix which is being optimally represented by the row points, and—by implication—the inter-row distances which depend on the scalar products.

Covariance biplot

If the singular values are assigned totally to the right singular vectors in (6.2), then we get an alternative biplot called the *covariance biplot*, because it shows the inter-variable covariance structure. It is also called the *column-metric-preserving biplot*. The left and right matrices are then defined as (cf. (6.3)):

– Coordinates in covariance biplot: $\mathbf{\Phi} = I^{1/2}\mathbf{U}$ and $\mathbf{G} = J^{1/2}\mathbf{V}\mathbf{D}_{\beta}$ (6.7)

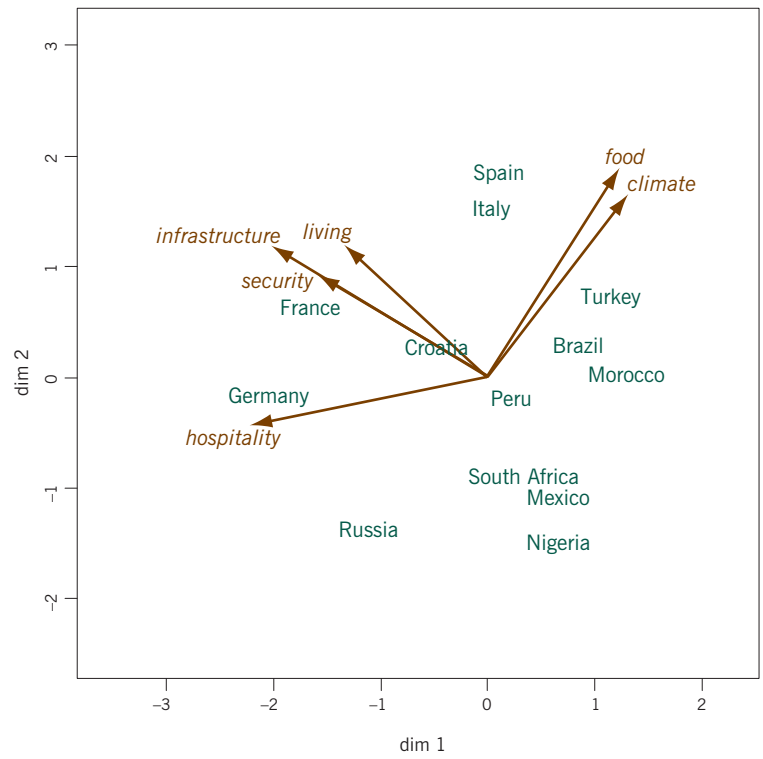
The covariance biplot is shown in Exhibit 6.2.

Apart from the changes in scale along the respective principal axes in the row and column configurations, this biplot hardly differs from the form biplot in Exhibit 6.1. In Exhibit 6.1 the countries have weighted sum of squares with respect to each axis equal to the corresponding squared singular value, while in Exhibit 6.2 it is the attributes that have weighted sum of squares equal to the squared singular values. In each biplot the other set of points has unit normalization on both principal axes.

In the covariance biplot the covariance matrix $\mathbf{Y}^{\mathsf{T}}\mathbf{D}_w\mathbf{Y} = (1/I)\mathbf{Y}^{\mathsf{T}}\mathbf{Y}$ between the variables is equal to the scalar product matrix between the column points using all the principal axes:

$$\mathbf{Y}^{\mathsf{T}}\mathbf{D}_w\mathbf{Y} = \mathbf{G}\mathbf{\Phi}^{\mathsf{T}}\mathbf{D}_w\mathbf{\Phi}\mathbf{G}^{\mathsf{T}} = \mathbf{G}\mathbf{G}^{\mathsf{T}} \tag{6.8}$$

Exhibit 6.2:
PCA biplot of the data in Exhibit 4.3, with the columns in principal coordinates, and the rows in standard coordinates, as given in (6.7). This is the column-metric-preserving biplot, or covariance biplot



and is thus approximated in a low-dimensional display using the major principal axes. Hence, the squared lengths of the vectors in Exhibit 6.2 approximate the variances of the corresponding variables—this approximation is said to be "from below", just like the approximation of distances in the classical scaling of Chapter 4. By implication it follows that the lengths of the vectors approximate the standard deviations and also that the cosines of the angles between vectors approximate the correlations between the variables. If the variables in $\mathbf{Y}$ were normalized to have unit variances, then the lengths of the variable vectors in the biplot would be less than one—a unit circle may then be drawn in the display, with vectors extending closer to the unit circle indicating variables that are better represented.

Connection with regression biplots

It can be easily shown that in both the form and covariance biplots the coordinates of the variables, usually depicted as vectors, are the regression coefficients of the variables on the dimensions of the biplot. For example, in the case of the covariance biplot, the regression coefficients of $\mathbf{Y}$ on $\mathbf{\Phi} = I^{\frac{1}{2}}\mathbf{U}$ are:

$$(\mathbf{\Phi}^{\mathsf{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathsf{T}}\mathbf{Y} = (I\,\mathbf{U}^{\mathsf{T}}\mathbf{U})^{-1}(I^{\frac{1}{2}}\mathbf{U})^{\mathsf{T}}(IJ)^{\frac{1}{2}}\mathbf{U}\mathbf{D}_{\beta}\mathbf{V}^{\mathsf{T}} \qquad \text{(from (6.2))}$$

$$= J^{\frac{1}{2}}\mathbf{D}_{\beta}\mathbf{V}^{\mathsf{T}} \qquad \text{(because } \mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I}\text{)}$$

which is the transpose of the coordinate matrix $\mathbf{G}$ in (6.7). In this case the regression coefficients are not correlations between the variables and the axes (see Chapter 2), because the variables in $\mathbf{Y}$ are not standardized—instead, the regression coefficients are the covariances between the variables and the axes. These covariances are equal to the correlations multiplied by the standard deviations of the respective variables. Notice that in the calculation of covariances and standard deviations, sums of squares should be divided by I (=13 in this example) and not by $I - 1$ as in the usual computation of sample variance.

Dual biplots

The form biplot and the covariance biplot defined above are called *dual biplots*: each is the *dual* of the other. Technically, the only difference between them is the way the singular values are allocated, either to the left singular vectors in the form biplot, which thus visualizes the spatial form of the rows (cases), or to the right singular vectors in the covariance biplot, visualizing the covariance structure of the columns (variables). Substantively, there is a big difference between these two options, even though they look so similar. We shall see throughout the following chapters that dual biplots exist for all the multivariate situations treated. An alternative display, especially prevalent in correspondence analysis (Chapter 8), represents both sets of points in principal coordinates, thus displaying row and column structures simultaneously, that is both row- and column-metric-

preserving. The additional benefit is that both sets of points have the same dispersion along the principal axes and avoid the large differences in scale that are sometimes observed between principal and standard coordinates. However, this choice is not a biplot and its benefits go at the expense of losing the scalar-product interpretation and the ability to project one set of points onto the other. This loss is not so great when the singular values on the two axes being displayed are close to each other—in fact, the closer they are to each other, the more the scalar-product interpretation remains valid and intact. But if there is a large difference between the singular values, then the scalar-product approximation of the data becomes degraded (see the Epilogue for further discussion of this point).

Squared singular values are eigenvalues

From (6.6) and (6.8) it can be shown that the squared singular values are eigenvalues. If (6.6) is multiplied on the right by $\mathbf{D}_w\mathbf{F}$ and (6.8) is similarly multiplied on the right by $\mathbf{D}_q\mathbf{G}$, and using the normalizations of $\mathbf{F}$ and $\mathbf{G}$, the following pair of eigenequations is obtained:

$$\mathbf{Y}\mathbf{D}_q\mathbf{Y}^\mathsf{T}\mathbf{D}_w\mathbf{F} = \mathbf{F}\mathbf{D}_\beta^2 \text{ and } \mathbf{Y}^\mathsf{T}\mathbf{D}_w\mathbf{Y}\mathbf{D}_q\mathbf{G} = \mathbf{G}\mathbf{D}_\beta^2, \text{ where } \mathbf{F}^\mathsf{T}\mathbf{D}_w\mathbf{F} = \mathbf{G}^\mathsf{T}\mathbf{D}_q\mathbf{G} = \mathbf{D}_\beta^2 \quad (6.9)$$

The matrices $\mathbf{Y}\mathbf{D}_q\mathbf{Y}^\mathsf{T}\mathbf{D}_w$ and $\mathbf{Y}^\mathsf{T}\mathbf{D}_w\mathbf{Y}\mathbf{D}_q$ are square but non-symmetric. They are easily symmetrized by writing them in the equivalent form:

$$\mathbf{D}_w^{1/2}\mathbf{Y}\mathbf{D}_q\mathbf{Y}^\mathsf{T}\mathbf{D}_w^{1/2}\mathbf{D}_w^{1/2}\mathbf{F} = \mathbf{D}_w^{1/2}\mathbf{F}\mathbf{D}_\beta^2 \text{ and } \mathbf{D}_q^{1/2}\mathbf{Y}^\mathsf{T}\mathbf{D}_w\mathbf{Y}\mathbf{D}_q^{1/2}\mathbf{D}_q^{1/2}\mathbf{G} = \mathbf{D}_q^{1/2}\mathbf{G}\mathbf{D}_\beta^2$$

which in turn can be written as:

$$\mathbf{S}\mathbf{S}^\mathsf{T}\mathbf{U} = \mathbf{D}_\beta^2 \text{ and } \mathbf{S}^\mathsf{T}\mathbf{S}\mathbf{V} = \mathbf{V}\mathbf{D}_\beta^2, \text{ where } \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I} \quad (6.10)$$

These are two symmetric eigenequations with eigenvalues $\lambda_k = \beta_k^2$ for $k = 1, 2, \ldots$

Scree plot

The eigenvalues (i.e., squared singular values) are the primary numerical diagnostics for assessing how well the data matrix is represented in the biplot. They are customarily expressed relative to the total sum of squares of all the singular values—this sum quantifies the total variance in the matrix, that is the sum of squares of the matrix decomposed by the SVD (the matrix $\mathbf{S}$ in (6.2) in this example). The values of the eigenvalues and a bar chart of their percentages of the total are given in Exhibit 6.3—this is called a *scree plot*.

It is clear that the first two values explain the major part of the variance, 57.6% and 34.8% respectively, which means that the biplots in Exhibits 6.1 and 6.2 explain 92.4% of the variance. The pattern in the sequence of eigenvalues in the bar chart is typical of almost all matrix approximations in practice: there are a few

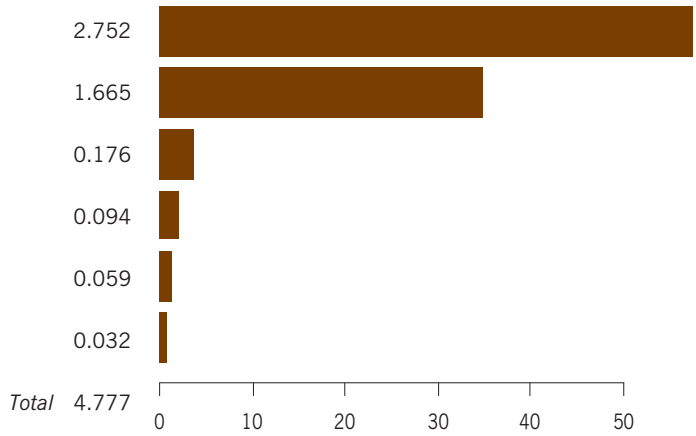


Exhibit 6.3:
Scree plot of the six squared singular values $\lambda_1, \lambda_2, \ldots, \lambda_6$, and a horizontal bar chart of their percentages relative to their total

eigenvalues that dominate and separate themselves from the remainder, with this remaining set showing a slow "dying out" pattern associated with "noise" in the data that has no structure. The point at which the separation occurs (between the second and third values in Exhibit 6.3) is often called the *elbow* of the scree plot. Other rules of thumb for deciding which axes reflect "signal" in the data, as opposed to "noise", is to calculate the average variance per axis, in this case 4.777/6 = 0.796. Axes with eigenvalues greater than the average are generally considered worth plotting.

Just like the eigenvalues quantify how much variance is accounted for by each principal axis, usually expressed as a percentage, so we can decompose the vari-

Contributions to variance

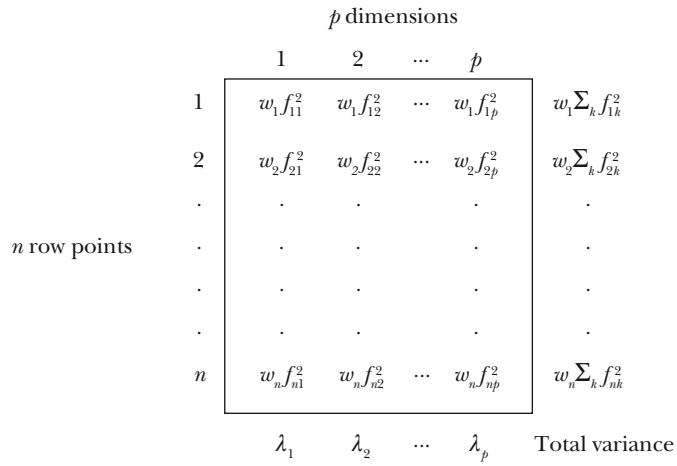


Exhibit 6.4:
Decomposition of total variance by dimensions and points: the row sums are the variances of the row points and the columns sums are the variances of the dimensions

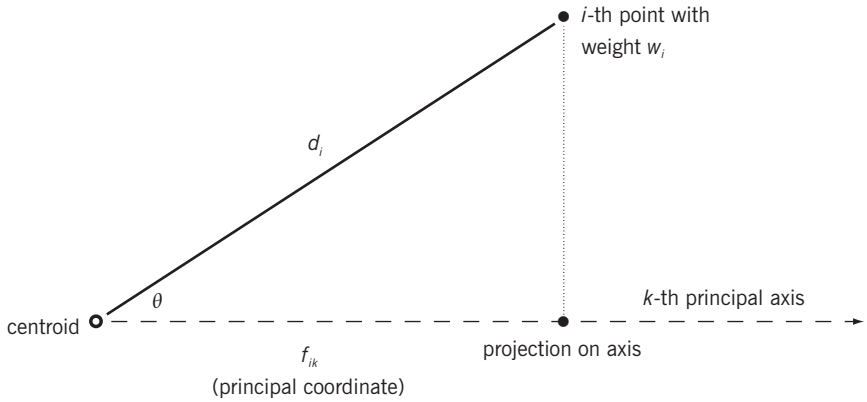


Exhibit 6.5:
Geometry of variance contributions: f_{ik} is the principal coordinate of the i-th point, with weight w_i, on the k-th principal axis. The point is at distance $d_i = \sum_k f_{ik}^2$ from the centroid of the points, which is the origin of the display, and θ is the angle between the point vector (in the full space) and the principal axis. The square cosine of θ is $\cos^2(\theta) = f_{ik}^2 / d_i^2$ (i.e., the proportion of point i's variance accounted for by axis k) and $w_i f_{i1}^2$ is the contribution of the i-th point to the variance on the k-th axis

ance of each individual point, row or column, along principal axes. But we can also decompose the variance on each axis in terms of contributions made by each point. This leads to two sets of contributions to variance, which constitute important numerical diagnostics for the biplot. These contributions are best understood in terms of the principal coordinates of the rows and columns (e.g., **F** and **G** defined above). For example, writing the elements $w_i f_{ik}^2$ in an $n \times p$ matrix (n rows, p dimensions) as shown in Exhibit 6.4. The solution in two dimensions, for example, displays $100(\lambda_1 + \lambda_2)/(\lambda_1 + \lambda_2 + \cdots + \lambda_p)\%$ of the total variance. On the first dimension $100\, w_i f_{i1}^2 / \lambda_1 \%$ of this dimension's variance is accounted for by point i (similarly for the second dimension)—this involves expressing each column in Exhibit 6.4 relative to its sum. Conversely, expressing each row of the table relative to its row sum, $100\, w_i f_{i1}^2 / w_i \sum_k f_{ik}^2 \%$ of point i's variance is accounted for by the first dimension and $100\, w_i f_{i2}^2 / w_i \sum_k f_{ik}^2 \%$ is accounted for by the second dimension. Notice that in this latter case (row elements relative to row sums) the row weights cancel out. The ratio $w_i f_{i1}^2 / w_i \sum_k f_{ik}^2$, for example, equals $f_{i1}^2 / \sum_k f_{ik}^2$, which is equal to the squared cosine of the angle between the i-th row point and the first principal axis, as illustrated in Exhibit 6.5.

The contribution biplot

In the principal component biplots defined in (6.3) and (6.7) the points in standard coordinates are related to their contributions to the principal axes. The following result can be easily shown. Suppose that we rescale the standard coordinates by the square roots of the respective point weights, that is we recover the corresponding singular vectors:

$$\text{from (6.3): } (1/J)^{1/2}\mathbf{\Gamma} = (1/J)^{1/2} J^{1/2}\mathbf{V} = \mathbf{V},$$

$$\text{and from (6.7): } (1/I)^{1/2}\mathbf{\Phi} = (1/I)^{1/2} I^{1/2}\mathbf{U} = \mathbf{U}$$

Then these coordinates display the relative values of the contributions of the variables and cases respectively to the principal axes. In the covariance biplot, for example, the squared value u_{ik}^2 (where u_{ik} is the (i,k)-th element of the singular vector, or rescaled standard coordinate of case i on axis k) is equal to $(w_i f_{ik}^2)/\lambda_k$, where $w_i = 1/I$ here. This variant of the covariance biplot (plotting **G** and **U** together) or the form biplot (plotting **F** and **V** together) is known as the *contribution biplot*[4] and is particularly useful for displaying the variables. For example, plotting **F** and **V** jointly does not change the direction of the biplot axes, but changes the lengths of the displayed vectors so that they have a specific interpretation in terms of the contributions. In Exhibit 6.1, since the weights are all equal, the lengths of the variables along the principal axes are directly related to their contributions to the axes, since they just need to be multiplied by a constant $(1/J)^{1/2} = (1/6)^{1/2} = 0.41$. For PCA this is a trivial variation of the original definition —because the point masses are equal, it just involves an overall rescaling of the standard coordinates. But in other methods where the point masses are different, for example in log-ratio analysis and correspondence analysis, this alternative biplot will prove to be very useful—we will return to this subject in the following chapters.

SUMMARY: Principal Component Analysis Biplots

1. Principal component analysis of a cases-by-variables matrix reduces to a singular value decomposition of the centred (and optionally variable-standardized) data matrix.
2. Two types of biplot are possible, depending on the assignment of the singular values to the left or right singular values of the decomposition. In both the projections of one set of points on the other approximate the centred (and optionally standardized) data.
3. The *form biplot*, where singular values are assigned to the left vectors corresponding to the cases, displays approximate Euclidean distances between the cases.
4. The *covariance biplot*, where singular values are assigned to the right vectors corresponding to the variables, displays approximate standard deviations and correlations of the variables. If the variables had been pre-standardized to have standard deviation equal to 1, a unit circle is often drawn on the covariance biplot because the variable points all have lengths less than or equal to 1 —the closer a variable point is to the unit circle, the better it is being displayed.

4. This variation of the scaling of the biplot is also called the *standard biplot* because the projections of points onto vectors are approximations to the data with variables on standard scales, and in addition because it can be used across a wide spectrum of methods and wide range of inherent variances.

5. The *contribution biplot* is a variant of the form or covariance biplots where the points in standard coordinates are rescaled by the square roots of the weights of the respective points. These rescaled coordinates are exactly the square roots of the part contributions of the respective points to the principal axes, so this biplot gives an immediate idea of which cases or variables are most responsible for the given display.

CHAPTER 7

Log-ratio Biplots

All the techniques described in this book are variations of generalized principal component analysis defined in Chapter 5, and in Chapter 6 we demonstrated the simplest version of principal component analysis. As mentioned at the end of Chapter 6, when variables are on different scales they are usually standardized in some way to equalize the roles of the variables in the analysis: this can be thought of either as a pre-transformation of the data or equivalently as a reweighting of the variables. Many other pre-transformations of the data are possible: for example, to visualize multiplicative differences in the data the logarithmic transformation can be applied, in which case no further standardization is required. Data that are proportions are often transformed by the arcsine function (i.e., the inverse sine function) or by some power function such as the square root. In this chapter we treat the log-ratio transformation which is applicable to a common situation in practice: when data are all measured on the same units and strictly positive. The biplots that result have some special properties and this approach deserves a wider usage, hence a whole chapter is devoted to it.

Contents

Interval and ratio scales

What has not been stated or explained up to now is that PCA assumes the data are on *interval* scales. By this we mean that when we compare two values, we look at their (interval) differences. For example, if we compare a temperature of 3.6 degrees with 3.1 degrees, we say that the difference is 0.5 degrees and this difference is comparable to the difference between 21.9 and 21.4 degrees. On the oth-

er hand, many variables are measured on *ratio* scales where we would express the comparison as a multiplicative, or percentage, difference. For example, a factory worker obtains an increase in salary of 50 euros a month, but his salary before was only 1000 euros a month, so this is in fact a 5% increase; if he were earning 3000 euros before, the increase would be 1.67%. Here it is relevant to compare the ratios of the numbers being compared: 1050/1000 and 3050/3000, not their differences. One has to carefully consider whether the observed variables are on interval or ratio scales, as this affects the way we analyze them. In practice, the values of the variable may be so far away from the zero point of their scale that the distinction between interval and ratio scale is blurred: for example, a 50-unit increase on the much higher value of 100,000 is not much different, percentage-wise, from a 50-unit increase on the higher value of 110,000.

The logarithmic transformation

The classic way to treat ratio-scale variables is to apply the logarithmic transformation, so that multiplicative differences are converted to additive differences: $\log(x/y) = \log(x) - \log(y)$. If the variables are all ratio-scale in nature but in different measurement units, a blanket logarithmic transformation on all of them is an excellent alternative to variable standardization. For example, an economist might analyze the behaviour of several stock market indices such as Dow-Jones, Financial Times, Nikkei, CAC40, etc (the variables) over time (the rows). Each variable has its own inherent scale but differences between them are evaluated multiplicatively (i.e., percentage-wise). The logarithmic transformation will put them all on comparable interval scales, perfect for entering into a PCA, without any standardization necessary. In fact, standardization would be incorrect in this case, since we want to compare the natural variation of the indices on the logarithmic scale, and not equalize them with respect to one another. If we biplotted such data, then the variables would be calibrated non-linearly on a logarithmic scale, reminiscent of the biplots described in Chapter 3.

Log-ratios

Log-ratios are a bit more specialized than logarithms. Not only are values compared within each variable on a ratio scale, but also values within each case are compared across the variables. This means that all the variables must be measured on the same scale. This approach originated in compositional data analysis in fields such as chemistry and geology, where the variables are components of a sample and the data express proportions, or percentages, of the components in each sample (hence the values for each sample sum to 1, or 100%). It is equally applicable to data matrices of strictly positive numbers such as values all in dollars, measurements all in centimetres, or all counts. The R data set `USArrests` has the 50 US states as rows, and the columns are the numbers of arrests per 100,000 residents of three violent crimes: murder, assault and rape. The "ratio" in log-ratio analysis can refer either to ratios within a state or ratios within a crime. The first five rows of this data set are:

```
> USArrests[1:5,c(1,2,4)]
           Murder     Assault       Rape
Alabama      13.2         236       21.2
Alaska       10.0         263       44.5
Arizona       8.1         294       31.0
Arkansas      8.8         190       19.5
California    9.0         276       40.6
```

By ratios within a row (state) we mean the three unique ratios Murder/Assault, Murder/Rape and Assault/Rape, which for Alabama are (to four significant figures) 0.05593, 0.6226 and 11.13 respectively. By ratios within a column (crime) we mean the $50 \times 49/2 = 1225$ pairwise comparisons between states, of which the first four for the column Murder are Alabama/Alaska: 1.320, Alabama/Arizona: 1.630, Alabama/Arkansas: 1.500, Alabama/California: 1.467. The basic idea in log-ratio analysis is to analyze all these ratios on a logarithmic scale, which are interval differences between the logarithms of the data. In general, for a $I \times J$ matrix there are $\frac{1}{2}I(I-1)$ unique ratios between rows and $\frac{1}{2}J(J-1)$ unique ratios between columns. Fortunately, we do not have to calculate all the above ratios—there is a neat matrix result that allows us to work on the original $I \times J$ matrix and effectively obtain all the log-ratios in the resulting map.

Log-ratio analysis

The algorithm for performing *log-ratio analysis* (LRA) relies on a double-centring of the log-transformed matrix and a weighting of the rows and columns proportional to the margins of the data matrix $\mathbf{N}$:

– Let the row and column sums of $\mathbf{N}$, relative to its grand total $n = \sum_i \sum_j n_{ij}$ be denoted by $\mathbf{r}$ and $\mathbf{c}$ respectively:

$$\mathbf{r} = (1/n)\mathbf{N1},\ \mathbf{c} = (1/n)\mathbf{N}^{\mathsf{T}}\mathbf{1} \tag{7.1}$$

– Logarithmic transformation of elements of $\mathbf{N}$: $\mathbf{L} = \log(\mathbf{N})$ (7.2)

– Weighted double-centring of $\mathbf{L}$: $\mathbf{Y} = (\mathbf{I} - \mathbf{1r}^{\mathsf{T}})\mathbf{L}(\mathbf{I} - \mathbf{1c}^{\mathsf{T}})^{\mathsf{T}}$ (7.3)

– Weighted SVD of $\mathbf{Y}$: $\mathbf{S} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{Y}\,\mathbf{D}_c^{\frac{1}{2}} = \mathbf{U}\mathbf{D}_\varphi\mathbf{V}^{\mathsf{T}}$ (7.4)

– Calculation of coordinates:

Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}\mathbf{D}_\varphi$, of columns: $\mathbf{G} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}\mathbf{D}_\varphi$ (7.5)

Standard coordinates of rows: $\mathbf{\Phi} = \mathbf{D}_r^{-\frac{1}{2}}\mathbf{U}$, of columns: $\mathbf{\Gamma} = \mathbf{D}_c^{-\frac{1}{2}}\mathbf{V}$ (7.6)

(The above analysis is the weighted form of LRA, which is usually preferred above the unweighted form, which has equal weights on the rows and columns;

that is, unweighted LRA uses the same steps (7.1) to (7.6) but with $\mathbf{r} = (1/I)\mathbf{1}$, $\mathbf{c} = (1/J)\mathbf{1}$.)

As before, two biplots are possible, but in this case they have completely symmetric interpretations—in our terminology of Chapter 6, they are actually both form biplots. The row-metric preserving biplot of $\mathbf{F}$ and $\boldsymbol{\Gamma}$ plots the rows according to their log-ratios across columns, while the column-metric preserving biplot of $\mathbf{G}$ and $\boldsymbol{\Phi}$ plots the columns according to their log-ratios across rows. The points displayed in standard coordinates represent all the log-ratios by vectors between pairs of points, called *link* vectors. It is the double-centring in (7.3) which gives this special result that analyzing the matrix of I rows and J columns yields the representation of all the pairwise log-ratios. To demonstrate these geometric features with a simple example, we show the row-principal $(\mathbf{F}, \boldsymbol{\Gamma})$ LRA biplot of the USArrests data in Exhibit 7.1.

The double-centring removes one dimension from the data, hence the dimensionality of this 3-column matrix is equal to 2 and Exhibit 7.1 displays 100% of the variance, equal to 0.01790. This can be explained alternatively by the fact that any of the three log-ratios is linearly dependent on the other two, hence the rank of the matrix of log-ratios is 2. In this LRA biplot it is not the positions of the three columns that are of interest but the link vectors joining them, which represent the pairwise log-ratios. For example, the link from Rape to Assault represents the

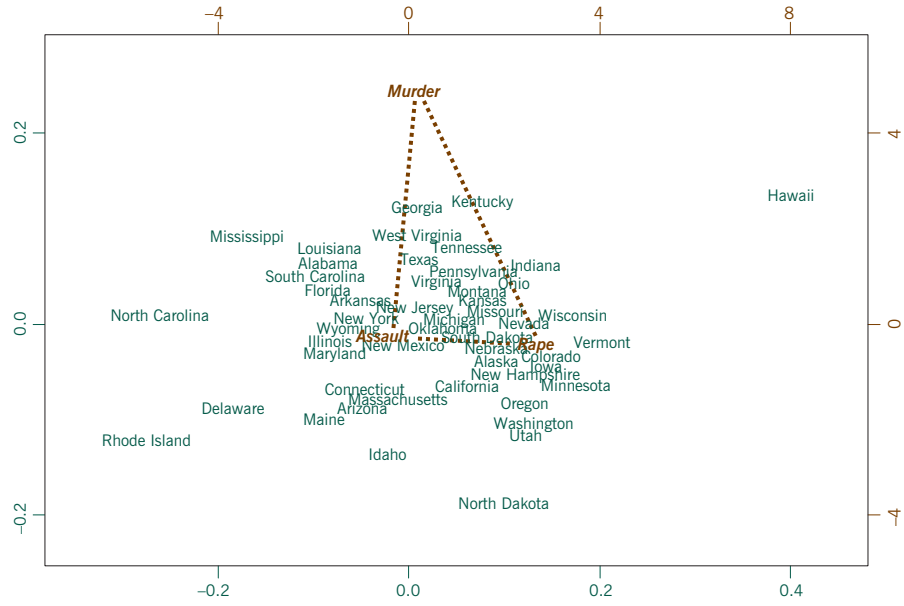


Exhibit 7.1: *Log-ratio biplot of the* `USArrests` *data set from the* R *package, with rows in principal and columns in standard coordinates. The columns are connected by links which represent the pairwise log-ratios. 100% of the log-ratio variance is displayed. Notice the different scales for the two sets of points*

logarithm of Rape/Assault, and the link in the opposite direction represents the negative of that log-ratio, which is the logarithm of Assault/Rape. A biplot axis can be drawn through this link vector and the states projected onto it. The position of Hawaii separates out at the extreme right of this Rape/Assault axis—in fact, Hawaii has an average rate of Rape, but a very low rate of Assault, so the Rape/Assault ratio is very high. Likewise, to interpret Murder/Assault log-ratios, project the states onto this vector which is almost vertical. Hawaii projects high on this axis as well, but it projects more or less at the average of the Murder/Rape link. To know what "average" is, one has to imagine the link translated to the origin of the biplot, which represents the average of all the log-ratios; that is, draw an axis through the origin parallel to the Murder/Rape link—the projection of Hawaii is then close to the origin. Alternatively, project the origin perpendicularly onto the links and this indicates the average point on the corresponding log-ratios.

Log-ratio distance and variance

The positions of the states in Exhibit 7.1 are a representation of *log-ratio distances* between the rows. This distance is a weighted Euclidean distance between the log-ratios within each row, for example the squared distance between rows i and i':

$$d^2_{ii'} = \sum\sum_{j<j'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{ij'}} - \log \frac{n_{i'j}}{n_{i'j'}} \right)^2 \tag{7.7}$$

This can be written equivalently as:

$$d^2_{ii'} = \sum\sum_{j<j'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{i'j}} - \log \frac{n_{ij'}}{n_{i'j'}} \right)^2 \tag{7.8}$$

showing that log-ratios can be considered between the pair of values in corresponding columns. Both (7.7) and (7.8) can be written equivalently in terms of the logarithms of odds-ratios for the four cells defined by row indices i,i' and column indices j,j':

$$d^2_{ii'} = \sum\sum_{j<j'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{i'j}} \frac{n_{i'j'}}{n_{ij'}} \right)^2 \tag{7.9}$$

The log-ratio distances $d^2_{jj'}$ between columns in the alternative column-principal biplot are the symmetric counterparts of (7.7), (7.8) or (7.9), with index i substituting j in the summations, and r_i substituting c_j.

The total variance of the data is measured by the sum of squares of (7.4), which can be evaluated as a weighted sum of squared distances $\sum\sum_{i<i'} r_i r_{i'} d^2_{ii'}$, for example using the definition (7.9) of squared distance in terms of the odds-ratios:

$$\sum\sum_{i<i'}\sum\sum_{j<j'} r_i r_{i'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{i'j}} \frac{n_{ij'}}{n_{ij'}} \right)^2 \tag{7.10}$$

In the above example the total variance is equal to 0.01790.

Data set "morphology"

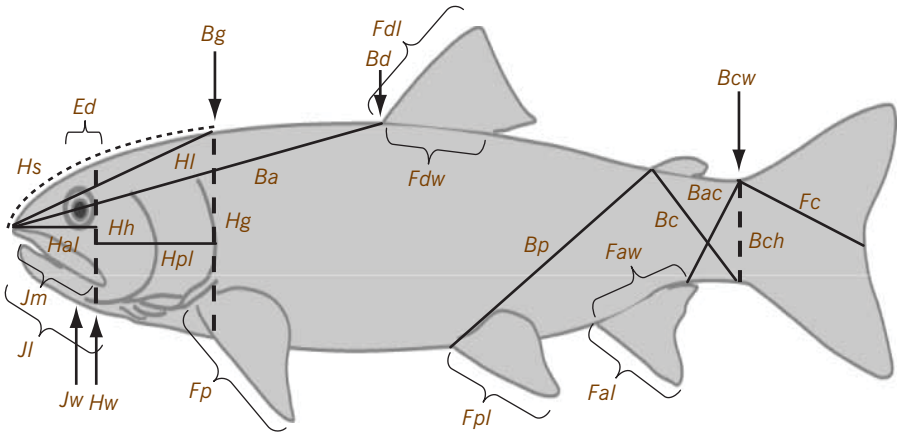
The LRA biplot works well for any strictly positive data that are all measured on the same scale, and for which multiplicative comparisons of data elements, row- or column-wise, make more sense than additive (interval) comparisons. Morphometric data in biology are an excellent candidate for this approach, so we show an application to a data set of 26 measurements on 75 *Arctic charr* fish. The data come from a study of the diet and habitat of the fish and their relationships to their body form and head structure.[5] Exhibit 7.2 shows the abbreviated names of the measurements.

The total variance in these data is 0.001961, much lower than the previous example, indicating that the fish are quite similar to one another in an absolute sense, which is not surprising since they are all of the same species. Nevertheless there are interesting differences between them which may be related to their environment and diets. In Exhibit 7.3 shows the row-principal LRA biplot, where the scale of the low-variance fish points has been enlarged 50 times to show them more legibly. The fish have been labelled according to their sex (f = female, m = male) and habitat where they were caught (L = littoral near shore, P = pelagic in open sea). There does not seem to be any apparent connection with the distribution of the fish and these labels—this can be tested formally using a permutation test, described in the Computational Appendix, while in Chapter 11 the topic of direct comparison of groups of cases is treated (as opposed to comparisons between individual cases, which is what is being analyzed here).

Diagnosing equilibrium relationships

The variable points have no relevance *per se*, rather it is the links between all pairs of variables that approximate the log-ratios—in fact, one could imagine all these links transferred to the origin as vectors representing the pairwise log-ratios. Thus the logarithm of the ratio *Bc/Hpl* (caudal peduncle length/posterior head length) has one of the highest variances—its calibrations, proportional to the inverse of

5. Data provided by Prof. Rune Knudsen and the freshwater biology group of the Department of Arctic and Marine Biology at the University of Tromsø, Norway.



Jl: lower jaw length; *Jw*: lower jaw width; *Jm*: upper jaw length; *Ed*: eye diameter; *Hw*: head width; *Hh*: head height; *Hg*: head height behind gills; *Hpl*: posterior head length; *Hl*: head length; *Hal*: anterior head length; *Hs*: snout length, head curvature from the snout to back of the gills; *Ba*: anterior body length from the snout to the dorsal fin; *Bp*: posterior body length from the anal fin to the adipose fin; *Bch*: caudal height; *Bc*: caudal peduncle length, from the adipose fin to ventral caudal height; *Bac*: caudal peduncle length, anal fin to dorsal caudal height; *Bg*: body width at gills; *Bd*: body width at dorsal fin; *Bcw*: caudal body width; *Fc*: caudal fin length; *Fp*: pectoral fin length; *FdI*: dorsal fin length; *Fdw*: dorsal fin width; *FpI*: pectoral fin length; *Fal*: anal fin length; *Faw*: anal fin width.

Exhibit 7.2:
Morphological characteristics from the left side of Arctic charr fish. Dashed lines indicate heights, arrows indicate widths

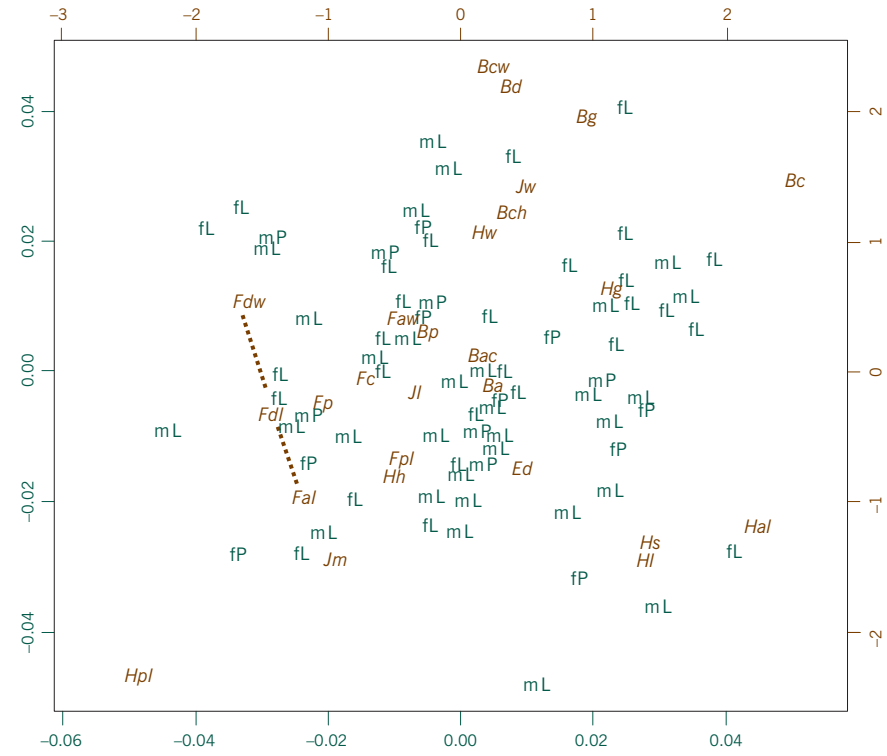


Exhibit 7.3:
Log-ratio biplot of the "morphology" data set, with rows in principal and column in standard coordinates. Labels for the fish are:
fL = *female littoral;*
mL = *male littoral;*
fP = *female pelagic;*
mP = *male pelagic.*
34.5% of the total variance is explained in this map

the vector length would be very close together, and thus the projections of the fish onto this direction would vary greatly in value.

A further interesting property of log-ratio analysis is that a certain class of models can be diagnosed between subsets of variables if they line up in straight lines in the biplot. In Exhibit 7.3 there is a lining up of the three variables Fdw (dorsal fin width), Fdl (dorsal fin length) and Fal (anal fin length), indicated by the dotted line. This means that the log-ratios formed from Fdw, Fdl and Fal could be linearly related—we say "could be" because only 34.5% of the variance is explained by the map and this lining up in the low-dimensional projection of the biplot does not necessarily mean that the points line up in the full space (however, not lining up in the biplot means they definitely do not line up in the full space). To investigate this possible relationship, Exhibit 7.4 shows a scatterplot of two log-ratios, and indeed there is a high positive correlation of 0.700 ($R^2 = 0.490$).

To quantify the relationship we find the best-fitting straight line through the points (this is the first principal axis of the points, not the regression line), and

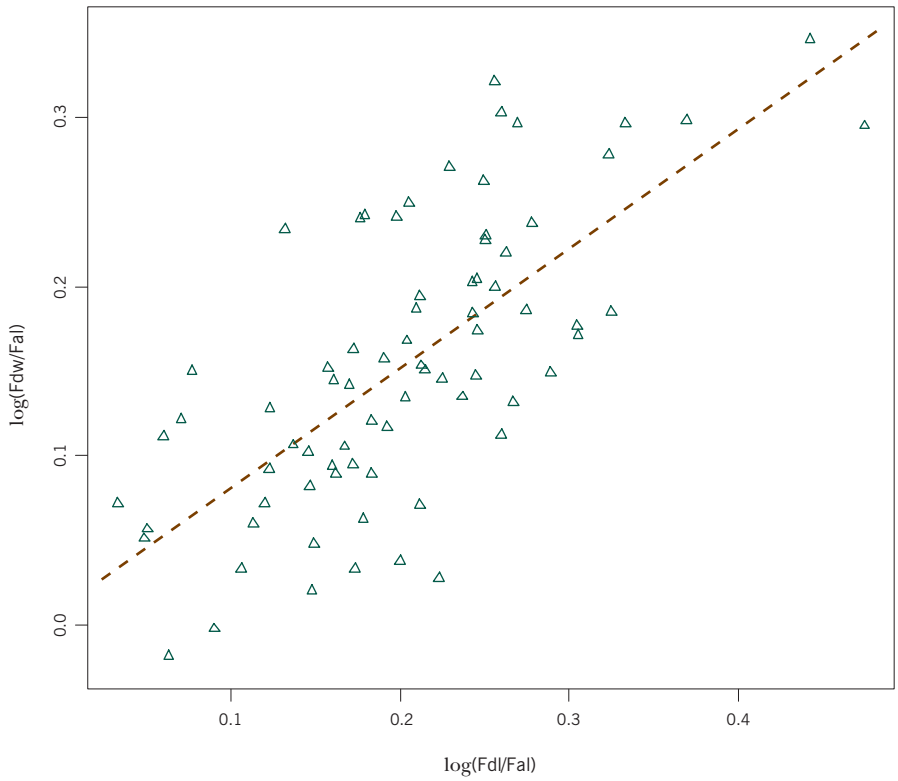


Exhibit 7.4: *Plot of two log-ratios diagnosed from Exhibit 7.3 to be possibly in a linear relationship (the correlation is 0.70). The best-fitting line through the scatterplot has slope equal to 0.707 and intersection 0.0107*

this line turns out to have slope 0.707, and intersection with the vertical axis at 0.0107. So the relationship simplifies to:

$$\log(\mathsf{Fdw}/\mathsf{Fal}) = 0.707 \log(\mathsf{Fdl}/\mathsf{Fal}) + 0.0107$$

Exponentiating:

$$\mathsf{Fdw}/\mathsf{Fal} = 1.0108 \times (\mathsf{Fdl}/\mathsf{Fal})^{0.707}$$

Simplifying:

$$\mathsf{Fdw} = 1.0108 \times \mathsf{Fdl}^{0.707} \times \mathsf{Fal}^{0.293} \tag{7.11}$$

Then, calculating the predicted values of Fdw, the dorsal fin width, as a function of Fdl (dorsal fin length) and Fal (anal fin length), we get a good fit (correlation of 0.750, $R^2 = 0.562$) between the predicted and observed values (Exhibit 7.5).

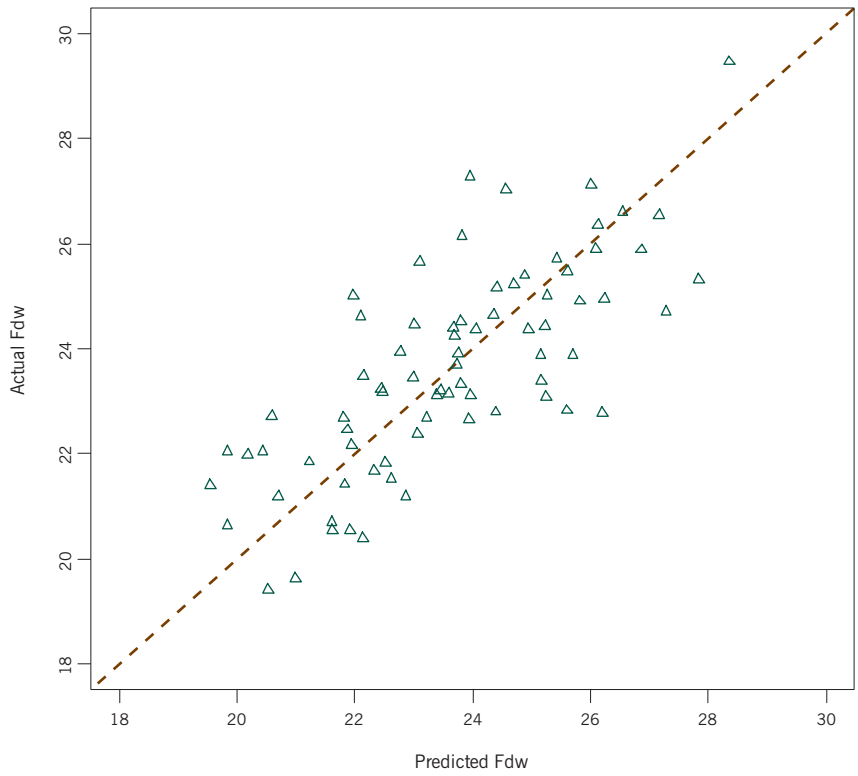


Exhibit 7.5: *Predicted versus actual values of* Fdw *(dorsal fin width) based on the model of (7.11)*

Notice that the model of (7.11) really only has two parameters, the multiplicative constant and a single parameter for the two "predictor" variables, because their power coefficients sum to 1. It is this type of restricted parameter model that is called an "equilibrium model" in certain areas of research such as genetics, chemistry and geology.

SUMMARY: Log-ratio Biplots

1. Log-ratio analysis applies to any table of strictly positive data, where all data entries are measured on the same scale.
2. The original $I \times J$ table is logarithmically transformed and then double-centred, where the rows and columns are weighted proportionally to their marginal sums, followed by a SVD decomposition. The form biplot, where singular values are assigned to the left vectors corresponding to the cases, displays approximate Euclidean distances between the cases based on all the pairwise log-ratios of the variables.
3. Log-ratio biplots represent the pairwise log-ratios between all the columns, or between all the rows, as the case may be. These are the vectors that connect the pairs of columns or pairs of rows.
4. If a subset of columns, for example, line up in straight lines, this diagnoses possible equilibrium relationships in that subset, in the form of a multiplicative power model relating the columns.

CHAPTER 8

Correspondence Analysis Biplots

Correspondence analysis is the most versatile of the methods based on the SVD (singular value decomposition) for visualizing data. It applies primarily to a cross-tabulation (also called a contingency table) of two categorical variables but can be extended to frequency tables, ratio-scale data in general, binary data, preferences and fuzzy-coded continuous data. Like log-ratio analysis, correspondence analysis treats the rows and columns of a table in a symmetric fashion. There are several equivalent asymmetric ways of thinking about the analysis, however, and there are different associated biplots depending on whether the rows or columns are regarded as the "variables" of the table. In this chapter we define and illustrate the basic algorithm of correspondence analysis and list its properties and relationships to principal component analysis, log-ratio analysis, multidimensional scaling and regression biplots. In subsequent chapters various extensions of correspondence analysis will be described: the multiple form in Chapters 9 and 10, and the constrained form in Chapter 12.

Contents

Profiles, masses and chi-square distances: data set "smoking"

All the methods in this book are based on what the French call a *triple* (triplet) of information for a data set: the definition of (1) objects in a multidimensional space, (2) their weights, and (3) the distances between them. In MDS (multidimensional scaling) the distances form the original data set and an approximate map of the objects is produced. The objects could, however, have different

weights and the analysis would then represent distances involving points with higher weight better than those with lower weight. In PCA (principal component analysis), the original data is in the form of a rectangular data matrix and each row (or column) defines a point in multidimensional space. The points could be assigned different weights here as well, and if the distance function between the points is Euclidean, then the SVD provides the solution for the low-dimensional visualization of the points. In correspondence analysis (CA), these three concepts of points, weights and distances are called profiles, masses and chi-square distances, respectively. We review their definitions using the classic "smoking" data set (available in the **ca** package in R), given in Exhibit 8.1. The first table is the cross-tabulation of all 193 staff members of an organization according to their category in the organization and their level of smoking.

The row *profiles*, given in the second table, are the frequencies in the rows divided by their row sums (e.g., 0.364 = 4/11). The last row contains the average row profile, which is the profile of the column sums of the original table (e.g., 0.316 = 61/193). Similarly, the column profiles in the third table are the frequencies in the columns divided by the column sums and the average column profile in the last column is the profile of the row sums in the original table (e.g., 0.057 = 11/193). The row profiles are the points visualized in the row problem, and the column profiles are those visualized in the column problem.

Each profile has a weight called a *mass*, equal to the marginal sum of that row or column as the case may be, divided by the grand total of the table. For example, the first row profile has mass 11/193 = 0.057, which is identical to the first element of the average column profile. Thus the average column profile contains the row masses, and the average row profile contains the column masses. The masses are used to weight the profiles in the analysis, so that profiles based on larger counts have a stronger role in the analysis.

Distances between profiles are calculated using the chi-square distance, which has already been introduced in Chapter 4. The average row profile, for example, apart from serving to centre the row profiles, defines the distance function between row profiles, using the inverses of its values. For example, the distance between the first two row profiles is:

$$\sqrt{\frac{(0.364-0.222)^2}{0.316}+\frac{(0.182-0.167)^2}{0.233}+\frac{(0.273-0.389)^2}{0.321}+\frac{(0.182-0.222)^2}{0.130}}=0.345$$

This is a natural default standardization for frequency data, which tend to have higher variances if their means are higher. Similarly, chi-square distances can be

Exhibit 8.1:
Data set "smoking" and its row and column profiles, as well as their respective average profiles

Original cross-tabulation:

Staff group		Smoking class				
		None	*Light*	*Medium*	*Heavy*	*Sum*
Senior managers	SM	4	2	3	2	*11*
Junior managers	JM	4	3	7	4	*18*
Senior employees	SE	25	10	12	4	*51*
Junior employees	JE	18	24	33	13	*88*
Secretaries	SC	10	6	7	2	*25*
	Sum	*61*	*45*	*62*	*25*	*193*

Row profiles:

	Smoking class			
	None	*Light*	*Medium*	*Heavy*
SM	0.364	0.182	0.273	0.182
JM	0.222	0.167	0.389	0.222
SE	0.490	0.196	0.235	0.078
JE	0.205	0.273	0.375	0.148
SC	0.400	0.240	0.280	0.080
Average	*0.316*	*0.233*	*0.321*	*0.130*

Column profiles:

	Smoking class				
	None	*Light*	*Medium*	*Heavy*	*Average*
SM	0.066	0.044	0.048	0.080	*0.057*
JM	0.066	0.067	0.113	0.160	*0.093*
SE	0.410	0.222	0.194	0.160	*0.264*
JE	0.295	0.533	0.532	0.520	*0.456*
SC	0.164	0.133	0.113	0.080	*0.130*

defined between the column profiles, using the inverses of the elements of the average column profile.

Correspondence analysis (CA)

CA has many equivalent definitions and we give just one of them here. It is—at the same time—a generalized PCA of the row profiles and a generalized PCA of the column profiles, and the treatment of the rows and columns is the same, just as in LRA (log-ratio analysis) of the previous chapter. And again, both the row and column problems rely on the same matrix decomposition, as follows:

– First divide the original data table $\mathbf{N}$ by its grand total $n = \sum_i \sum_j n_{ij}$: $\mathbf{P} = (1/n)\mathbf{N}$

  Denote by $\mathbf{r}$ and $\mathbf{c}$ the marginal sums of $\mathbf{P}$: $\mathbf{r} = \mathbf{P1}$, $\mathbf{c} = \mathbf{P}^\mathsf{T}\mathbf{1}$ (8.1)
  (these are identical to $\mathbf{r}$ and $\mathbf{c}$ defined in (7.1)).

– Calculate the matrix of standardized residuals $\dfrac{p_{ij} - r_i c_j}{\sqrt{r_i c_j}}$ and its SVD:

$$\mathbf{S} = \mathbf{D}_r^{-½}(\mathbf{P} - \mathbf{r}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-½} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T} \tag{8.2}$$

– Calculate the coordinates:

  Principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-½}\mathbf{U}\mathbf{D}_\alpha$, of columns: $\mathbf{G} = \mathbf{D}_c^{-½}\mathbf{V}\mathbf{D}_\alpha$ (8.3)

  Standard coordinates of rows: $\mathbf{\Phi} = \mathbf{D}_r^{-½}\mathbf{U}$, of columns: $\mathbf{\Gamma} = \mathbf{D}_c^{-½}\mathbf{V}$ (8.4)

Notice how similar this algorithm is to that of log-ratio analysis, formulated in (7.1)–(7.6) of Chapter 7; in fact, the two algorithms can be seen to be even more similar if the $\mathbf{S}$ matrix in (8.2) is rewritten in the equivalent form:

$$\text{[matrix for SVD in CA]} \qquad \mathbf{S} = \mathbf{D}_r^{½}(\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\,(\mathbf{D}_r^{-1}\mathbf{P}\,\mathbf{D}_c^{-1})\,(\mathbf{I} - \mathbf{1}\mathbf{c}^\mathsf{T})^\mathsf{T}\mathbf{D}_c^{½} \tag{8.5}$$

whereas in log-ratio analysis, from (7.2), (7.3) and (7.4):

$$\text{[matrix for SVD in LRA]} \qquad \mathbf{S} = \mathbf{D}_r^{½}(\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\log(\mathbf{N})\,(\mathbf{I} - \mathbf{1}\mathbf{c}^\mathsf{T})^\mathsf{T}\mathbf{D}_c^{½} \tag{8.6}$$

So the difference is that CA analyzes the contingency ratios $p_{ij}/(r_i c_j)$ in $\mathbf{D}_r^{-1}\mathbf{P}\,\mathbf{D}_c^{-1}$, whereas LRA analyses the logarithms of the data $\log(\mathbf{N})$. Since the double-centring removes any additive row or column constant, $\log(\mathbf{N})$ in (8.6) can be replaced by $\log(\mathbf{D}_r^{-1}\mathbf{P}\,\mathbf{D}_c^{-1})$ without changing the matrix for the SVD. So the only real difference between LRA and CA is the logarithmic transformation!

Asymmetric maps

As in LRA, there are two biplots that result in CA: the row-principal biplot ($\mathbf{F}, \mathbf{\Gamma}$) and the column-principal biplot ($\mathbf{G}, \mathbf{\Phi}$). In CA, however, the points in standard coordinates have an additional geometric interpretation: they are the extreme *unit profiles* or *vertices* of the profile space. Consider the row profiles of the "smoking" data, for example, and their associated biplot coordinates: $\mathbf{F}$ for the rows and $\mathbf{\Gamma}$ for the columns. In a two-dimensional display (using the first two columns of $\mathbf{F}$ and $\mathbf{\Gamma}$) the five row points are projections of the row profiles onto the best-fitting plane. The four column points, in standard coordinates, are the projections onto the same plane of the unit profiles [1 0 0 0], [0 1 0 0], [0 0 1 0] and [0 0 0 1]. Since any row profile $[p_1\ p_2\ p_3\ p_4]$ with elements adding up to 1 can be expressed as p_1 [1 0 0 0] + p_2 [0 1 0 0] + p_3 [0 0 1 0] + p_4 [0 0 0 1], it follows that the row profiles are at weighted averages of the column points, the weights being the profile elements. It is this weighted average (or centroid) property that makes CA so popular in ecological applications—if the columns follow an ecological gradient (for example, rainfall in a botanical study) then the weighted averages of the columns points for each row profile would situate the row on that gradient. Because the row and column points in this biplot lie in the same space, with the column points defining the most extreme profiles possible, the resultant display is also called a *map*, specifically an *asymmetric map*.

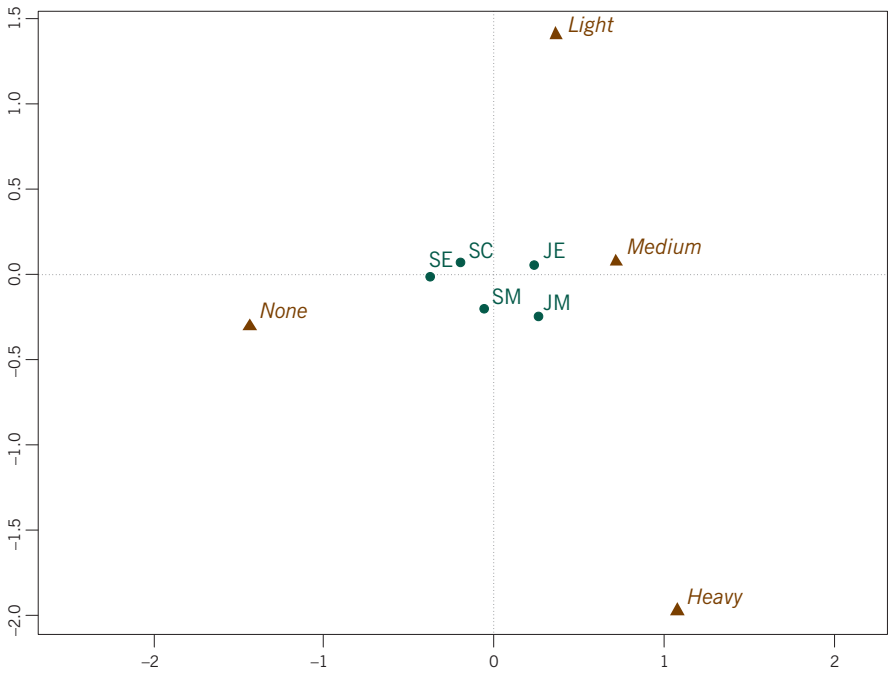


Exhibit 8.2:
Row asymmetric CA map (i.e., row principal biplot) of the "smoking" data, with rows in principal coordinates and columns in standard coordinates. This map is reproduced directly from the **ca** *package in* R *—see the Computational Appendix*

Exhibit 8.2 shows the row asymmetric map of the "smoking" data set. Because the two sets of points co-exist in the same space, the amount of variation between the row profiles can be seen in relation to the extreme vertex profiles. The five row profiles actually lie inside a tetrahedron in three-dimensional space, which has vertices defined by the four column points. As explained above, each row profile is at the weighted average of the four vertex points in this three-dimensional space, and so also in the projected map of Exhibit 8.2. Thus secretaries (SE) lie to the left because they must have higher than average proportion of non-smokers, whereas junior employees and managers (JE and JM) lie to the right because they have higher than average levels of smokers, with junior managers tending towards the high smoking group. All these deductions from the map can be confirmed in the data of Exhibit 8.1. In fact, we can be absolutely sure of these conclusions because 99.5% of the variance in Exhibit 8.1 is displayed in the map.

Connection with PCA and MDS

In (8.1)–(8.5) the algorithm is presented as a decomposition of a matrix where rows and columns have been treated symmetrically, but the same decomposition can also be thought of asymmetrically as an analysis of rows or an analysis of columns, as we in fact introduced it originally in the context of Exhibit 8.1. The matrix formulation $\mathbf{S}$ in (8.2) (equivalently in (8.5)) can be written either as:

$$\mathbf{S} = \mathbf{D}_r^{-1/2}(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-1/2} \tag{8.7}$$

or, in transposed form:

$$\mathbf{S}^\mathsf{T} = \mathbf{D}_c^{-1/2}(\mathbf{D}_c^{-1}\mathbf{P}^\mathsf{T} - \mathbf{1}\mathbf{r}^\mathsf{T})\mathbf{D}_r^{-1/2} \tag{8.8}$$

These two formulations show that CA can be thought of either (in (8.7)) as a weighted PCA (see Chapter 5) of the row profiles in the rows of $\mathbf{D}_r^{-1}\mathbf{P}$, weighted by the row masses in $\mathbf{r}$, centred at their average profile $\mathbf{c}^\mathsf{T}$, in the chi-square metric defined by $\mathbf{D}_c^{-1}$; or (in (8.8)) as a weighted PCA of the column profiles in the rows of $\mathbf{D}_c^{-1}\mathbf{P}^\mathsf{T}$, weighted by the column masses in $\mathbf{c}$, centred at their average profile $\mathbf{r}^\mathsf{T}$, in the chi-square metric defined by $\mathbf{D}_r^{-1}$. In the former row problem, the asymmetric map represents the row profiles in principal coordinates, with the unit profiles representing the columns in standard coordinates; in the latter asymmetric map, the columns profiles are in principal coordinates with the unit profiles representing the rows in standard coordinates.

Exactly the same principal coordinates can be obtained if CA is formulated as a pair of MDS problems. For example, chi-square distances are calculated between row profiles using the metric $\mathbf{D}_c^{-1}$ and with row points weighted by the row masses in $\mathbf{r}$. Then by applying the classical MDS algorithm, with weights, in (5.11) and

(5.12), the principal row coordinates are recovered exactly. A symmetric result holds for the column profiles.

Connection with regression biplots

There are several ways to explain CA as a regression biplot, as described in Chapter 2—we explain it here in terms of definition (8.7), the weighted PCA of the row profiles, weighted by the row masses in the chi-square metric based on $\mathbf{D}_c^{-1}$ (in other words, the asymmetric map of the row profiles). The j-th column of the row profile matrix has elements p_{1j}/r_1, p_{2j}/r_2, ⋯, p_{Ij}/r_I. Centring is with respect to the average profile element c_j and the chi-square standardization implies dividing the centred profile by the square root of the corresponding average profile element, $c_j^{1/2}$. An appropriate regression is then when these standardized values $((p_{ij}/r_i) - c_j)/c_j^{1/2}$ $(i = 1, \ldots, I)$ constitute the response variable and the standard row coordinates on the first two dimensions, say, form the explanatory variables, then applying weighted least-squares fitting with weights equal to the row masses. The solution gives a constant equal to 0 and coefficients equal to $c_j^{1/2}$ times the principal coordinates of the j-th column (this result is illustrated in the Computational Appendix). This result implies that, if the regression is performed on the principal row coordinates rather than the standard ones, then the coefficients in the solution will be exactly the coordinates of the columns in the contribution biplot (see later in this chapter).

Inertia and inertia decomposition

The total variance in CA has a close connection with the chi-square statistic χ^2 often calculated on cross-tabulations as a measure of statistical association between rows and columns. The total variance of the **S** matrix decomposed in (8.2) (equivalently, (8.5), (8.7) or (8.8)) is:

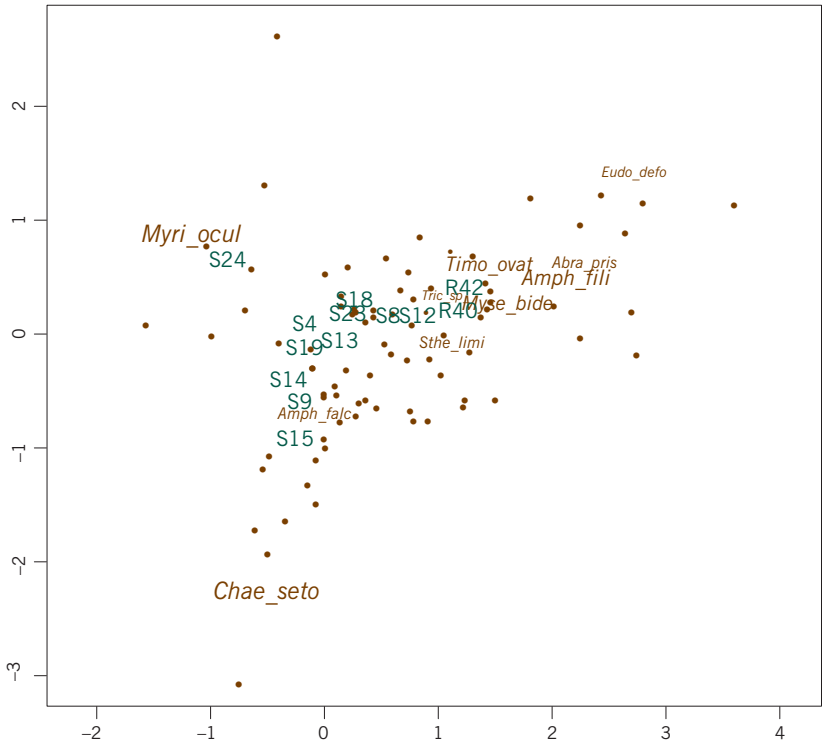
$$\sum_i \sum_j s_{ij}^2 = \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j} \tag{8.9}$$

In CA terminology this quantity is called the *total inertia* of the data matrix, or simply the *inertia*. It is easily shown that multiplying the inertia by the grand total n of the matrix gives the chi-square statistic: $\chi^2 = n \times$ inertia. As explained in general in Exhibit 6.4, there is a decomposition of total variance across points and across principal axes, leading to two ways of defining contributions for the rows as well as for the columns. First, contributions of each profile point to the inertia of each axis (column proportions in Exhibit 6.4) are used to interpret each axis—in the **ca** package in R, these are denoted by the acronym CTR and expressed as permills (see Computational Appendix). Second, contributions of the axes to the inertia of each point (row proportions in Exhibit 6.4) are squared angle cosines between the axes and the points, interpreted as squared correlations or as proportions of inertia explained at the point level rather than for all points together—these are denoted by the acronym COR in the **ca** package and also multiplied by 1000.

Data set "benthos"

This data set consists of 13 columns, the sites at which samples have been taken on the sea bed in the North Sea near an oilfield to study the effect of oil exploration on marine life. In each sample the benthic ("sea bed") species have been identified and counted, leading to an ecological abundance table where the large number of variables (the species) form the rows and the smaller number of samples the columns: in this case, 92 rows (species) by 13 columns (sites). Two of the sites, labelled R40 and R42, are reference stations far from the oilfield and regarded as an unpolluted environment. CA is regularly applied to such abundance tables to visualize the sites in relation to their species composition (the "column problem", the way the matrix is organized here) or to visualize the species' distribution across sites (the "row problem"). Exhibit 8.3 shows the column principal asymmetric map, where sites in principal coordinates are at weighted averages of species points in standard coordinates. The abbreviated species names are shown only if they contribute more than 1% to the two-dimensional map—referring to Exhibit 6.4, this percentage is calculated as $100(w_i f_{i1}^2 + w_i f_{i2}^2)/(\lambda_1 + \lambda_2)$. Of the 92 species, only 10 contribute more than 1% each, totalling 85% of the two-dimensional solution, while the remaining 82 collectively contribute only 15%.

Exhibit 8.3:
Column principal CA biplot of the "benthos" data, with columns (sites) in principal coordinates and rows (species) in standard coordinates. The 10 species with abbreviated labels each make a contribution of more than 1% to the solution, the others are indicated by dots. Total inertia is 0.783, with 57.5% explained in the biplot



Contribution CA biplot

The points that contribute highly to a CA map such as Exhibit 8.3 are generally the high-frequency points, while the low-frequency points contribute very little. The low frequency points, however, often have unusual profiles and lie on the periphery of the map, giving an impression of high importance—for example, in the "benthos" application a very rare species, occurring in just two or three sites, will have a profile at the outer reaches of the profile space. If we are interested only in the direction vectors for the species in the biplot, then this is an excellent situation to use the contribution biplot (see the end of Chapter 6). Rather than use the standard coordinates to represent the species, as in Exhibit 8.3, these coordinates are multiplied by the square roots of the corresponding species masses, causing rare species to be pulled towards the centre, as shown in Exhibit 8.4. The species vectors now show which ones are important to interpret, because their lengths now reflect the contributions of the species to the solution. Exhibit 8.4 shows which are the important species that separate out the unpolluted sites R40 and R42 to the right, while species *Chaetesona setosa* is generally found at polluted sites, particularly S15 which is close to the oilfield. There is a very high abundance of *Myriochele oculata* at site S24 which is not related to the pollution

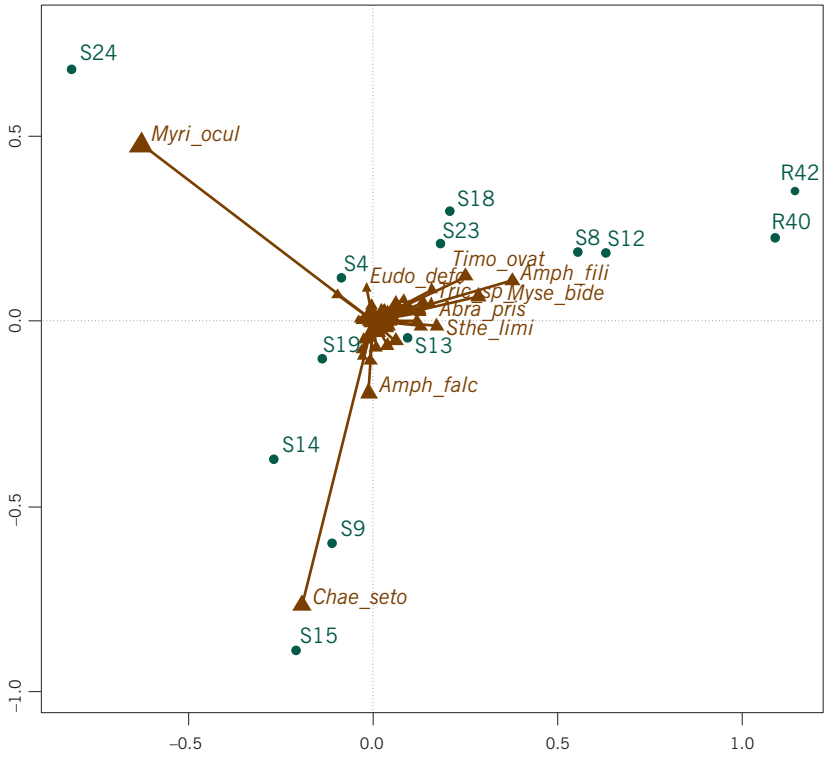


Exhibit 8.4:
Contribution CA biplot of the "benthos" data, with sites in principal coordinates and species in standard coordinates multiplied by the square roots of their masses. The position of each species on each axis is now directly related to its contribution to that axis. The 10 highly contributing species of Exhibit 8.3 (labelled) now stand out in the biplot and all the others collapse to the centre. In this graphic the size of the triangle at each species point, rather than the label, is related to the species total abundance level

gradient—this gradient emerges as a curve from the bottom (sites S15, S9 and S14) to the right (the reference stations).

As described earlier in this chapter, the coordinates of the species in this contribution biplot turn out to be the regression coefficients when the species are regressed on the two CA dimensions, where (i) the species "response" is defined by its elements in the matrix of site profiles, centred and normalized in a chi-square fashion, (ii) the "predictors" are the principal coordinates of the sites, and (iii) the regression is fitted by weighted least squares, using the site masses as weights. This is an additional interpretation of the contribution biplot in the case of CA, giving even more meaning to the positions of the species points in Exhibit 8.4.

SUMMARY: Correspondence Analysis Biplots

1. Correspondence analysis is applicable to a table of nonnegative data, the primary example being a cross-tabulation of two categorical variables, that is a contingency table.
2. The method can be thought of as an analysis of row or column *profiles* of the data matrix—these are the rows or columns expressed relative to their marginal totals.
3. Each profile receives a weight equal to the relative marginal total, called a *mass*.
4. Distances between profiles are defined by the *chi-square metric*. This is essentially a type of standardization of the profile values similar to that used in PCA, but using the average profile element as an estimate of variance rather than the variance itself.
5. The total variance, called *inertia*, in the data is numerically equal to the chi-square statistic for the table divided by the table's grand total.
6. Two types of asymmetric maps, both of which are biplots, are possible, depending on whether row or column profiles (and thus their interpoint chi-square distances) are visualized. Both are form biplots.
7. The contribution biplot can be particularly useful in CA applications, especially when there are quite different levels in rows or in columns (i.e., large differences in the masses). This biplot pulls in the points represented in standard coordinates by the square roots of their respective masses. For each such point, the squares of its rescaled coordinates are equal to the part contributions that the point makes to the respective principal axes.
8. In the contribution biplot, suppose that rows are in principal coordinates (i.e., row profiles are being visualized) and columns in "shrunken" standard coordinates. Then these latter coordinates for each column are also regression coefficients when the standardized values for that column in the row profile matrix are regressed on the principal coordinates of the rows, using weighted least squares with weights equal to the row masses.

CHAPTER 9

Multiple Correspondence Analysis Biplots I

Multiple correspondence analysis is the extension of simple correspondence analysis of a cross-tabulation of two categorical variables to the case of several variables. The method is used mostly in the visualization of social survey data, where respondents reply to a series of questions on discrete scales such as "yes/no" or "agree/unsure/disagree". A type of data that is intermediate between simple and multiple correspondence analysis is a concatenated, or stacked, table—this is a block matrix composed of several two-way cross-tabulations of the same sample of respondents, where each cross-tabulation is between a demographic and a substantive variable. In this chapter we show how CA biplots of a single table can be extended to concatenated tables and then, in the following chapter, to multiple correspondence analysis where several variables are cross-tabulated with one another. The way total variance is measured and how it is decomposed into parts is a recurrent theme in this area, and it will be shown how the biplot concept can clarify this issue.

Contents

Data set "women"

The *International Social Survey Program* (ISSP) is an annual co-operative program between many countries where social surveys are conducted to ask people in each country the same questions on a different theme. The data we shall consider in this chapter are from the third Family and Changing Gender Roles survey conducted in 2002. Even though data are available from more than 30 countries, we shall just treat the Spanish data here (see the second case study in Chapter 14 for

a more detailed analysis). Also, because we want to avoid the issue of missing values for the moment, we have deleted 364 cases with missing data, leaving 2107 of the original 2471 respondents. The questions we focus on are those related to working women and the effect on the family, specifically the following eight statements to which the respondents could either (1) strongly agree, (2) agree, (3) neither agree nor disagree, (4) disgree, or (5) strongly disagree:

A: a working mother can establish a warm relationship with her child
B: a pre-school child suffers if his or her mother works
C: when a woman works the family life suffers
D: what women really want is a home and kids
E: running a household is just as satisfying as a paid job
F: work is best for a woman's independence
G: a man's job is to work; a woman's job is the household
H: working women should get paid maternity leave

There were also several demographic variables, of which we retained the following:

g: gender (1 = male, 2 = female)
m: marital status (1 = married/living as married, 2 = widowed, 3 = divorced, 4 = separated, but married, 5 = single, never married)
e: education (0 = no formal education, 1 = lowest education, 2 = above lowest education, 3 = higher secondary completed, 4 = above higher secondary level, below full university, 5 = university degree completed)
a: age (1 = 16-25 years, 2 = 26-35, 3 = 36-45, 4 = 46-55, 5 = 56-65, 6 = 66 and older)

Abbreviations in the analyses that follow are constructed in the obvious way: for example, *C2* is an agreement to statement *C*, and e5 is category 5 of education. The only exception is for the variable *H* for which there were only two respondents who strongly disagreed—these were combined with the disagree category, leading to a new category denoted as *H4,5*. To demonstrate what is called *interactive coding* of two variables, a variable with 12 categories was constructed from the gender and age variables, with categories denoted by ma1 to ma6 (six age groups for males) and fa1 to fa6 (six age groups for females).

Concatenated table

In simple CA a single demographic variable would be cross-tabulated with a single substantive question, for example the cross-tabulation of education (6 categories) with question *A* (5 categories). The pairwise cross-tabulations of each of the demographic variables with each of the substantive questions can be assembled in a block matrix called a *concatenated table*. Exhibit 9.1 shows just a part of this 23 × 39 table (one less column because of the combining of *H4* and *H5*), with

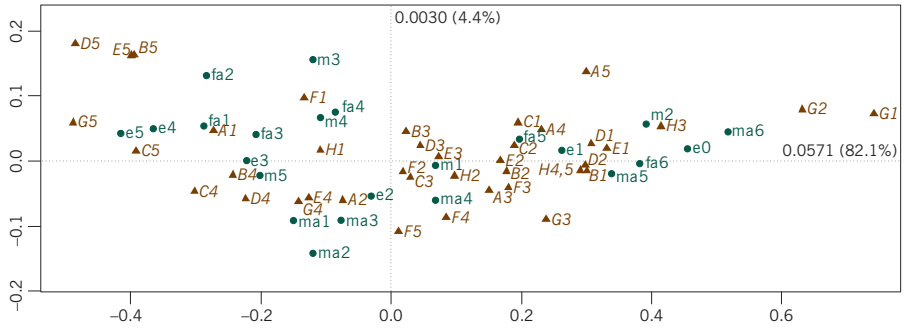
	A1	*A2*	*A3*	*A4*	*A5*	*B1*	*B2*	*B3*	*B4*	*B5*		*n*
m1	192	486	54	381	56	80	550	138	334	67	⋯	1169
m2	21	68	9	63	11	13	101	16	37	5	⋯	172
m3	10	17	2	12	4	5	16	10	12	2	⋯	45
m4	12	32	5	16	4	4	30	7	24	4	⋯	69
m5	162	329	21	126	14	22	259	66	258	47	⋯	652
e0	23	97	16	90	14	20	138	26	52	4	⋯	240
e1	76	203	21	192	30	39	286	62	114	21	⋯	522
e2	99	263	28	168	22	37	264	69	182	28	⋯	580
e3	100	203	16	95	17	17	178	43	160	33	⋯	431
e4	48	81	2	32	3	5	52	13	78	18	⋯	166
e5	51	85	8	21	3	6	38	24	79	21	⋯	168
ma1	38	80	3	27	3	6	70	13	55	7	⋯	151
ma2	41	116	9	53	7	10	92	23	92	9	⋯	226
ma3	30	94	13	37	7	14	77	18	60	12	⋯	181
ma4	25	65	9	40	4	13	68	12	43	7	⋯	143
ma5	16	54	0	50	5	15	72	17	18	3	⋯	125
ma6	15	43	9	64	15	18	95	13	18	2	⋯	146
fa1	48	83	7	39	6	5	56	21	84	17	⋯	183
fa2	59	92	5	58	13	10	82	23	86	26	⋯	227
fa3	46	81	9	51	6	7	78	21	68	19	⋯	193
fa4	37	75	7	60	3	7	76	30	55	14	⋯	182
fa5	21	52	5	42	6	6	65	15	34	6	⋯	126
fa6	21	97	15	77	14	13	125	31	52	3	⋯	224

Exhibit 9.1: *Part of the 23 × 39 concatenated table for the "women" data set, showing the first 10 columns corresponding to the response categories of questions A and B. The 40 column categories are reduced to 39 because H4 and H5 are combined. The sample size for each demographic category is given in the last column. There are 3 × 8 = 24 cross-tabulations in this concatenated table*

the rows being the categories of marital status (5 categories), education (6 categories) and the gender/age combinations (12 categories). Notice that the gender and age groups themselves are not included along with their combinations, although they can be added as so-called *supplementary points* in the analysis, a subject to be discussed later in this chapter.

This type of table is also called a *block matrix*, composed of 3 blocks row-wise and 8 blocks column-wise, that is 24 subtables in total which form the blocks, or subtables, of the matrix. Each of the 24 subtables has the same grand total, which is the number of respondents, equal to 2107. Each of the 8 subtables in a row block has the same row margins and each of the 3 subtables in a column block has the same column margins: for example, the row sums of the table cross-tabulating marital status (m1 to m5) with question *A* (*A1* to *A5*) are the same as the row sums of the table cross-tabulating marital status with question *B*—these row sums are the sample sizes given in column "*n*" of Exhibit 9.1. As a consequence of this equality of marginal sums, it is easy to show the useful result that the total inertia of the concatenated table is the average of the inertias of its 24 subtables.

Exhibit 9.2:
Symmetric CA map of the concatenated table of Exhibit 9.1. This is not a biplot since both the row and column points are displayed in principal coordinates



Symmetric CA map

The most common way of showing the results of a CA is in the form of the symmetric map, shown in Exhibit 9.2. In this map both rows and columns are displayed in principal coordinates, with the result that the graphic is strictly speaking not a biplot. However, interpoint chi-square distances are approximately displayed both between rows and between columns. Since in this case the result is very one-dimensional, with 82.1% of the inertia on the first dimension, we initially interpret only the left-to-right dispersion of the points. Clearly, categories on the left hand side correspond to attitudes favourable to women working while those on the right hand side correspond to the traditional view that they should not work but look after the household and children. Correspondingly there is a lining up of the demographic categories from left to right: for example, highest education is on left and lowest on the right, and the age groups similarly vary from youngest on the left to oldest on the right.

Asymmetric map/biplot for concatenated table

The total inertia in this example is equal to 0.06957, which is a very low value in absolute terms: on average the associations between the demographic variables and the question responses are low, which is quite typical for social science data. Geometrically, this means that the row profiles, for example, are scattered close to the average row profile, with the column vertex points at the outer extremities of the profile space very far out from the set of row profile points. This is clear in the asymmetric map/biplot of Exhibit 9.3, where the column points (in standard coordinates) are so far away from the demographic row points (in principal coordinates) that the latter are too close to one another to label.

Notice that for a concatenated table a vertex point consisting of zeros with a single 1 does not have the same geometric meaning as in simple CA because it is a point impossible to observe—a sample can not be present just for one variable and non-existent elsewhere. But one can think about the row–column relationship as an average of separate CA-type relationships across the column variables as follows. Each row, for example education group `e5`, has a profile across each

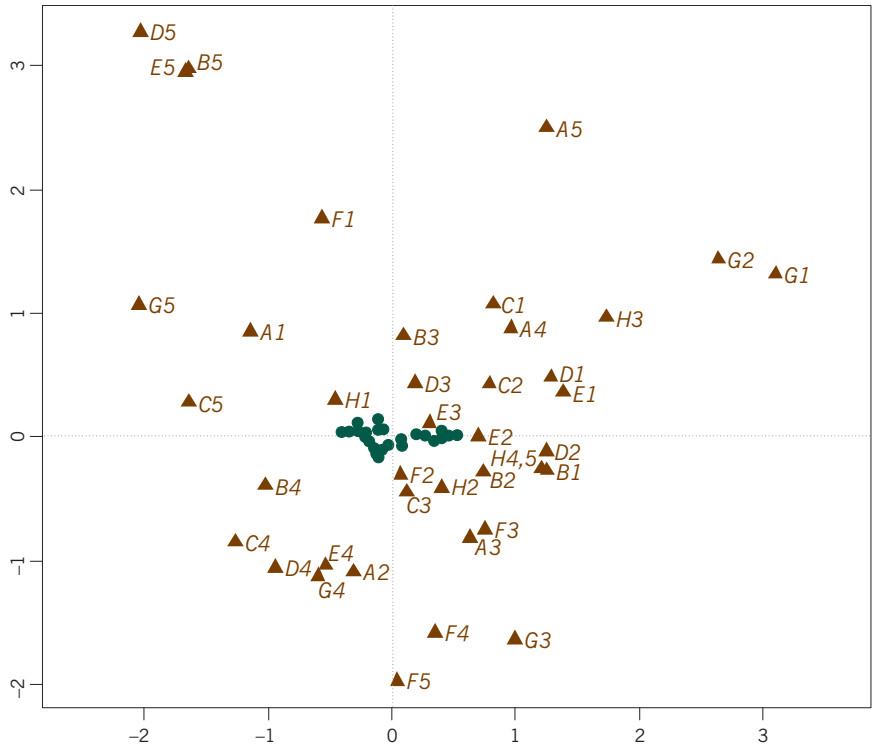


Exhibit 9.3:
Asymmetric CA map of the concatenated table of Exhibit 9.1. The positions of the row points (in green) are identical to those in Exhibit 9.2, as well as the inertias and percentages of inertia

column variable. There are thus 8 different weighted average positions that one could compute for e5, computed the same way as in a simple CA—e5's position is then the ordinary average of these 8 positions. Exhibit 9.4 illustrates the idea by showing the eight averages around e5 and also those around ma4, males in the fourth age group. There is much more variance around e5 compared to ma4. Respondents in the highest education group e5 react to statement *G* ("a man's job is to work; a woman's job is the household") with a relatively high level of disagreement, whereas with respect to statement *E* ("running a household is just as satisfying as a paid job") they are closer to the average opinion. Taking all 8 statements into account, their average position is the most extreme on the liberal side of the map. The attitudes to the individual questions by males in the 45-55 years age group, on the other hand, are much more similar, slightly to the conservative/traditional side of average.

Between- and within category inertia

This way of showing each demographic category's position as an average across the questions suggests an interesting decomposition of inertia for each demographic category, into a "within-category" component across the 8 questions and a "between-category" component. The "between" component is nothing else but

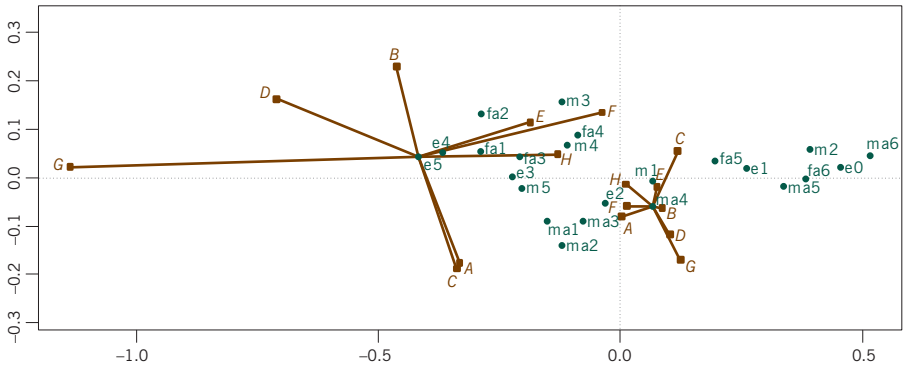


Exhibit 9.4: *Map of row points (demographic categories) of the concatenated table, illustrating two points,* e5 *and* ma4*, at the average of their positions with respect to the 8 variables (for example, "G" for point* e5 *is the weighted average position of* e5 *with respect to the categories for question G, using the profile values for* e5 *across G)*

the usual measure of inertia, which is the measure of dispersion of the demographic categories, whereas the "within" component is a measure of the dispersion across the 8 questions for each demographic category. The table in Exhibit 9.5 summarizes how much each demographic category contributes to the total "within" component, measured in permills (thousandths). This table shows that ma4 has the smallest contribution (5/1000) of all categories, and e5 an above average contribution (61/1000). The lowest two education groups and the oldest age group, both male and female, have the highest contributions—thus e0, for example, would show a much higher dispersion than e5 across the questions if its 8 individual points (of which it is the average) were drawn as in Exhibit 9.4.

Contribution biplot for concatenated table

Both the symmetric map of Exhibit 9.2 and the asymmetric biplot of Exhibit 9.3 have their particular advantages but neither tells the analyst which categories of the variables are driving the solution—this can be seen using the contribution biplot (see Chapter 8), which multiplies the standard coordinates of the column categories by the square roots of their corresponding masses. The contribution

Exhibit 9.5: *Permill contributions of each category to the dispersion across the questions, or "within-category" inertia*

Marital Status		*Education*		*Gender × Age*			
				Male		*Female*	
m1	14	e0	149	ma1	11	fa1	37
m2	81	e1	94	ma2	16	fa2	50
m3	8	e2	17	ma3	7	fa3	22
m4	5	e3	64	ma4	2	fa4	12
m5	55	e4	37	ma5	37	fa5	11
		e5	61	ma6	103	fa6	107
TOTALS	*163*		*422*		*176*		*239*

biplot is shown in Exhibit 9.6—the directions of the category points are identical to those of Exhibit 9.3 (thus, calibrations along these directions would be identical), but now the squares of their coordinates are equal to their contributions to the respective principal axes. Immediately it is clear that variable *G* ("a man's job is to work; a woman's job is the household"), especially categories *G5* opposed to *G2*, is the biggest contributor to the first (horizontal) axis—in fact, these two categories alone contribute 31% to the first axis, which is itself explaining 82.1% of the total inertia in the data. Notice that it is the "agree" category on the right which opposes the "strongly disagree" on the left—in Exhibits 9.2 and 9.3, which show this category's positional information as a scale value, the "strongly agree" category *G1* is situated further to the right than *G2*, as expected, but as far as contributing to this axis is concerned *G1* is less important, probably due to the fact that not many people give this response.

Now that we see which categories of attitude are driving the solution, there is an interesting interpretation of the vertical second dimension on the left hand side of Exhibit 9.6, even though this dimension explains only 4.4% of the total inertia. The biggest contributors are (at the top) *D5*, *E5* and *F1*, expressing the strongest support for women working, whereas at the bottom we have *G4*, *A2*, *D4* and *E4*, expressing moderate support for women working. The corresponding contrast is between the divorced marital group and the female groups up to the ages of 55 on top, and the male groups up to age 45. This contrast between males and females

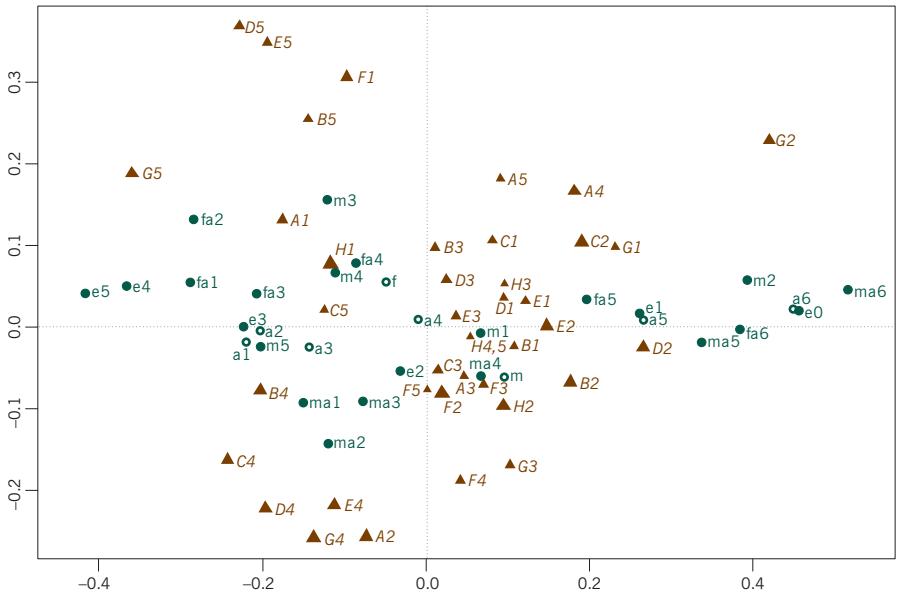


Exhibit 9.6:
Contribution biplot of the concatenated table of Exhibit 9.1, with column coordinates equal to the standard coordinates multiplied by the square roots of the respective column masses. The gender and age groups have been added as supplementary points (empty circle symbols). The positions of the row points (in green) are identical to those in Exhibits 9.2 and 9.3, as well as the inertias and percentages of inertia

does not exist in the older age groups on the right hand side of the display, where the demographic groups are closer together on the vertical axis.

Supplementary points

Exhibit 9.6 shows some additional points, for the two gender and six age groups. In the analysis these two variables had been combined interactively to form 12 groups, but the original categories can also be displayed. From the graphical viewpoint these points are just the weighted averages (centroids) of their displayed component groups: for example, the point `a1`, denoting the youngest age group, is between the female and male points for this group, `fa1` and `ma1`. It is at the weighted average of these two points, weighted by the numbers in the respective female and male subgroups. Similarly, the point `f` for females, is at the weighted average of the six female subgroups, `fa1` to `fa6`.

Analytically, supplementary points define additional profiles that are not used to establish the solution space, but are projected onto that space afterwards. The coordinates of the supplementary row points in this example are obtained by computing scalar products between the profile elements and the standard column coordinates: $\mathbf{D}_r^{-1}\mathbf{P\Gamma}$ in the notation of Chapter 8, where the profile is calculated across all Q variables (i.e., summing to 1 across all the variables). Equivalently, following the way the joint display in Exhibit 9.3 was interpreted, compute for each column variable the weighted average position of the row using its profile just across that variable (i.e., summing to 1 across that variable), and then average these positions (8 of them in this case) to situate the supplementary point.

SUMMARY: Multiple Correspondence Analysis Biplots I

1. A *concatenated table* is a block matrix composed of several contingency tables cross-tabulating the same sample of cases between two sets of variables. If there are L variables in the first set and Q in the second set, then there are $L \times Q$ subtables constituting the concatenated table, and each subtable has the same grand total, equal to the sample size.
2. The CA of a concatenated table is an average picture of the pairwise relationships between the two sets of variables. Its inertia is the average of the inertia of the subtables and the graphical display is the best approximation to all the subtables. One can think of this analysis as a compromise among all possible simple CAs of the subtables, using only one set category points for each row and column variable.
3. The asymmetric biplot of a concatenated table usually shows the set of points in principal coordinates close to the origin and far from the other set in standard coordinates, in which case a separate plot of the "inner" set of points is required.
4. Each category in principal coordinates, say a row category, is the average of a mini-cloud of points, one point for each of the column variables. It is useful to

measure the dispersion within each of these mini-clouds because this gives information about the variance of the category across the column variables.

5. The contribution biplot of a concatenated table displays one set of points in principal coordinates to show their interpoint distances, and the other points in standard coordinates multiplied by the square root of the respective masses (these are the usual masses that sum to 1 across all the variables). The latter set of points then indicates how they contribute to the construction of the axes of the representation space.
6. A *supplementary point* is an additional row or column of data with a profile that is displayed afterwards by projection onto the biplot. One can think of this row or column being in the analysis from the start but with zero mass assigned to it, hence having no influence on the solution.

CHAPTER 10

Multiple Correspondence Analysis Biplots II

Multiple correspondence analysis (MCA) is the CA of a special type of concatenated table, where a set of variables is cross-tabulated with itself. Therefore we can say that, whereas in Chapter 9 the concatenation was made *between* two sets of different variables, in this chapter the concatenation is made of variables *within* one single set. This concatenated matrix has identical rows and columns and is thus square symmetric. It includes cross-tabulations between each variable and itself, which are diagonal matrices of perfect association. These perfect associations are impossible to represent in a low-dimensional display, so any biplot as described in Chapter 9 of this particular concatenated matrix would be degraded as far as overall quality of data representation is concerned. This problem is avoided with a simple adjustment of the solution (or by using an alternative approach called joint correspondence analysis). In this chapter we show some possibilities for biplots in this MCA context, where the rows and columns are identical, and also how individual case points, or group averages of cases, can be displayed.

Contents

Burt matrix

Using the same "women" data set for Spain as in Chapter 9, Exhibit 10.1 shows part of the concatenated table of all the cross-tabulations of the questions with one another (here, for the moment, we do not consider the demographic variables). This is a square symmetric block matrix, since the cross-tabulation of variable q with variable s is the transpose of that of variable s with variable q. This ma-

Exhibit 10.1: *Part of the Burt matrix of the eight variables of the "women" data set, showing the first three variables cross-tabulated with one another, including the cross-tabulations of perfect association between each variable and itself down the diagonal blocks*

	A1	*A2*	*A3*	*A4*	*A5*	*B1*	*B2*	*B3*	*B4*	*B5*	*C1*	*C2*	*C3*	*C4*	*C5*	
a1	397	0	0	0	0	19	113	42	132	91	37	91	45	154	70	…
a2	0	932	0	0	0	18	362	126	405	21	21	405	128	359	19	…
a3	0	0	91	0	0	2	44	22	21	2	8	44	25	13	1	…
a4	0	0	0	598	0	40	411	42	101	4	48	422	36	88	4	…
a5	0	0	0	0	89	45	26	5	6	7	51	26	3	6	3	…
b1	19	18	2	40	45	124	0	0	0	0	80	34	4	3	3	…
b2	113	362	44	411	26	0	956	0	0	0	52	673	65	154	12	…
b3	42	126	22	42	5	0	0	237	0	0	12	82	82	59	2	…
b4	132	405	21	101	6	0	0	0	665	0	8	182	79	378	18	…
b5	91	21	2	4	7	0	0	0	0	125	13	17	7	26	62	…
c1	37	21	8	48	51	80	52	12	8	13	165	0	0	0	0	…
c2	91	405	44	422	26	34	673	82	182	17	0	988	0	0	0	…
c3	45	128	25	36	3	4	65	82	79	7	0	0	237	0	0	…
c4	154	359	13	88	6	3	154	59	378	26	0	0	0	620	0	…
c5	70	19	1	4	3	3	12	2	18	62	0	0	0	0	97	…
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	

trix is called a *Burt matrix*, after the psychometrician Cyril Burt, who first considered this type of data structure.

Like the concatenated table of Chapter 9, this one has the property that each subtable has a sum equal to the sample size, and the row margins are all the same across the variables in each row block and the column margins are all the same across the variables in each column block. Hence, the inertia of the Burt matrix of, in general, Q variables is the average of the inertias in the Q^2 subtables. In this eight-variable example there are eight cross-tabulations down the diagonal, each variable cross-tabulated with itself, which are diagonal matrices of perfect association, each having an inertia of 1 less than the number of categories, i.e. 4 in this example (a 5×5 diagonal matrix has four principal inertias, each equal to 1). These high values inflate the total inertia considerably, so we do not include them in the computation of the total inertia, preferring to define the total inertia to be explained as the average of all the other ("off-diagonal") subtables, of which there are $Q^2 - Q = Q(Q - 1)$. This is called the *adjusted inertia* in MCA. In fact, since the matrix is symmetric we can compute the adjusted inertia as the average of the upper or lower triangle of subtables, of which there are $\frac{1}{2}Q(Q - 1)$. If the usual total inertia of the complete Burt matrix, inertia($\mathbf{B}$), is available, the constant amounts due to the problematic diagonal matrices can be simply subtracted, and the adjusted inertia can be shown to be:

$$\text{adjusted inertia of Burt matrix} = \frac{Q}{Q-1}\left(\text{inertia}(B) - \frac{J-Q}{Q^2}\right) \qquad (10.1)$$

where Q = number of variables, J = total number of categories in all the variables. For example, in this example the total inertia of **B** is equal to 0.6776, but when we make the calculation removing the contributions of the eight diagonal matrices it reduces to (remembering that we combined two categories of *H*, so there are 39 categories in total):

$$\text{adjusted inertia of Burt matrix} = \frac{8}{7}\left(0.6776 - \frac{39-8}{64}\right) = 0.2208$$

(If one computes the individual inertias in the $\frac{1}{2} \times 8 \times 7 = 28$ cross-tabulations between pairs of variables and averages them, the result is identical.)

Indicator matrix

In Exhibit 10.2, on the left, the first five records of the original respondent-level data are shown, first the response categories to the eight questions, then the demographic groups for gender, marital status, education, age and the gender-age combination. On the right we see an alternative way of coding the data, where the columns are *dummy variables*, one column for each category of response, and coded with a 1 to indicate the response category, otherwise 0. The matrices on the right are called *indicator matrices*. The indicator matrices can be used to construct concatenated tables: for example, if we denote the 39-column indicator matrix for the eight questions *A* to *H* as **Z**, then the Burt matrix **B** is simply:

$$\mathbf{B} = \mathbf{Z}^{\mathsf{T}}\mathbf{Z} \qquad (10.2)$$

A	B	C	D	E	F	G	H	g	m	e	a	ga
2	4	3	3	4	1	4	1	1	1	3	4	4
3	2	2	3	4	1	3	2	2	2	1	6	12
2	4	4	4	2	2	5	1	1	5	4	2	2
1	3	2	3	2	4	4	2	2	1	5	5	11
2	4	2	3	4	4	5	1	2	1	5	3	9
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

A1	A2	A3	A4	A5	B1	B2	B3	B4	B5	⋯	g1	g2	m1	m2	m3	m4	m5	⋯
0	1	0	0	0	0	0	0	1	0	⋯	1	0	1	0	0	0	0	⋯
0	0	1	0	0	0	1	0	0	0	⋯	0	1	0	1	0	0	0	⋯
0	1	0	0	0	0	0	0	1	0	⋯	1	0	0	0	0	0	1	⋯
1	0	0	0	0	0	0	1	0	0	⋯	0	1	1	0	0	0	0	⋯
0	1	0	0	0	0	0	0	1	0	⋯	0	1	1	0	0	0	0	⋯
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋯

Exhibit 10.2:
Data for first five respondents (out of 2107) in the "women" data set, showing on the right the corresponding indicator coding of some of the variables

The total inertia of an indicator matrix of Q variables with a total of J categories can be shown to be equal to a constant which depends only on J and Q:

$$\text{total inertia of indicator matrix} = (J - Q)/Q \tag{10.3}$$

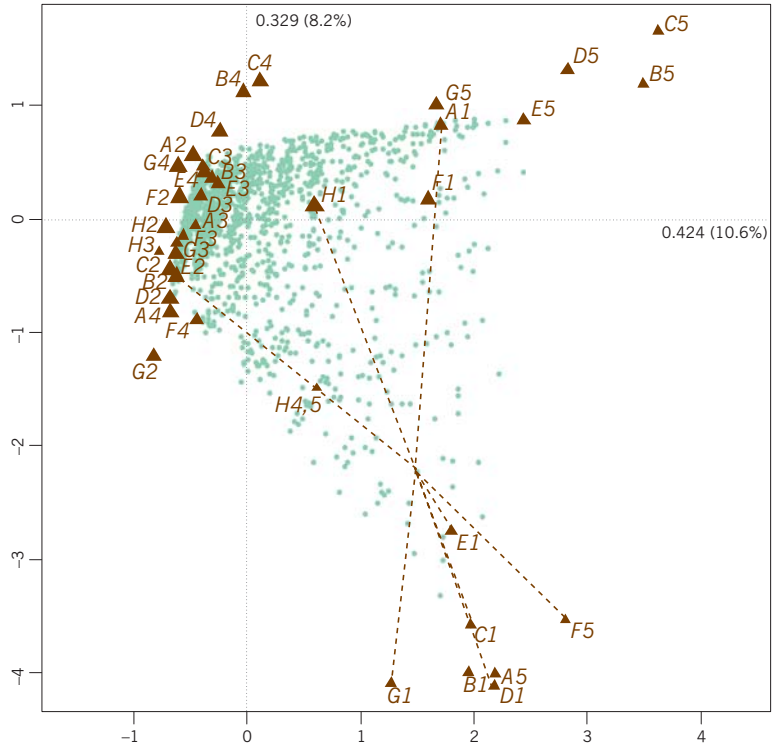
In the present example, this inertia would be equal to $(39 - 8)/8 = 3.875$.

Indicator matrix biplot

All the above results are relevant to understanding biplots for multiple correspondence analysis (MCA). MCA is classically defined as the CA algorithm applied to either the indicator matrix **Z** or the Burt matrix **B**. It is well-known that the two alternatives lead to exactly the same standard coordinates for the variable categories, while the singular values in the CA of **B** are the squares of those in the CA of **Z**. What interests us here is how to biplot the results meaningfully so that data are successfully recovered in the graphical representation. For example, consider the asymmetric map of the indicator matrix, with rows in principal coordinates and columns in principal coordinates, shown in Exhibit 10.3.

In contrast to the biplots of Chapter 9, which showed relationships between the question responses and the demographic categories, this biplot shows the relationships amongst the response categories themselves. All the extreme categories (the 1's and 5's) are on the right and all the moderate categories (2's and 4's) and

Exhibit 10.3: *Asymmetric map/biplot of the 2107 × 39 indicator matrix of the eight questions of the "women" data set. Each respondent point is at the average of the corresponding eight response categories —an example is shown of a respondent linked to her responses* A1, B2, C1, D1, E1, F5, G1, H1



middles (3's) are on the left. The main feature of the data is thus the opposition between respondents giving extreme opinions on the issue of working women and those with moderate attitudes. The conservative-liberal dimension in the responses is reflected in the vertical spread of the response categories, with conservative views at the bottom and liberal views at the top, on both extreme and moderate sides of the map.

Each row (respondent) is displayed as a green dot, and because each row of the indicator matrix consists of 0s apart from eight 1s in the response category positions, its profile consists of 0s except for the value 1/8 in those positions. The barycentric (or weighted average) property in the asymmetric map implies that each row point is at the ordinary average position of its eight response categories, for example the point shown in Exhibit 10.3, which is situated towards the extreme conservative region of the map. It is clear that in such a biplot there is no question of calibrating axes or trying to recover the 0/1 data—this is borne out by the fact that percentages of inertia are very low (10.6% and 8.2% respectively on the axes), which is typical if the indicator matrix is analyzed.

While it is interesting to see the whole cloud of respondent points, the biplot of Exhibit 10.3 can be made more meaningful by representing subgroups of points, for example the demographic groups such as male/female, or high/low education. This is explained below using the supplementary point idea, but basically the idea is to remove the individual case points and rather display average points for individuals in specified demographic groups.

Adjusted inertias

A biplot based on the Burt matrix is similar to the one we had for concatenated tables in Chapter 9, except for two major differences: the row and column variables are the same, and the tables of perfect association down the diagonal of the Burt matrix need to be avoided in some way. Our objective is to achieve a biplot that reconstructs the profiles of all tables apart from those in the diagonal blocks, and we have already proposed an adjusted total inertia that omits these cross-tabulations of perfect association. A simple adjustment of the singular values turns out to be just what is necessary so that the MCA solution best fits the off-diagonal tables. If λ_k denotes the k-th principal inertia of **B**, then only those axes for which $\sqrt{\lambda_k}$ are larger than $1/Q$ are retained (notice that the $\sqrt{\lambda_k}$'s are exactly the principal inertias of the CA of the indicator matrix). The adjustment is as follows:

$$\frac{Q}{Q-1}\left(\sqrt{\lambda_k}-\frac{1}{Q}\right),\ k=1,\ 2,\ldots,\quad \text{for}\ \sqrt{\lambda_k}>\frac{1}{Q} \tag{10.4}$$

The values in (10.4) effectively replace the singular values in the MCA and their squares are the *adjusted principal inertias*, which can be expressed relative to the adjusted total inertia of (10.1) to obtain percentages of inertia.[6]

Burt matrix biplot

We now have all the results necessary to define a biplot of the Burt matrix. First, we perform the MCA of the complete Burt matrix, giving us the standard coordinates identical to those used to plot the categories in Exhibit 10.3. The number of principal inertias λ_k that satisfy $\sqrt{\lambda_k} > 1/8$ is equal to 9. Applying the adjustment (10.4), the first two are equal to:

$$\frac{8}{7}\left(\sqrt{0.1801}-\frac{1}{8}\right) = 0.3422 \quad \text{and} \quad \frac{8}{7}\left(\sqrt{0.1079}-\frac{1}{8}\right) = 0.2326$$

These two values replace the singular values in the MCA and then the asymmetric biplot in Exhibit 10.4 can be drawn, with the rows, say, in principal coordinates (standard coordinates multiplied by the above adjusted singular values) and the columns in standard coordinates. Their squared values quantify the amount of inertia accounted for by each axis: $0.3422^2 = 0.1171$ and $0.2326^2 = 0.0541$, and relative to the adjusted total inertia of 0.2208 calculated previously, they explain 53.0% and 24.5% respectively.

One might ask what the benefit is of representing the categories twice, since the row principal coordinates are at the same relative positions along the principal axes as the column standard coordinates. The answer is that, like in any biplot, a biplot axis can be drawn through the point *C5*, for example, in standard coordinates, and then the profile values of all other categories on *C5*—except for the categories of the variable *C* itself—can be lined up by their projections onto that axis. So the fact that categories *C5* and *B5* are close means that all other response categories have similar profile values on these two categories. A possibly more interesting biplot of the two sets of identically labelled points is provided by the contribution biplot—since we are only interested in the directions of the biplot axes, we can change the lengths of the set of points in standard coordinates to reflect the contributions to the principal axes.

MCA contribution biplot

Since the row configuration in principal coordinates gives us the essential information for interpreting inter-category associations, we can use the column con-

6. The set of dimensions for which $\sqrt{\lambda_k} > 1/Q$ will usually account for less than 100% of the inertia in the off-diagonal cross-tabulations. To account for 100% of the inertia, another form of MCA called *joint correspondence analysis* needs to be used (not treated in this book).

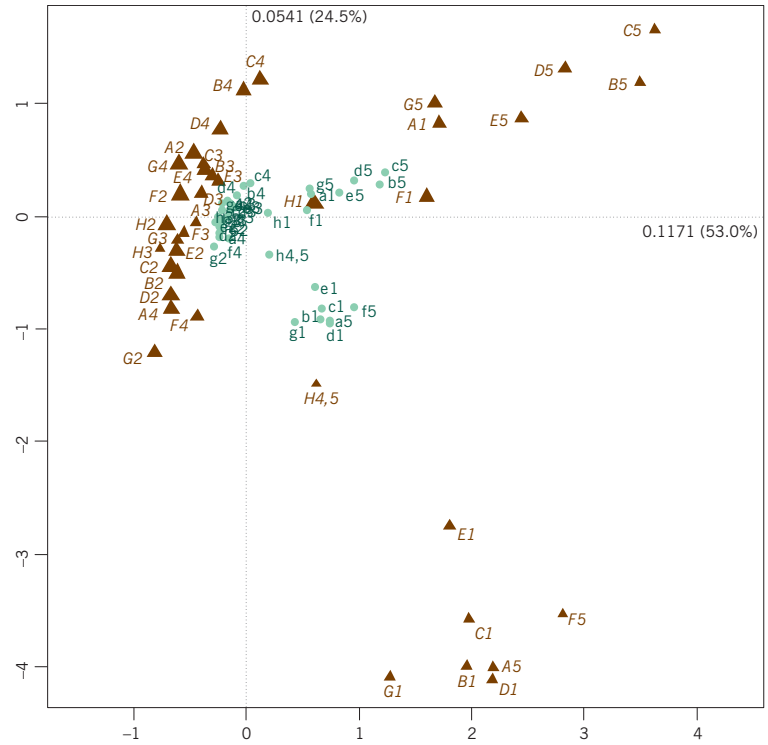


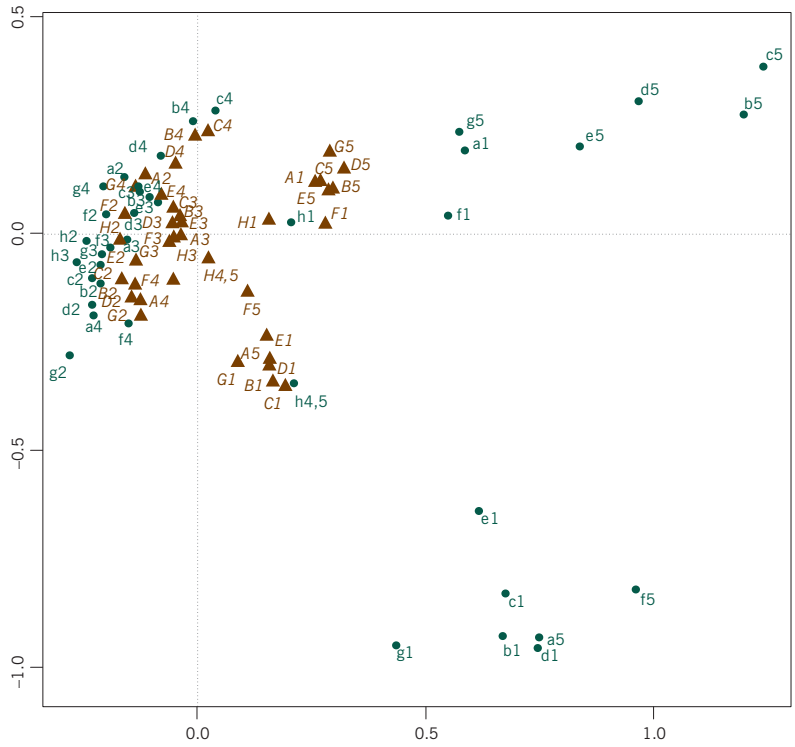
Exhibit 10.4:
Asymmetric map/biplot of the Burt matrix: columns in standard coordinates and rows in principal coordinates using adjusted principal inertias. Percentages of inertia on the two axes are 53.0% and 24.5% respectively

figuration to show which variables and which categories are contributing the most to the solution. The column points in standard coordinates are thus multiplied by the square roots of their masses to obtain the standard MCA biplot of Exhibit 10.5. Notice that the contribution coordinates of the middle categories (3's) are close to the centre, so these categories play a minor role in this biplot. However, they all contribute strongly as a group to the third dimension (not shown here). The separation of the middle response categories is a common phenomenon in survey data and is the theme of the case study in Chapter 14.

Supplementary points

As we saw in Exhibit 10.3, every respondent (a row of the indicator matrix) has a position in the biplot depending on that respondent's particular choice of responses. Groups of respondents can be displayed by their mean positions and, optionally, some type of confidence region for each mean. Geometrically, this is simply finding the average of all respondents in the lowest education group, for example, in the display and this gives a point e0, or finding the average of all females in the age group up to 25 years, and this gives a point fa1. Analytically, this is achieved by adding an extra row to the data matrix which accumulates all the frequencies for males across the variables, or all females in the first age group, in

Exhibit 10.5:
MCA contribution biplot. The row points (principal coordinates, in green—same as in Exhibit 10.4) show chi-square distances between the categories, while the column points (contribution coordinates, in brown) serve as directions for biplot axes as well as quantifying the contributions of the categories to each dimension



other words, exactly the rows of the concatenated matrix in Exhibit 9.1. The profiles of these rows define points in the MCA space which are the means of the corresponding demographic group.

Exhibit 10.6 shows the demographic group means added to the display of Exhibit 10.5 (categories shown in contribution coordinates only).

Because the upper right represents the most liberal attitude towards working women and the lower right the most conservative (and lower left the more moderate conservative attitude), one can see that the oldest male group and lowest education group are the most conservative while at the top end it is the two highest education groups and the younger female groups that are the most liberal. Not only have the combinations of gender and age been added but also the points representing all males (`m`) and all females (`f`) and each age been group (`a1` to `a6`). A similar result to that observed in Exhibit 9.6 can be seen at the top where the young females (e.g., `fa1`) tend to be strongly liberal, whereas the corresponding male group (`ma1`) tends to the more moderate liberal side (upper left). The point `a1` representing the age group as a whole is between these two points. At

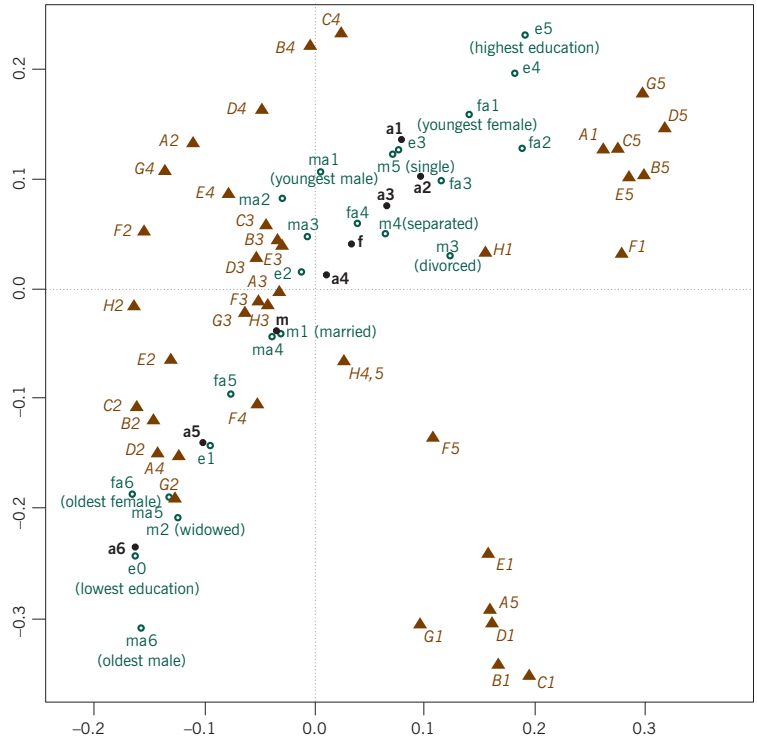


Exhibit 10.6:
MCA contribution biplot, showing variables in their contribution positions and supplementary points added for the demographic groups

bottom left we have a6 lying between fa6 and ma6 (but closer to fa6 since there are more females in this age group) but here the difference is simply that males are more conservative than females and there is not so much the strong versus moderate attitude observed in the youngest groups. In fact, there is no specific demographic group tending to the extreme conservative attitudes at bottom right.

In summary, Exhibit 10.6 is displaying many aspects of the original 2107 × 12 data matrix (8 questions and 4 demographic variables (see Chapter 9). In the first instance it shows the principal axes of the question categories (green points in Exhibit 10.5) in such a way as to maximize the inertia accounted for in all the pairwise relationships between the questions. Second, it shows the question categories (brown points) with lengths related to their contributions to the solution axes and with directions that define biplot axes, onto which the green points (in Exhibit 10.5) can be projected (excluding the green points for the same question, because projecting a question onto itself has been purposely avoided in finding the solution). Third, demographic groups are displayed as supplementary mean points of the respondents in each respective group.

The MCA in Exhibit 10.6 is primarily focused on displaying the associations among the eight questions, and secondarily on showing how the demographic groups relate to these dimensions. Exhibit 9.6, on the other hand, is focused on the associations between the demographic variables and the questions. From the biplot point of view the difference can be explained as follows. In Exhibit 10.6 if a biplot axis is drawn through *G5*, for example, then the projections of all the question categories a1 to a5, b1 to b5, etc. (but not g1 to g5), will be approximating the profile values of these categories on *G5*—the overall quality of display of all these profiles with respect to the biplot axes is 77.5%, the inertia explained by the first two axes. The demographic categories can also be projected onto directions such as *G5*, but their displayed profile values have not been specifically optimized—the overall quality of display of these supplementary profiles is only 58.1% in Exhibit 10.6. In Exhibit 9.6, on the other hand, the projections of the demographic categories onto the biplot axes defined by the response categories (for example, onto *G5*), were optimal and there the overall quality of display was 86.5%.

SUMMARY: Multiple Correspondence Analysis Biplots II

1. One of the ways of defining and thinking about MCA is as the CA of the concatenated table of cross-tabulations of a set of categorical variables with themselves. This square symmetric block matrix is called the *Burt matrix*.
2. The Burt matrix includes down its diagonal blocks the cross-tabulations of each variable with itself, and these tables inflate the total inertia of the problem, leading to low percentages of inertia explained on the principal axes if the Burt matrix is displayed.
3. A simple adjustment of the principal inertias and the total inertia optimizes the solution to the off-diagonal tables that cross-tabulate distinct pairs of variables.
4. Because the rows and columns of the Burt matrix are identical, the contribution biplot is particularly useful: one of the sets, for example the rows, shows the category points in principal coordinates and so displays inter-profile distances, while the other set can display the categories both as biplot axes and with lengths related to their contributions to the solution.
5. In all MCA biplots the respondent points can also be displayed, but it is usually more interesting to show various average positions of groups of respondents in terms of their demographic characteristics. These are added as supplementary points.

CHAPTER 11

Discriminant Analysis Biplots

Discriminant analysis is characterized by a classification of the cases in a data set into groups, and the study of these groups in terms of the variables observed on the cases. The ultimate aim is to discover which variables are important in distinguishing the groups, or to develop a classification rule for predicting the groups. Up to now biplots have been displaying, as accurately as possible, data on individual case points. One exception, which is a type of discriminant analysis, was in Chapter 9 where the biplot was designed to show differences between demographic groups in the data rather than show individual differences. This biplot of group differences did not take into account correlations between the variables, while other approaches—such as Fisher's linear discriminant analysis—use a distance function between cases which does take into account correlations. In this chapter the geometry of these two approaches is explained and biplots are developed to display the cases and their group averages, along with the variables, in a discriminant space that shows the group differences optimally.

Contents

Analyzing groups

A common feature of a cases-by-variables data matrix is the classification of the cases (or the variables) into groups. For example, in the case of the *Arctic charr* fish in the "morphology" data set, each fish is classified as male or female, and also whether it was caught near the shore or in the open waters. Are these groups of fish different in terms of their morphology? In the "women" data set we displayed differences between individual respondents in Exhibit 10.3, whereas in Exhibit 9.3 group differences were displayed. In all of these methods where individ-

ual-level differences are displayed, it is also possible to display aggregate-level differences. Analyzing group differences based on multivariate data is called *discriminant analysis*, or DA. The basic idea underlying DA is the analysis of group means, or centroids, rather than the individual data. In other words, what we have been doing up to now in analyzing N cases, say, can be thought of as a DA of N groups, each consisting of 1 case, whereas now we consider the DA of G groups, of size $N_1, \ldots, N_G$, where $N_1 + \cdots + N_G = N$. Although not so common, the same idea can be applied to the variables: instead of analyzing all the morphometric variables individually in the "morphology" data set, for example, the variables can be amalgamated by summation or averaging into predetermined groups.

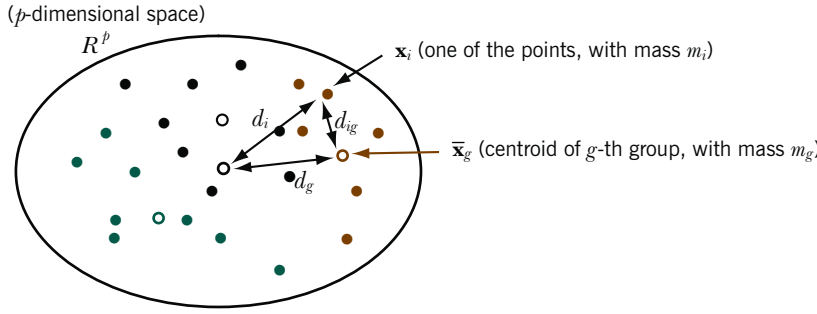
Between- and within-group variance/inertia

When cases are grouped, there is a decomposition of the total variance (or inertia in CA and LRA when points have different weights) into two parts: variance between groups, which quantifies how different the groups are, and variance within groups, which quantifies how internally heterogeneous the groups are:

$$\text{Total variance} = \text{Between-group variance} + \text{Within-group variance} \qquad (11.1)$$

The greater the between-group variance, the more homogeneous the groups must be—in the extreme case where groups consist of single cases, between-group variance is the total variance and within-group variance is zero. At the other extreme where there is no group separation, between-group variance is zero and the within-group variance is the total variance. The decomposition in (11.1) is the basis of analysis of variance for one variable, whereas in our context that variance is measured in multidimensional space, using the distance measure of the particular method, be it PCA, CA/MCA or LRA.

Exhibit 11.1: *The open circles represent the centroids of three groups (coloured in green, black and brown). Points have a distance d_i to the overall centroid, represented by the bold open circle. The distance of a member of group g to its group centroid is d_{ig}, and the distance from the centroid of group g to the overall centroid is d_g. Points have masses m_i and the aggregated mass in group g is m_g, which is assigned to the respective group centroid*



$$\text{Total inertia} = \text{Between-group inertia} + \text{Within-group inertia}$$

$$\sum_i m_i d_i^2 = \sum_g m_g d_g^2 + \sum_g \left(\sum_i m_i d_{ig}^2\right)$$

In PCA (principal component analysis) each case point typically has weight $1/N$ and the weights assigned to the group average points $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_G$ are then N_1/N, N_2/N, …, N_G/N. In LRA and CA/MCA the case points can have varying weights, r_i, in which case the group means are weighted by the accumulated weights of their respective cases. Exhibit 11.1 illustrates the decomposition of variance/inertia in a schematic way—here the groups are shown almost perfectly separated, but in practice there is a high degree of overlap and the group averages, or centroids, separate out much less. For example, in the "morphology" data set of Chapter 7, where LRA was used to visualize the 75 *Arctic charr* fish, there were four groups of fish defined by the combinations of sex (male/female, abbreviated as m/f) and habitat (littoral/pelagic, abbreviated as L/P)—these were labelled in Exhibit 7.3 as fL, mL, fP and mP. The total inertia in the full multidimensional space of the LRA was equal to 0.001961. The decomposition of inertia (11.1) with respect to the four sex/habitat groups turns out to be:

$$0.001961 = 0.000128 + 0.001833$$

The between-group inertia (0.000128) is only 6.5% of the total. In the Computational Appendix a permutation test is performed, showing that the between-group differences, although small, are statistically significant ($p = 0.015$) and worth investigating. A small between-group variance or inertia does not mean that there is no meaningful separation of the groups—groups can be separable and still have a high percentage of within-group variance. In Exhibit 7.3, however, the objective of the biplot was to separate the individual fish optimally in the two-dimensional view, not the groups of fish—separating the groups optimally in a low-dimensional display is the job of DA.

Example: LRA-DA biplot

The log-ratio discriminant analysis (LRA-DA) biplot of the four fish groups is shown in Exhibit 11.2. This is achieved by performing a regular LRA on the 4×26 matrix of group centroids, weighted by their respective aggregated masses (remember that in LRA, as in CA, the mass of a point is proportional to its marginal sum, so that the mass r_i of each fish is proportional to the total of its morphometric values). The dimensionality of the four centroids is three, so that dimension reduction to two dimensions means sacrificing only one dimension. The biplot shows that the main difference (along the horizontal axis) is that between the two littoral groups on the left and the two pelagic groups on the right. The second axis separates the females from the males, especially the female and male littorals.

To find out which ratios might be associated with these separations, the log-ratio of *Bc* relative to *Jw* is the longest horizontal link corresponding to the left-to-right littoral-pelagic contrast in Exhibit 11.2. Performing a two-group *t*-test between the

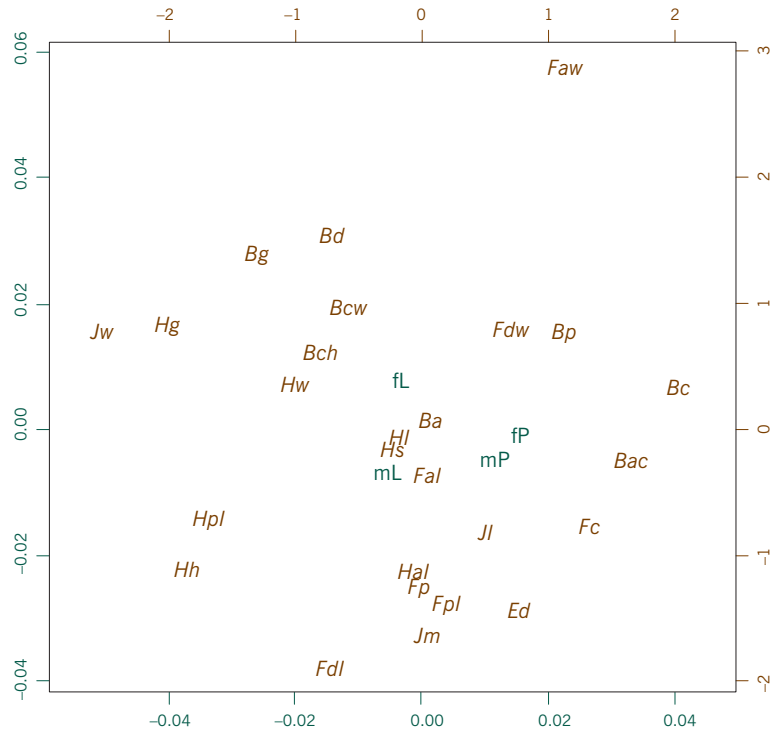


Exhibit 11.2: *LRA-DA biplot of the four fish groups in the "morphology" data set, with fish group centroids in principal coordinates and variables in standard coordinates. Because of the very small inertia of the centroids (0.000128), they are shown on a different scale. 79.9% of this inertia of the centroids is explained in the biplot*

pelagic and littoral groups on this log-ratio gives a difference on the log-scale of 0.083, which translates to a 8.7% increased ratio of *Bc*/*Jw* in pelagics compared to littorals, significant at $p = 0.006$. As for the difference between females and males on the vertical axis, the biplot suggests the log of *Faw*/*Fdl* to be a possible candidate. Indeed, the *t*-test shows a difference between females and males of 0.073 on the log-scale, which is a 7.6% increase in the *Faw*/*Fdl* ratio in females compared to males, with $p < 0.0001$.

Example: CA-DA biplot

Exhibit 7.3 was already an example of a DA between groups in a CA context, although there we showed the differences between several groupings of the cases simultaneously. To focus on only one set of groups, let us consider visualize and quantify the differences in attitudes between the marital status groups m1 to m5: respectively, "married", "widowed", "divorced", "separated" and "single". What we call a CA-DA is just a CA of the concatenated tables with the 5 rows being the marital groups, and 39 columns the categories of the 8 variables (remember that *H4* and *H5* were combined to give one column *H4,5*). Exhibit 11.3 shows the CA contribution biplot, comparable to Exhibit 9.6. Again, *G2* is an important category for the conservative end of the scale, typified by the group "widowed". In Exhibit 9.6 we saw that disagreement (category 4) to statements *C*, *D*, *E* and *G* all corre-

lated and were associated with younger males. In Exhibit 11.3 these categories are more spread out vertically, with *C4* an important category for "single" (disagree that family life suffers when a woman works) while *E4* and *G4* are important for "divorced" (disagree that running a household is just as satisfying as a paid job, and disagree that a man's job is to work and a woman's is the household).

Again, even though these between-group differences are meaningful and actually highly significant statistically (chi-square tests between marital status and each question all have $p < 0.0001$), the between-group inertia relative to the total is very small. This is because the total inertia of the original data in indicator matrix form is a fixed value and very high—see (10.3)—equal to 3.875 in this example. The between-group inertia could only attain this value in the extreme case that each of the five marital status groups gave identical responses within each group. In practice, the inertias of condensed tables like this one are very much smaller than the indicator matrix: in this example the total inertia of the five groups is 0.03554, which is only 0.917% of the total inertia of the indicator matrix. In the Computational Appendix we explain how to perform a permutation test to quantify the statistical significance of the between-group inertia.

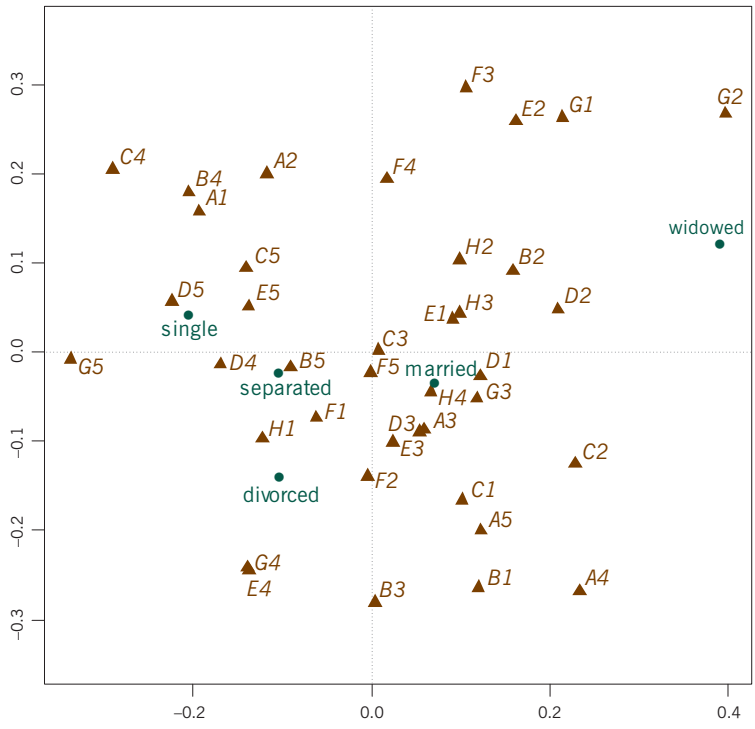


Exhibit 11.3:
CA-DA of marital status groups in the "women" data set, in terms of the 8 questions on women working. 90.7% of the inertia is displayed here

Mahalanobis distance

In the different variants of discriminant analysis described above, no mention was made at all of the effect of correlations between variables, which can distort the distances between group centroids. Suppose there are two groups of bivariate points, indicated by "x" and "o" in Exhibit 11.4, with their group means in boldface. The coloured ellipses in the left hand picture summarize the spread of each group, and show that the two variables are highly correlated within each group. The two centroids are at a certain distance apart. Now suppose that the "o" group lay in the position indicated by the dashed ellipse—its centroid is at the same distance from the "x" centroid, but the groups overlap much more. This anomaly can be removed by performing a transformation on the variables to de-correlate them, shown on the right hand side. The elliptical dispersion of the points is now spherical and, indeed, the "o" centroid is now much further away from the "x" centroid compared to the alternative in the dashed circle. The transformation involved is more than a standardization of the variables, because in the left hand picture of Exhibit 11.4 the two variables have the same variance. Rather, what is needed is a stretching out of the points in a direction more or less at right-angles to the axis of dispersion of the points—this is achieved by defining what is called the *Mahalanobis distance* between the cases, named after a famous Indian statistician.

Suppose **C** is the *within-groups covariance matrix* between the variables (this is defined in the next section). Then the Mahalanobis distance between two points **x** and **y** in multivariate space is

$$\text{Mahalanobis distance} = \sqrt{(\mathbf{x}-\mathbf{y})^{\mathsf{T}}\mathbf{C}^{-1}(\mathbf{x}-\mathbf{y})} \qquad (11.2)$$

If we omit the off-diagonal covariances in **C** so that **C** is the diagonal matrix of variances, then (11.2) is just the regular standardization of the variables. The presence of the off-diagonal covariances in **C** decorrelates the variables.

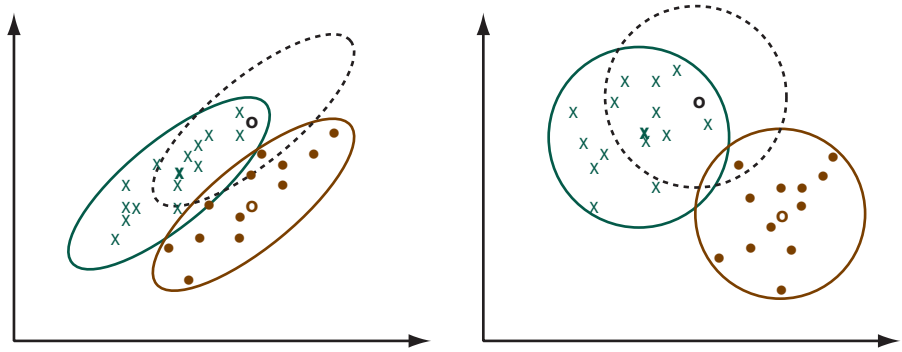


Exhibit 11.4: *The effect of high correlation between variables on the measure of between-group distance. On the right a transformation has been performed to remove the correlation—now the distances between points are Mahalanobis distances*

Linear discriminant analysis (LDA)

With this distance function in the space of the cases the same analysis of the centroids is performed as before—this is called *linear discriminant analysis* (LDA), attributed to the statistician R.A. Fisher, so sometimes called *Fisher discriminant analysis*. Given an cases × variables data matrix $\mathbf{X}$ ($I \times J$) where the cases are classified into G groups, denote by $\mathbf{x}_{ig}$ the vector of observations for the i-th case in the g-th group, with weight (mass) w_{ig}. The group masses are $w_1, w_2, \ldots, w_G$ ($w_g = \sum_i w_{ig}$) and the centroids $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \ldots, \bar{\mathbf{x}}_G$ ($\bar{\mathbf{x}}_g = \sum_i (w_{ig}/w_g)\mathbf{x}_{ig}$). Then the within-groups covariance matrix $\mathbf{C}$ is the weighted average of the covariance matrices computed for the groups separately:

$$\mathbf{C} = \sum_g w_g \mathbf{C}_g, \text{ where } \mathbf{C}_g = \sum_{i=1}^{N_g} (w_{ig}/w_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)(\mathbf{x}_{ig} - \bar{\mathbf{x}}_g)^\mathsf{T} \tag{11.3}$$

The theory of generalized PCA and contribution biplots in Chapters 5 and 6 applies here:

- Centre the group means: $\bar{\mathbf{Y}} = \bar{\mathbf{X}} - \mathbf{1}\mathbf{w}^\mathsf{T}\bar{\mathbf{X}} = (\mathbf{I} - \mathbf{1}\mathbf{w}^\mathsf{T})\bar{\mathbf{X}}$, where $\bar{\mathbf{X}}$ is the matrix of centroids in the rows, $\mathbf{w}$ is the vector of group weights; since the overall centroid $\bar{\mathbf{x}}^\mathsf{T}$ (written as a row vector) of the individual-level data in $\mathbf{X}$ is identical to the centroid $\mathbf{w}^\mathsf{T}\bar{\mathbf{X}}$ of the group centroids, we could also write $\bar{\mathbf{Y}} = \bar{\mathbf{X}} - \mathbf{1}\bar{\mathbf{x}}^\mathsf{T}$.

- Transform to Mahalanobis distance and weight by the masses before computing the SVD (singular value decomposition):

$$\mathbf{S} = \mathbf{D}_w^{½}\bar{\mathbf{Y}}\mathbf{C}^{-½}(1/J)^{½} = \mathbf{U}\mathbf{D}_\eta\mathbf{V}^\mathsf{T} \tag{11.4}$$

  where $\mathbf{C}^{-½}$ is the inverse of the *symmetric square root* of $\mathbf{C}$—this is calculated using the eigenvalue decomposition[7] of $\mathbf{C}$: $\mathbf{C} = \mathbf{X}\mathbf{D}_\lambda\mathbf{X}^\mathsf{T}$, hence $\mathbf{C}^{-½} = \mathbf{X}\mathbf{D}_\lambda^{-½}\mathbf{X}^\mathsf{T}$.

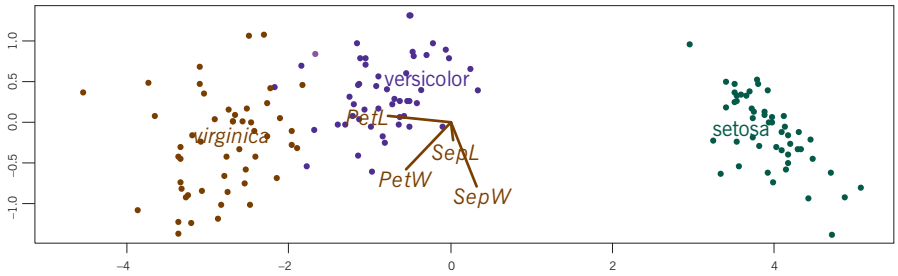
- Calculate the principal coordinates of the group centroids: $\mathbf{F} = \mathbf{D}_w^{-½}\mathbf{U}\mathbf{D}_\eta$ and the coordinates of the variables for the contribution biplot (for example): $\boldsymbol{\Gamma} = \mathbf{V}$.

This theory is illustrated with Fisher's famous "iris" data set, available in the R package (see the Computational Appendix). In this case, the decomposition of variance is:

$$9.119 = 8.119 + 1.000$$

7. Notice that the eigenvalue decomposition of a square symmetric matrix is the same as the singular value decomposition when the eigenvalues are all non-negative, as is the case here for the covariance matrix $\mathbf{C}$. Thus it is possible to calculate the square roots of the eigenvalues of $\mathbf{C}$.

Exhibit 11.5:
LDA contribution biplot of Fisher "iris" data. 99.1% of the variance of the centroids is explained by the first axis, on which PetL (petal length) is the highest contributor



Notice that the within-group variance is equal to 1, by construction. In this example the variance of the centroids accounts for 89.0% of the total variance, and the separation is excellent, as shown in Exhibit 11.5. The first principal axis of the centroids totally dominates, accounting for 99.1% of their variance. Petal length (*PetL*) and petal width (*PetW*) are seen to be the most important variables on this axis.

Notice that the individual points have been added to this LDA biplot, as supplementary points. To derive the coordinates of the individuals, notice from the algorithm above that the principal coordinates $\mathbf{F} = \mathbf{D}_w^{-½}\mathbf{U}\mathbf{D}_\eta$ of the group centroids are equivalently obtained by the transformation of the column variable points $\mathbf{V}$: $\bar{\mathbf{Y}}\mathbf{C}^{-½}(1/J)^{½}\mathbf{V}$. The coordinates $\mathbf{F}_{\text{case}}$ of individual case points are obtained in a similar way, using the centred matrix $\mathbf{Y}$ for the original individual-level data:

$$\mathbf{F}_{\text{case}} = \mathbf{Y}\mathbf{C}^{-½}(1/J)^{½}\mathbf{V} \tag{11.5}$$

In a similar fashion all the individual cases could have been added to the DA biplots of Exhibits 11.2 and 11.3, using the appropriate relationship in each analysis between the row and column points—this relationship is often referred to as the *transition formula* between rows and columns.

SUMMARY: Discriminant Analysis Biplots

1. Discriminant analysis is the analysis of group means, or centroids, of a set of multivariate points classified into pre-specified groups.
2. The centroids have masses (weights) equal to the sum of the masses of their members.
3. There is a decomposition of total variance/inertia of the set of points into that of the centroids, the between-group variance/inertia, plus the weighted average variance/inertia within the groups, the within-group variance/inertia. (Inertia is simply the alternative term for variance when the points have different weights; or conversely, variance is the special case of inertia when the weights of all points are equal).

4. The dimension reduction of the centroids follows the same algorithm as the corresponding PCA, LRA or CA/MCA method.
5. *Linear discriminant analysis* (LDA) is also a dimension-reduction method on a set of centroids, but uses the Mahalanobis distance based on the within-groups covariance matrix to decorrelate the data.
6. In all these variations of DA the contribution biplot displays the centroids in an optimal map of their positions, along with the variables so that the most important (i.e., most discriminating) variables are quickly identified.

CHAPTER 19

Constrained Biplots and Triplots

The pervading theme of this book has been the visualization of the maximum amount of information in a rectangular data matrix through a graphical display of the rows and columns, called the biplot. Often the rows are cases, displayed as points, and the columns are variables, displayed as vectors, and thanks to the scalar product property the projections of row points onto axes defined by the column vectors lead to approximations of the original data. Up to now no condition has been imposed on the solution apart from certain normalization conditions on the coordinates because of the indeterminacy of the matrix decomposition. In this final chapter we look at several ways of constraining the biplot display to have some additional condition on its solution. Imposing restrictions on a biplot necessarily makes it sub-optimal in representing the original data matrix, but in many situations such constraints add value to the interpretation of the data in relation to external information available about the rows or the columns.

Contents

More than a supplementary point

The idea of a constrained biplot can be illustrated using the "morphology" data set, the measurements of the 75 *Arctic charr* fish, and the log-ratio (LRA) biplot of Exhibit 7.3. The LRA biplot explained 37.5% of the variance (20.9% on the first axis, 16.6% on the second) of the 75 × 26 data matrix, which was logarithmically transformed and double-centred, called the *log-ratio transformation*. The body weight of each fish was also available, and it would be interesting to see if the body weight is related to the solution. This is achieved using the regression biplot of Chapter 2, where continuous variables can be added to an existing plot using

their regression coefficients on the dimensions of the map. A regression analysis is performed, of body weight as response and the fish coordinates on the two axes as explanatory variables, and the two standardized regression coefficients are used to draw this supplementary variable's direction. The coefficients turn out to be 0.116 and 0.203, with an R^2 of 0.055, and we could add a short vector to Exhibit 7.3 pointing towards the upper right (0.116 on the first axis, 0.203 on the second) to show the relationship of this additional variable to the biplot. An explained variance of 5.5% of the variable body weight, however, shows that this variable has little relationship with the biplot. The idea of adding the body weight variable was to see if there was any relationship between this variable and the shape of the fish (remember that it is the shape that the biplot is visualizing, not the size, thanks to the log-ratio transformation). For example, perhaps fish that are generally wider than they are longer may be heavier (this is frequently the case for humans!).

A more direct way of investigating this possible relationship is to *constrain* the first dimension of the biplot to be linearly related to body weight, so that body weight will coincide exactly with the first axis—that is, it will be 100% explained by the first axis—while the second axis will be the optimal axis not related to body weight but still trying to visualize the morphometric data as accurately as possible. As a spin-off we obtain a measure of exactly how much variance in the (log-ratio transformed) morphometric data is explained by body weight. Before we look at how the solution is obtained technically, let us look at the result of imposing the constraint, shown in Exhibit 12.1—body weight is now perfectly correlated with the first axis, pointing to the right. Body weight explains 4.0% of the variance of the morphometric variables (in the Computational Appendix we shall show that this percentage is highly significant statistically, with a p-value of 0.001), while the second axis (which is the first axis of the unconstrained space) explains 20.7%. A log-ratio link that is lying in this horizontal direction and which is long suggests the ratio *Bcw/Ed*, caudal body width relative to eye diameter—one might say the fat fish are heavy-tailed and beady eyed! Plotting body weight against this ratio does show a significant correlation of 0.489, and the slope of the relationship estimates a 1.84% increase in the ratio *Bcw/Ed* for every 10g increase in body weight (since $\exp(0.00182 \times 10) = 1.0184$). The variable *Bd*, body width at dorsal fins, is also in the same direction as *Bcw*, again supporting the not surprising result that heavier fish are wider. In a separate analysis a much weaker relationship was found with the morphological variables and body length.

Constraining by a categorical variable

Suppose that we want to constrain the biplot to be related to an external categorical variable; for example, the four-category sex-habitat variable for the fish data again.

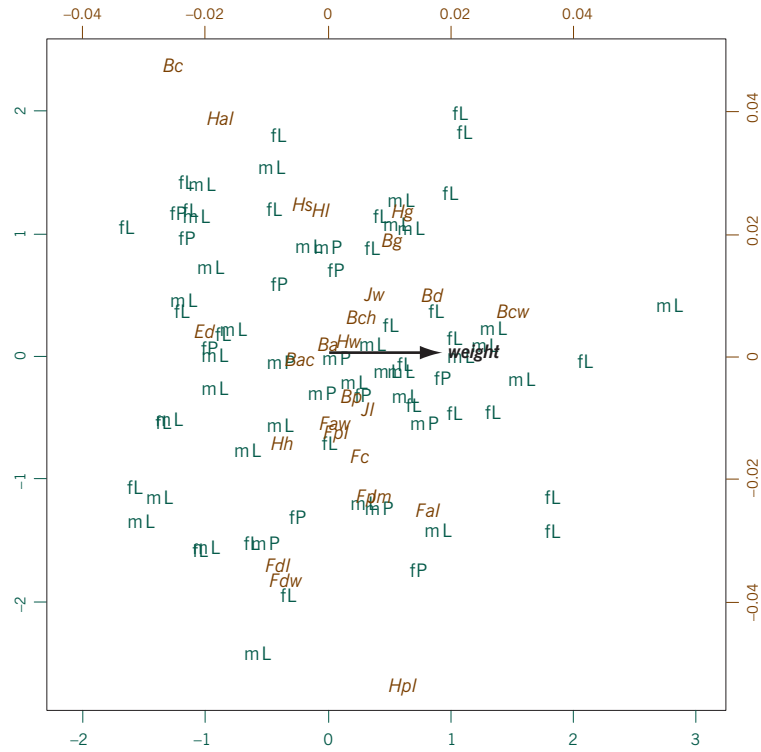


Exhibit 12.1:
The log-ratio biplot with the first axis constrained to be body weight. Rows (fish) are in standard coordinates, columns (morphometric variables) in principal coordinates. The constraining variable body weight follows the scale of the rows

This turns out to be equivalent to the discriminant analysis of Chapter 11. The sex-habitat variable is coded as four dummy variables, and it can be shown that the constrained space is exactly the space of the four centroids, thus the LRA-DA of Exhibit 11.2 is obtained. In a similar way, if the "women" data set is analyzed at an individual level, that is the indicator matrix **Z** is analyzed, and if the solution is constrained to be linearly related to the five dummy variables for marital status, this results in the CA (also called CA-DA) of the aggregated data in Exhibit 11.3.

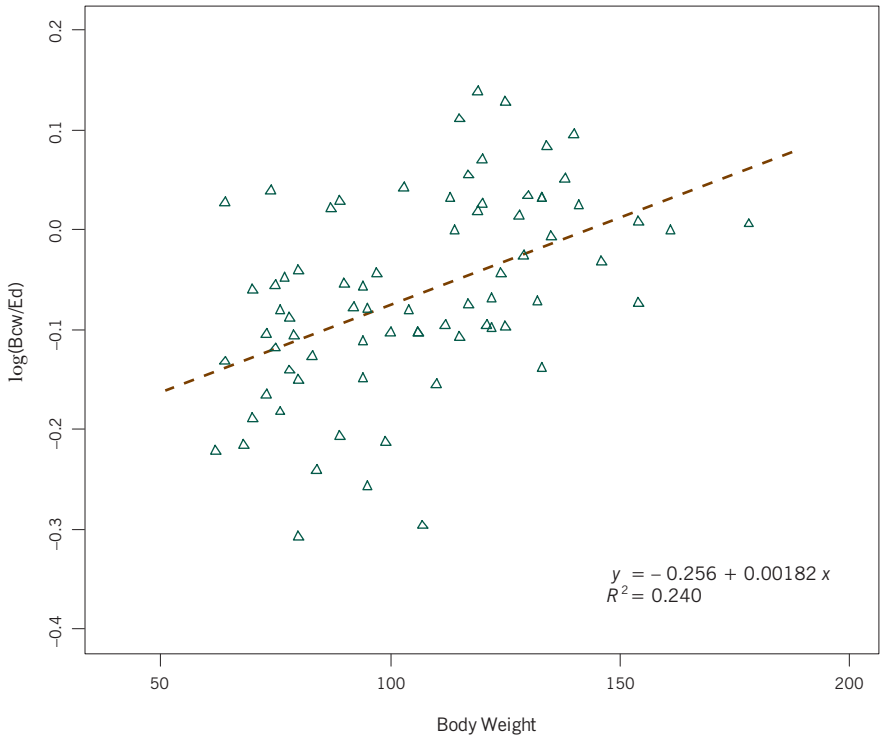
Constrained biplots

Whether it is a PCA, LRA or CA that is being performed, the imposition of the constraint is done in the same way. We give the theory for constrained CA, also known as *canonical correspondence analysis* (CCA), and then explain how this applies to the other methods. Regular unconstrained CA involves the SVD of the matrix of standardized residuals—see (8.2):

$$\mathbf{S} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{P} - \mathbf{r}\mathbf{c}^{\mathsf{T}})\mathbf{D}_c^{-\frac{1}{2}} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^{\mathsf{T}} \tag{12.1}$$

Suppose that $\mathbf{X}$ denotes the $I \times M$ matrix of external constraining variables, and suppose that the columns of $\mathbf{X}$ have already been standardized. Notice that, be-

Exhibit 12.2:
The possible relationship between the log-ratio of Bcw to Ed and body weight that was diagnosed in the biplot



cause the rows are weighted by the masses in **r**, all calculations of mean and variance are performed using these masses, so the columns of **X** have weighted mean of 0 and weighted variance (inertia) of 1. Constraining the solution linearly means projecting **S** onto the space of **X**. The projection matrix is defined as follows (again, the masses are taken into account):

$$\mathbf{Q} = \mathbf{D}_r^{1/2}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{D}_r^{1/2} \tag{12.2}$$

(one can easily check that **Q** satisfies the condition of a projection matrix: $\mathbf{QQ} = \mathbf{Q}$, i.e. applying the projection twice is the same as applying it once). The constrained (or restricted) version of **S** is then:

$$\mathbf{S}^* = \mathbf{QS} \tag{12.3}$$

From here on the calculations continue just as for CA, first calculate the SVD and then the principal and standard coordinates—see (8.2) to (8.4):

SVD: $$\mathbf{S}^* = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T} \tag{12.4}$$

Principal coordinates of rows: $\mathbf{F}^* = \mathbf{D}_r^{-½}\mathbf{U}\mathbf{D}_\alpha$, of columns: $\mathbf{G}^* = \mathbf{D}_c^{-½}\mathbf{V}\mathbf{D}_\alpha$ (12.5)

Standard coordinates of rows: $\mathbf{\Phi}^* = \mathbf{D}_r^{-½}\mathbf{U}$, of columns: $\mathbf{\Gamma}^* = \mathbf{D}_c^{-½}\mathbf{V}$ (12.6)

The above solution has as many principal axes as there are variables (or one less in the case of dummy variables).

There is a similar sequence of calculations to find the principal axes of the unconstrained space. Projection takes place onto the space orthogonal to (i.e., uncorrelated with) the variables in $\mathbf{X}$. This projection matrix is just $\mathbf{I} - \mathbf{Q}$, so the unconstrained (or unrestricted) part of $\mathbf{S}$ is now:

$$\mathbf{S}^\perp = (\mathbf{I} - \mathbf{Q})\mathbf{S} \tag{12.7}$$

(hence $\mathbf{S}$ has been split into two parts: $\mathbf{S} = \mathbf{S}^* + \mathbf{S}^\perp$). The same steps now proceed, where we re-use the same notation $\mathbf{U}$, $\mathbf{D}_\alpha$ and $\mathbf{V}$ for the SVD components, although they are numerically different here, of course:

SVD: $\mathbf{S}^\perp = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T}$ (12.8)

Principal coordinates of rows: $\mathbf{F}^\perp = \mathbf{D}_r^{-½}\mathbf{U}\mathbf{D}_\alpha$, of columns: $\mathbf{G}^\perp = \mathbf{D}_c^{-½}\mathbf{V}\mathbf{D}_\alpha$ (12.9)

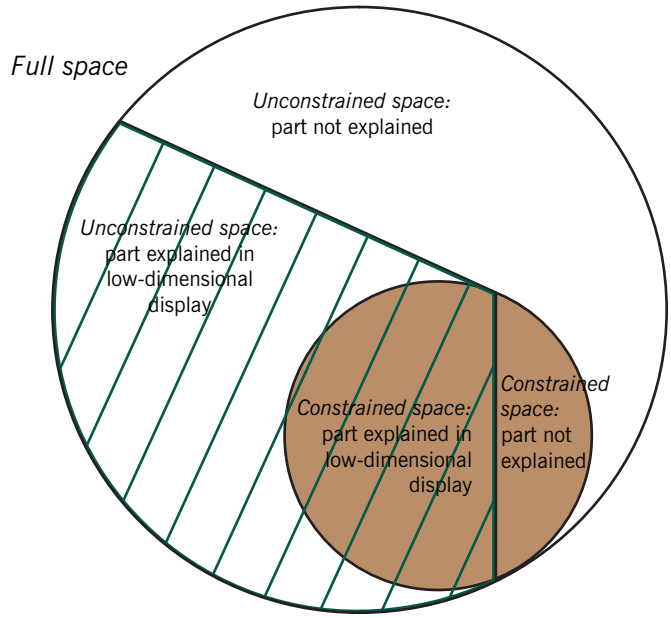
Standard coordinates of rows: $\mathbf{\Phi}^\perp = \mathbf{D}_r^{-½}\mathbf{U}$, of columns: $\mathbf{\Gamma}^\perp = \mathbf{D}_c^{-½}\mathbf{V}$ (12.10)

Constrained LRA is almost identical to the above, starting with the double-centred matrix of log-transformed data, and using the same row and column masses as weights (in Chapter 15 we give the exact formulation). For unweighted LRA, these weights would just be $1/I$ for the rows and $1/J$ for the columns. Similarly for PCA, where the weights are generally equal, implementing the constraints involves starting with the centred (and optionally standardized) matrix, and applying the above steps using weights $r_i = 1/I$ and $c_j = 1/J$. Linearly constrained PCA has also been called *redundancy analysis* in the literature.

Decomposition of variance

When there are two constraining variables, they will still be perfectly explained by the plane of the first two constrained axes, but neither variable will necessarily be identified exactly with a principal axis. For three or more constraining variables the two-dimensional constrained space of representation does not display the constraining variables perfectly. In this case there are two levels of approximation of the data matrix, as depicted in Exhibit 12.3. First the data matrix is split into two parts: the part which is linearly related to the constraining variables, and the part that is not (i.e., $\mathbf{S} = \mathbf{S}^* + \mathbf{S}^\perp$ in the formulation above). Then dimension reduction takes place just as before, but in the constrained space (i.e., the principal

Exhibit 12.3:
The full space decomposition into the constrained space (brown) and unconstrained space (white). Within each space there is a part of the variance (or inertia) that is explained in the respective low-dimensional displays (area with green shading)



axes of $\mathbf{S}^*$ are identified), with constraining variables being displayed in the usual regression biplot style. Dimension reduction can similarly be performed in the unconstrained space by identifying the principal axes of $\mathbf{S}^{\perp}$. This decomposition scheme is illustrated in Exhibit 12.4 for the fish morphology analysis, where the first dimension is constrained by body weight. Since there is only one constraining variable, no dimension reduction is performed in the constrained space. Body weight is represented perfectly on the first dimension, and the second axis of the solution is the optimal first dimension of the unconstrained data space.

Triplots

In a constrained biplot there are three sets of points and the display is called a *triplot*. The third set of points added to the biplot consists of the constraining variables, and they are usually displayed in terms of their regression coefficients with respect to the dimensions of the biplot. Their directions will then be biplot axes onto which the sample points (usually rows) can be projected to give estimates of their values, as before. If the rows have been displayed in standard coordinates, then the constraining variables have directions equal to their correlation coefficients with the axes.

An application to the data set "benthos" illustrates the triplot when there are several explanatory variables. For each site the levels of six variables were measured: total hydrocarbon content (THC), total organic material (TOM), barium (Ba),

Exhibit 12.4:
The decomposition of variance (or inertia), first into the one-dimensional constrained space of body weight and the unconstrained space uncorrelated with body weight. The constrained space forms the first dimension of the biplot, which is only 4.0% of the total variance, and the first dimension of the unconstrained space forms the second dimension of the biplot, explaining 20.7% of the total variance

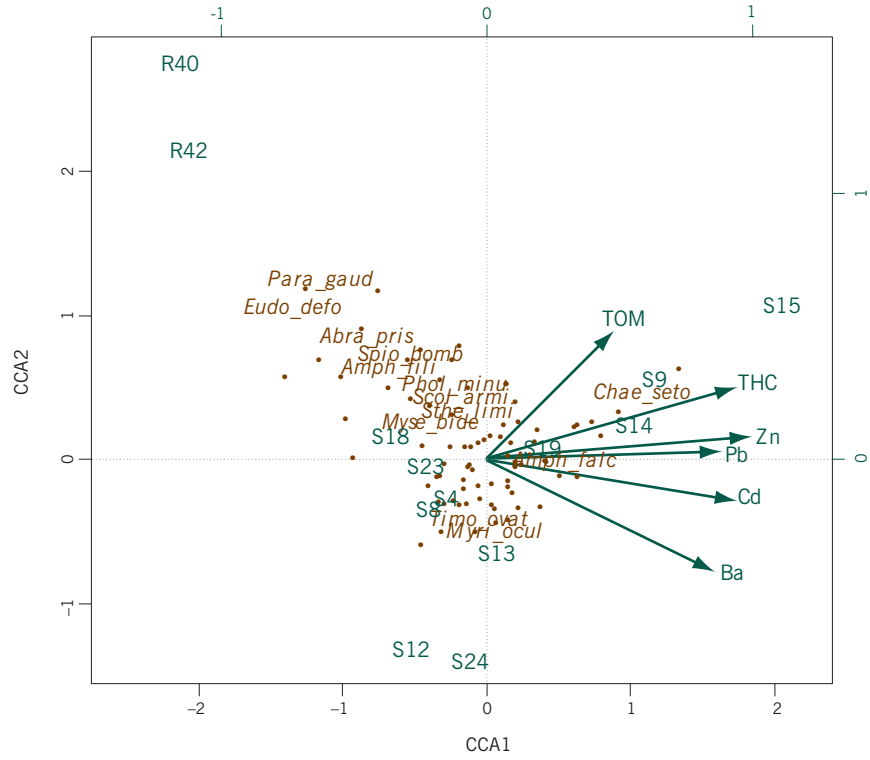
cadmium (Cd), lead (Pb) and zinc (Zn). It was preferable to log-transform these variables because of some very large values. Exhibit 12.5 shows the resultant triplot of the CCA restricted to the space of these six explanatory variables (in other words, dimension reduction has been performed from a six-dimensional space to a two-dimensional one). The sites in the triplot are in standard coordinates, and the species are at weighted averages of the sites. The explanatory variables are shown as vectors with coordinates equal to their regression coefficients on the axes (notice the different scale for these vectors). The reference stations are much more separated from the polluted stations now that the solution is constrained by variables that essentially measure pollution. Barium appears to be the variable that lines up the most with the separation of the reference stations from the others, pointing directly away from the unpolluted reference stations. The variable least associated with the unpolluted versus polluted contrast appears to be total organic material.

The corresponding decomposition of inertia is shown in Exhibit 12.6, showing that the six explanatory variables explain 65% of the total inertia in the data. Of this, 72.5% is explained by the two dimensions of the triplot. Because the explanatory variables are displayed using standardized regression coefficients, their lengths are related to how much of their variance is explained by the axes: the least is TOM (46% variance explained), and the most is Zn (96% variance explained).

Stepwise entry of the explanatory variables

As in multiple regression, the more explanatory variables that enter, the more variance is explained. When the number of explanatory variables equals the number of cases 100% of the variance would be explained and then there is effective-

Exhibit 12.5:
Triplot of the "benthos" data, showing the six constraining variables. Of the total inertia (0.7826) of the species abundance data, 65% is in the constrained space, of which 72.5% is displayed in the triplot



ly no constraint on the data and the analysis would be a regular biplot. To reduce the number of explanatory variables in such an analysis, a stepwise entry of explanatory variables is often performed, which ensures that only variables that explain a significant part of the variance are entered. At each step the variable that explains the most additional variance is entered and this additional variance is tested using a permutation test. The process continues until no variables entering produce a significant increase in explained variance. This procedure is illustrated in the case study of Chapter 15.

SUMMARY: Constrained Biplots and Triplots

1. Biplots, whether they are based on PCA, CA or LRA, display the data in a reduced dimensional space, usually a plane, with the objective of approximating the original data as closely as possible.
2. Often the data matrix can be regarded as responses to be explained by some explanatory variables that are available. The original biplot dimensions are not necessarily related to these explanatory variables, but an alternative approach constrains the principal axes of the biplot to be specifically related to these variables.

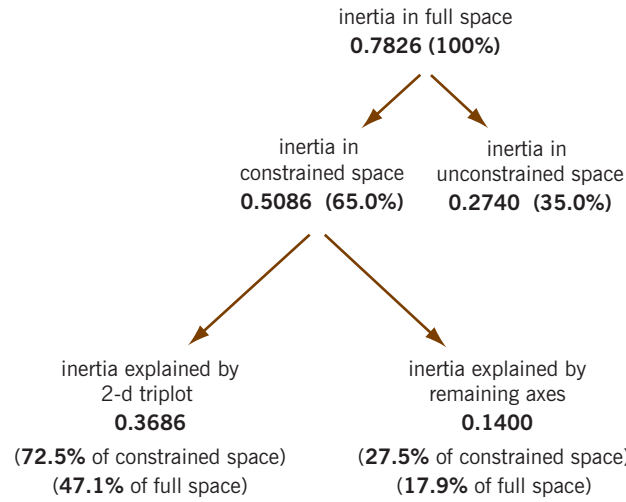


Exhibit 12.6:
The decomposition of inertia, first into the six-dimensional constrained space of the explanatory environmental variables and the unconstrained residual space that is uncorrelated with the explanatory variables. In the constrained space the first two dimensions explain 72.5% of the constrained inertia, which is 47.1% of the original grand total

3. The constraint is usually a linear one: the data are projected first into the constrained space which is linearly correlated with the explanatory variables, and then dimension reduction takes place as before.
4. The result of such an analysis with constraints is a triplot, showing the rows and columns of the original data matrix of interest, plus vectors indicating directions for the explanatory variables.
5. The dimensions of the residual, or unconstrained space, may also be of interest. In this space variance or inertia is explained in biplots that are uncorrelated with the explanatory variables.
6. The initial total variance or inertia of the data matrix is decomposed first into a constrained part (linearly related to the explanatory variables) and a residual unconstrained part (uncorrelated with the explanatory variables). Biplots can also be constructed for the unconstrained part of the data.
7. Explanatory variables are often entered stepwise, where the entering variable is the one that explains the most additional variance in the data, and this added variance can be tested for statistical significance.
8. For a single categorical variable as an explanatory variable, where the categories are coded as dummy variables, the constrained analysis is equivalent to a discriminant analysis between the categories.

CHAPTER 13

Case Study 1: Comparing Cancer Types According to Gene Expression Arrays

This first case study contains many aspects of biplots treated in this book. The context is a large data set of microarray data from tumour samples found in children. This is a very "wide" data set in the sense that there are only 63 samples but over 2000 variables in the form of genes expressed in the microarray experiments. The variables are on the same continuous scale and so the regular PCA biplot of Chapter 6 will be used to visualize the raw data. But because the samples are grouped we shall also apply the centroid biplot described in Chapter 11 to show separation of the tumour groups. There are two additional aspects to this case study. First, because of the large number of variables we will be interested in quantifying the contributions of each one to the biplots that we construct, with a view to reducing the gene set to the most important ones. Second, an additional sample of 20 tumours is available, which can be used to test whether the biplot provides a prediction rule capable of classifying these additional tumours correctly.

Contents

Data set "cancer"

This data set "cancer" is taken from the book *The Elements of Statistical Learning (second edition)* by Hastie, Tibshirani and Friedman and consists of a matrix of 2308 genes (columns) observed on 63 samples (rows)—see the Bibliography for a link to the book's website and accompanying data sets. The data arise from microarray experiments, a technology which has become important in genomic research,

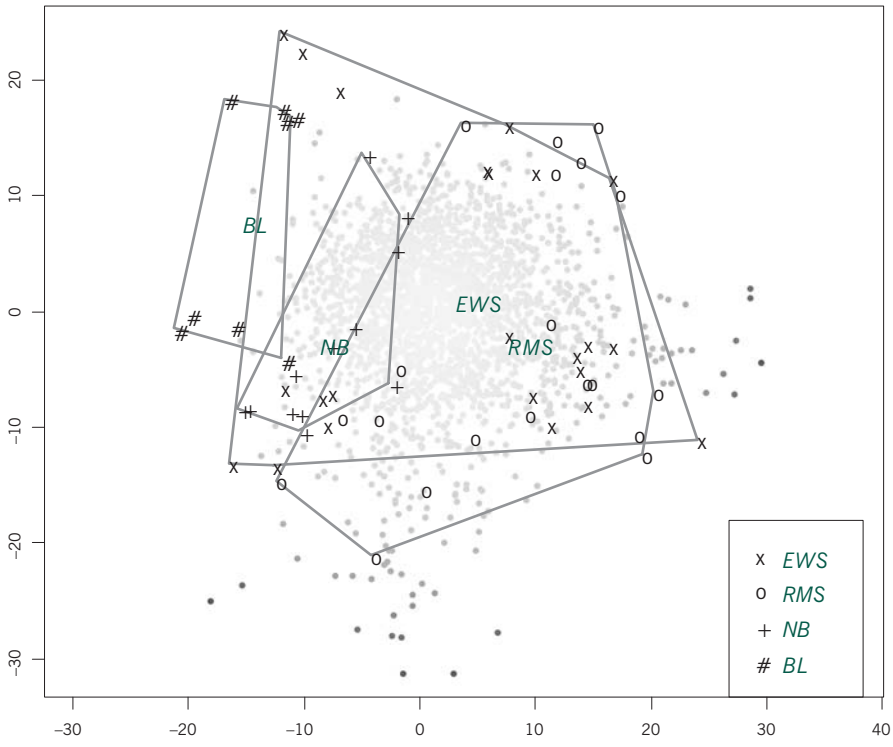
especially the relation of genes to various diseases. The samples are from small, round blue-cell tumours found in children. The genes are quantified by their expression values, the logarithm of the ratio R/G, where R is the amount of gene-specific RNA in the target sample that hybridizes to a particular (gene-specific) spot on the microarray, and G is the corresponding amount of RNA from a reference sample. The data set is called "wide" because of the large number of variables compared to the samples. The tumours fall into four major types: *EWS* (Ewing's sarcoma), *RMS* (rhabdomyosarcoma) *NB* (neuroblastoma) and *BL* (Burkitt lymphoma)—in this data set of 63 samples there are 23 *EWS*, 20 *RMS*, 12 *NB* and 8 *BL* tumours. There is an additional data set of 20 samples from these four cancer types, which we will use later in the case study to test a classification rule predicting cancer type.

Principal component biplot

The basic data are all on a logarithmic scale and do not require further standardization.

Notice that these logarithms of ratios are not log-ratios in the sense of Chapter 7, where the ratios are formed from all pairs of a set of observed variables. Because there are 2308 variables we will not use arrows to depict each one, but grey dots

Exhibit 13.1:
PCA contribution biplot of the data set "cancer", showing convex hulls around the four groups and labels at their centroids. Grey dots indicate the 2308 genes



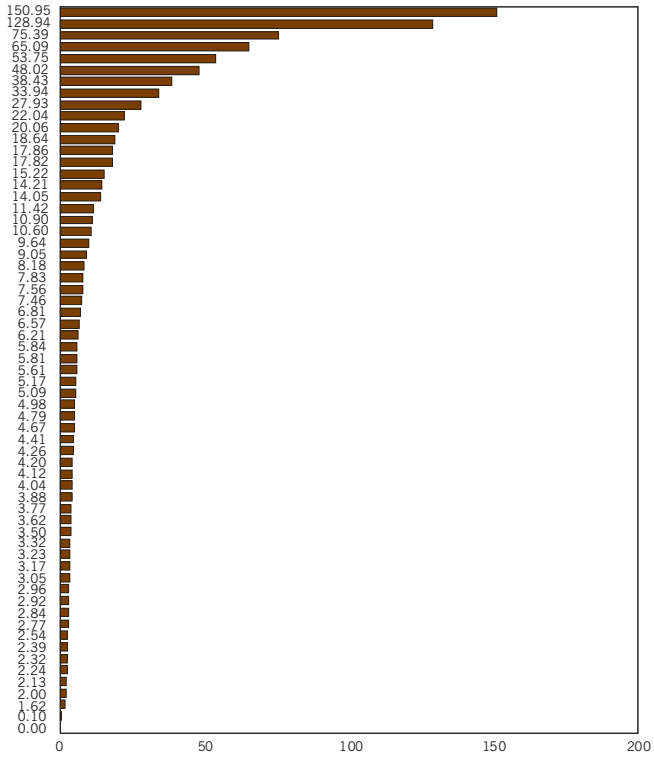


Exhibit 13.2:
Scree plot of the 63 eigenvalues in the PCA of the data set "cancer", showing the last one equal to 0 (there are 62 dimensions in this "wide" data set)

on a grey scale, where the darkness of the point is related to the gene's contribution to the solution—see Exhibit 13.1. Because this is a PCA biplot with no differential weights on the variables, the highly contributing genes will also be those far from the centre of the display.

The PCA biplot does not separate the cancer types very well, as seen by the large overlap of the four groups. Of course, this is not the objective of the PCA, which aims to maximize the between-sample dispersion, not the between-group dispersion. This sample-level biplot gives a first idea of how the samples lie with respect to one another and is useful for diagnosing unusual samples or variables, as well as spotting possible errors in the data. The dimensionality of this 63 × 2308 matrix is 62, determined by the number of samples minus 1 in this "wide" case rather than the number of variables. The percentage of variance accounted for by the two-dimensional solution is 28.5%. It is useful to look at the *scree plot* of the eigenvalues to try to assess the amount of noise in the data (Exhibit 13.2). The total variance in this data set is equal to 982.0, with an average per dimension of 982.0/62 = 15.8. By this criterion the first 14 dimensions are above average, although it is clear that the first two do separate clearly from the rest.

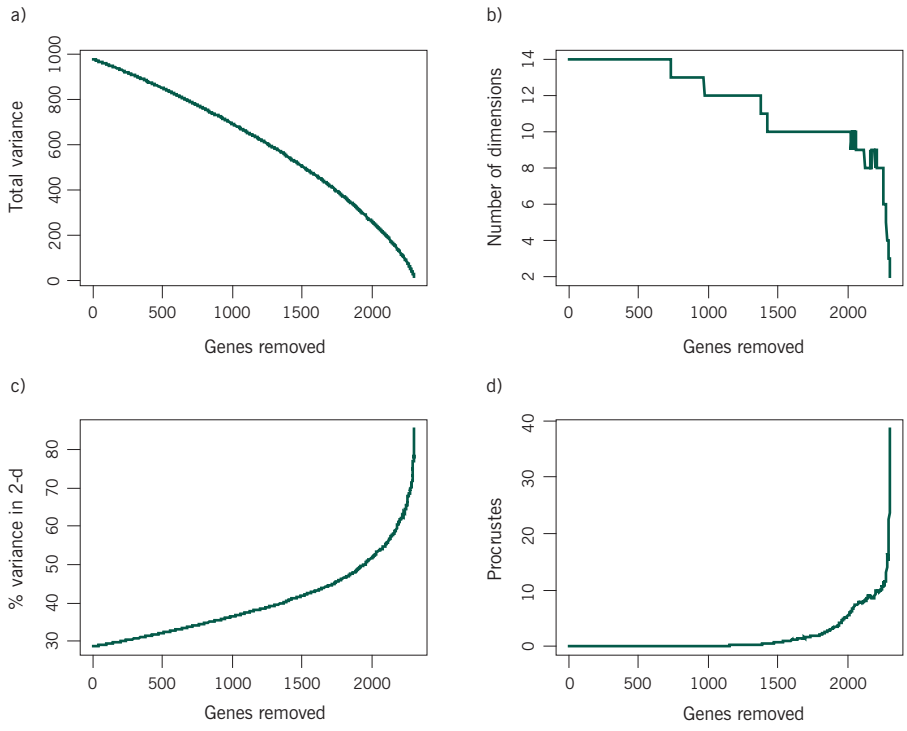
Reducing the number of variables

We have several tools at our disposition now to reduce the number of variables (genes) while keeping track of the effect this has on the visualization of the cancer samples. A possible strategy is to reduce the gene set one at a time, removing each time the gene that contributes the least to the solution. At each stage of the gene removal we measure the following aspects, shown in Exhibit 13.3:

a. The total variance and the average over the dimensions (the latter will be the former divided by 62 until the number of genes reduces below 62, in which case the dimensionality is determined by the number of genes).
b. The number of dimensions that are above the average.
c. The percentage of variance explained by the two-dimensional solution.
d. The Procrustes statistic on the configuration of sample points, compared to the initial solution (Exhibit 13.1)—this will quantify how much the configuration is changing.

Total variance (Exhibit 13.3a) obviously decreases as genes are removed—the decrease is less at the start of the process when the genes of very minor contribution to the solution are removed. The number of dimensions greater than the average also decreases (Exhibit 13.3b) but still remains fairly high until the end of the re-

Exhibit 13.3: *Monitoring of four statistics as the number of removed genes increases*



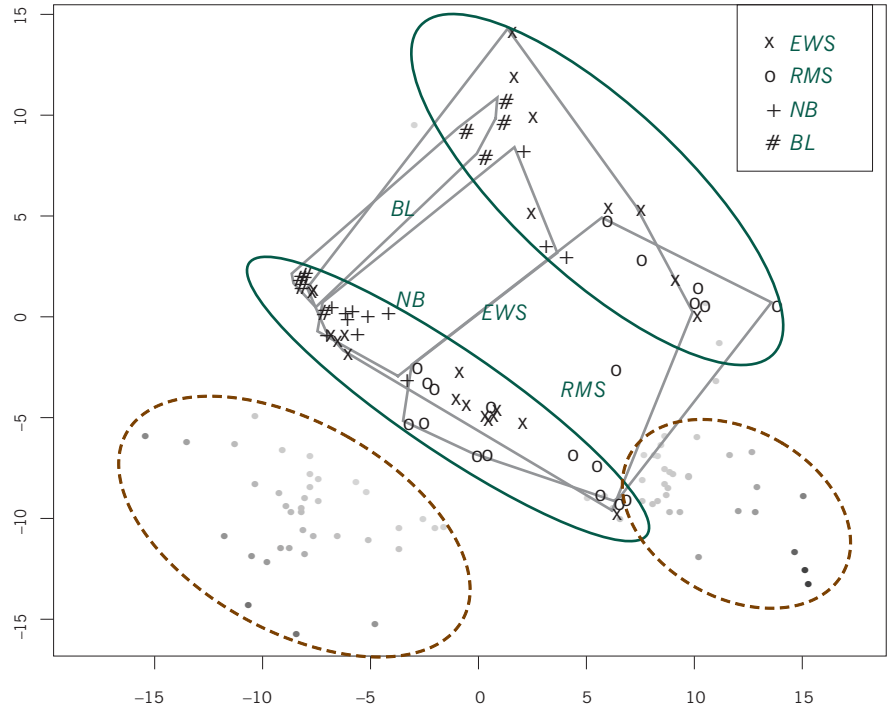


Exhibit 13.4:
PCA biplot of the reduced gene set (75 high-contributing genes, that is 2233 genes omitted), showing one set of genes (in dashed ellipse) at bottom right separating the group centroids (indicated by the labels) and another group at bottom left that is separating the total sample into two distinct groups (shown in the green ellipses), independent of their cancer types

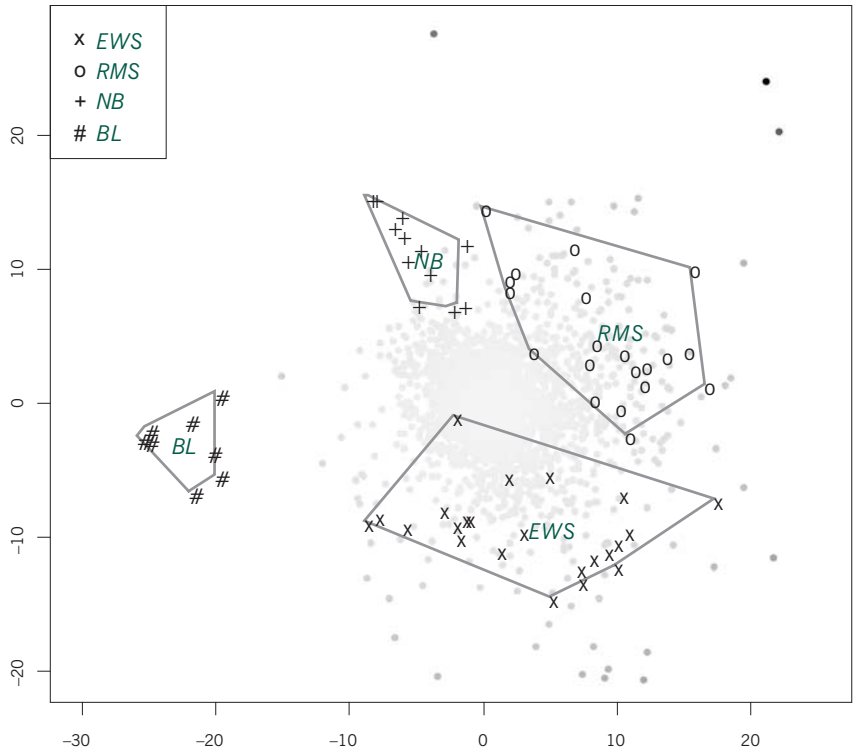
moval process. The percentage of variance on the first two axes increases as the "noisy" part of the data is removed (Exhibit 13.3c). According to the Procrustes analysis (Exhibit 13.3d) the two-dimensional configuration remains almost the same even when as many as 1500 genes are removed.

We chose a solution when the Procrustes statistic reached 10%, when 2233 genes were removed, leaving only 75 included in the PCA. Notice the gradual change in the Procrustes statistic (Exhibit 13.3d) up to this point, then a relative stability in the configuration at about 10% followed by more dramatic changes. Exhibit 13.4 shows the biplot with the reduced set of genes. The spread of the four groups, from *BL* to *RMS*, is retained (see Exhibit 13.1), just slightly rotated. What is evident here is the emergence of two groups of genes, one at bottom right which is responsible for the separation of the tumour groups, and another group at bottom left which separates the samples into two clear clusters independent of their groups—the only exception is an *RMS* tumour suspended between the two clusters.

Centroid biplot —all variables

In order to see the separation of the tumour groups better and to identify which genes are determining the difference between them, a biplot of the group centroids can be performed, as described in Chapter 11 on discriminant

Exhibit 13.5:
Centroid biplot of the four tumour groups, using all 2308 variables. The percentage of centroid variance displayed is 75.6%, with between-group variance in the plane 88.6% of the total



analysis (DA) biplots. Because there are four centroids, the space they occupy is three-dimensional; hence the planar display involves the loss of only one dimension. Exhibit 13.5 shows the centroid (or DA) biplot based on all 2308 genes.

The tumour groups are now very well separated, and the separation of the clusters observed in Exhibit 13.4 is no longer present. Of the total variance of the centroids in their full three-dimensional space, 75.6% is represented in the biplot. Of the total variance of the 63 samples represented in this two-dimensional biplot, 88.6% is between-group variance and 11.4% within-group variance.

Centroid biplot—reduced set of variables

Again, we are interested in reducing the number of genes to see which are the most determinant in separating the groups. By applying the same step-by-step reduction in the number of genes, always removing the gene with the least contribution to the group differentiation at each step, and by monitoring the percentage of variance displayed in the two-dimensional map as well as the proportion of total planar variance accounted for by the between-group part. It turns out that a

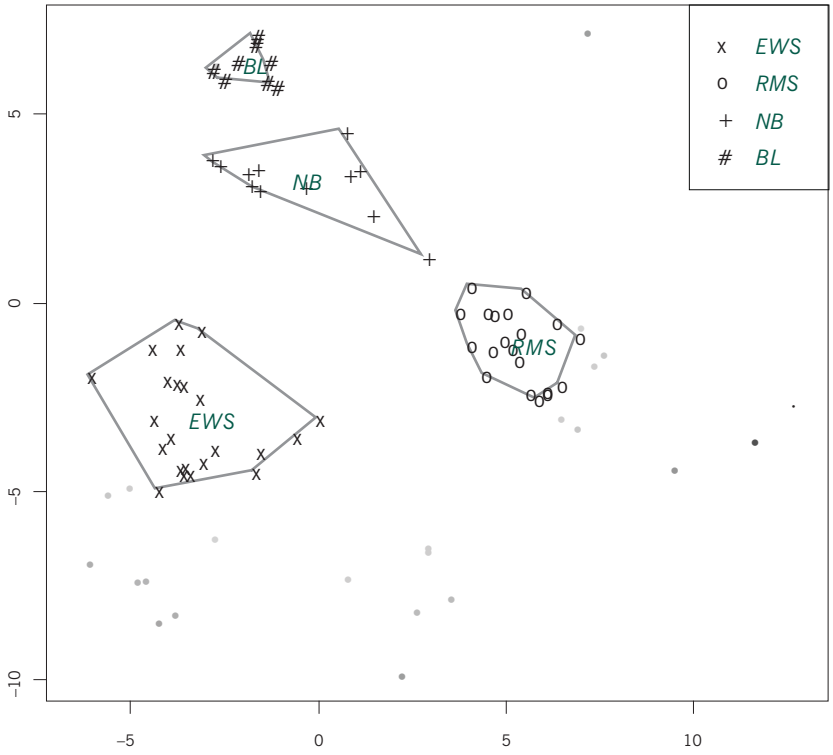


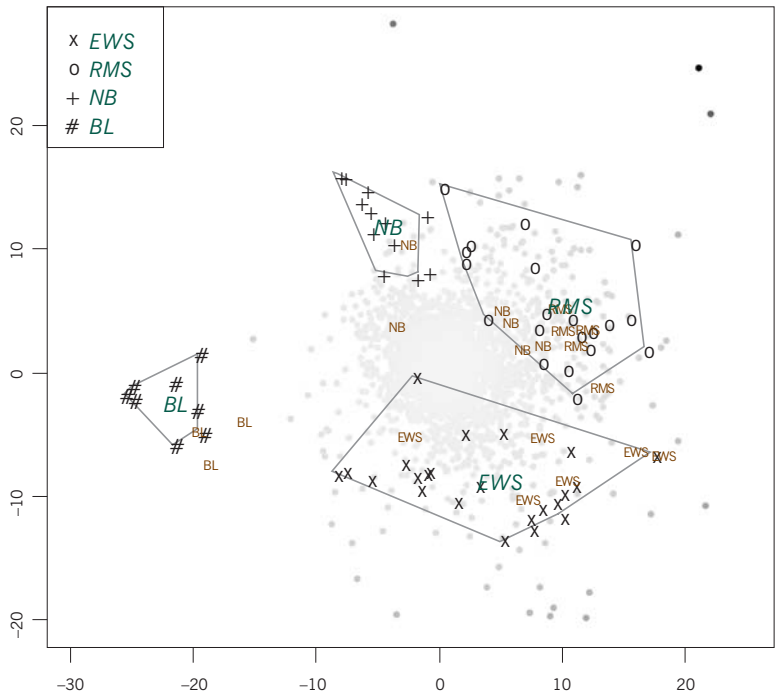
Exhibit 13.6:
Centroid biplot of the four tumour groups, using 24 highest contributing variables after stepwise removal. The percentage of centroid variance displayed is 94.9%, with between-group variance in the plane 90.5% of the total

maximum of the latter percentage is reached when we have reduced the gene set to 24 genes, for which the solution is shown in Exhibit 13.6. The between-group variance in the plane is 90.5% of the total, about 3 percentage points better than Exhibit 13.5. There is only 5.1% of the centroid variance in the third dimension now, as opposed to 24.4% in Exhibit 13.5.

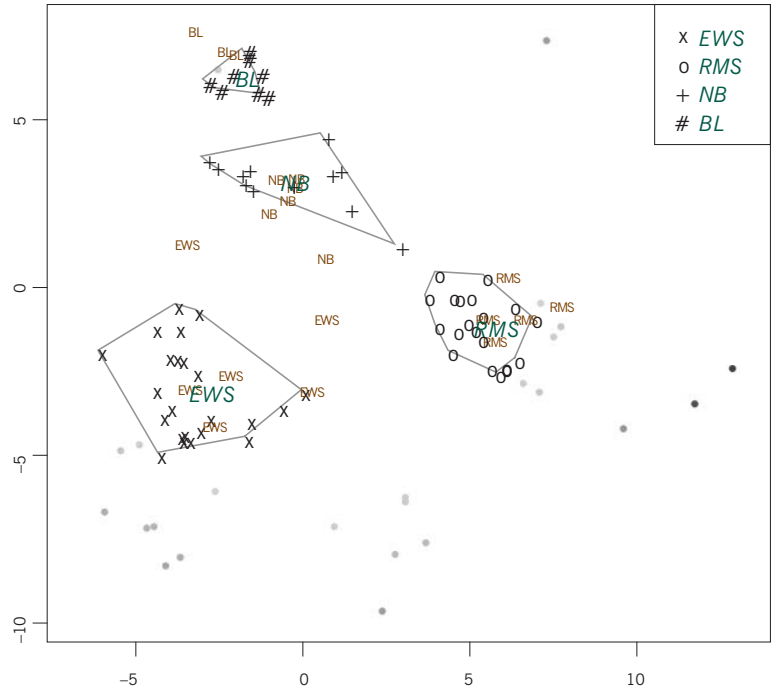
In Exhibit 13.6 we have thus achieved an optimal separation of the groups, while also reducing the residual variance in the centroids that is in the third dimension. Notice the lining up of the three tumour groups *BL*, *NB* and *RMS* from top left to bottom right, coinciding with the genes extending to the bottom right hand side: it will be these genes that distinguish these three groups, with increasing values from *BL* to *NB* to *RMS*. On the other hand the group *EWS* is situated at bottom left associated with high values of the group of genes at bottom left, and low values of the single gene that one finds at top right. There is a group of six genes at the bottom of the display that are separated from the group at bottom left, which no doubt not only separate *EWS* from the other groups but also contribute slightly to the left-to-right separation of the other three groups.

Exhibit 13.7:
The 20 additional tumours in the centroid solution space for all 2308 genes (upper biplot), and the reduced set of 24 genes (lower biplot)

a)



b)



Classification of additional samples

In addition to the 63 samples studied up to now, an additional sample of 20 tumours was available, and the type of tumour was known in each case. We can use our results in Exhibits 13.5 and 13.6 above to see whether accurate predictions of the tumour types are achieved in this test data set. We do this in a very simple way, just by situating the tumours in the two-dimensional solution space and computing their distances to the group centroids, and then predicting the tumour type by the closest centroid. Exhibit 13.7 shows the new tumours in the solution of Exhibit 13.5 using all 2308 genes (upper biplot) and then in the solution of Exhibit 13.6 using the reduced set of 24 genes (lower biplot). It is clear that in the upper biplot that four of the six *NB* tumours will be misclassified as *RMS*. In the lower biplot, the new tumours generally lie closer to their corresponding centroids, with just two *EWS* tumours being misclassified as *NB*, the one on the right being only a tiny bit closer (in the third significant digit) to the *NB* centroid than to the *EWS* one. It is a general principle that the elimination of irrelevant variables can improve the predictive value of the solution, and this is well illustrated here.

Improving prediction

As a final remark, it is possible to improve the predictive quality of this centroid classification procedure in two ways. First, there is some additional variance in the centroids in the third dimension, which we have ignored here. Calculating tumour-to-centroid distances in the full three-dimensional space of the four centroids will improve the classification. Second, in the area known as *statistical learning*, a branch of machine learning, the small subset of genes used to define the predictor space would be chosen in a more sophisticated way, using *cross-validation*. This involves dividing the *training set* of data (that is, our initial sample of 63 tumours) into 10 random groups, say, and then using 9 out of the 10 groups to determine the subset of variables that best predicts the omitted group, and then repeating this process omitting each of the other groups one at a time. There would thus be 10 ways of predicting new observations, which we would apply in turn to the *test set* (the 20 additional tumours), obtaining 10 predictions for each new tumour from which the final prediction is made by majority "vote". If these two additional improvements are implemented in our procedure it turns out that we can predict the group membership of all 20 tumours exactly.

SUMMARY

We have shown how biplots based on principal component analysis of both individual-level and aggregate-level data can be used to identify natural groups of observations in a large data set as well as distinguish between existing known groups. With respect to this data set which has a huge number of variables compared to observations:

1. In both the individual- and aggregate-level analyses, it is useful to reduce the number of variables to a smaller set that is the most determinant in showing

respectively (i) the patterns in the individual-level data, and (ii) the separation of the known groups.

2. One way of eliminating variables is to calculate each variable's contribution to the solution (a planar biplot in our application). The variable with the least contribution is eliminated, and the procedure is repeated over and over again until a small subset is found.

3. We decided to stop the variable elimination process in the individual-level analysis when the Procrustes statistic rose to 10%—this was an *ad hoc* decision, but was based on observing the evolution of the Procrustes statistic as variables were eliminated. This statistic increased very slightly and slowly up to this point, but reducing the variables beyond this stage the solution started to change dramatically

4. In the case of the aggregate-level analysis, we monitored the ratio of between-group variance to total variance in the low-dimensional solution as variables were eliminated, and stopped when this reached a maximum.

5. In the centroid analysis, the eventual space based on the smaller set of variables can be used to classify new observations, by calculating their distances in the solution to the centroids and then choosing the centroid that is closest as the group prediction.

CHAPTER 14

Case Study 2: Positioning the "Middle" Category in Survey Research

Offering a middle alternative, for example "neither agree nor disagree", as a response on an attitudinal scale can have various consequences in survey research. In this case study, after a general investigation of a large survey data set, special attention is given to the middle categories, specifically how they associate (a) with one another, (b) with the "adjacent" categories between which they are supposed to lie and (c) with demographic characteristics. Missing responses enter our study in a natural way as additional non-substantive responses and we can also see how these are interrelated and related in turn to the substantive responses. This approach is illustrated with the ISSP data on attitudes to women working (this is an expanded version of the data set "women" used in Chapter 9).

Contents

Data set "womenALL"

The data set "women" was introduced in Chapter 9, consisting of eight questions eliciting attitudes about working women. In that chapter, to simplify the explanation, data from only one country, Spain, were considered and all respondents with some missing data were eliminated. In this case study all the data from 46 638 respondents in 32 countries are considered, and no respondents are eliminated —this data set is referred to as "womenALL". Exhibit 14.1 lists the countries surveyed and the abbreviations used in the graphics.

Exhibit 14.1:
Countries surveyed in the third Family and Changing Gender Roles survey of the ISSP in 2002 (former West and East Germany are still sampled separately for research purposes). The abbreviations are used in subsequent biplots

AU	Australia	SE	Sweden	SK	Slovakia
DW	Germany (west)	CZ	Czech Republic	CY	Cyprus
DE	Germany (east)	SI	Slovenia	PT	Portugal
GB	Great Britain	PL	Poland	RC	China
NI	Northern Ireland	BG	Bulgaria	DK	Denmark
AT	Austria	NZ	New Zealand	CH	Switzerland
US	USA	RP	Philippines	FL	Belgium (Flanders)
HU	Hungary	IL	Israel	BR	Brazil
IE	Ireland	JP	Japan	SF	Finland
NL	Netherlands	ES	Spain	TW	Taiwan
NW	Norway	LV	Latvia		

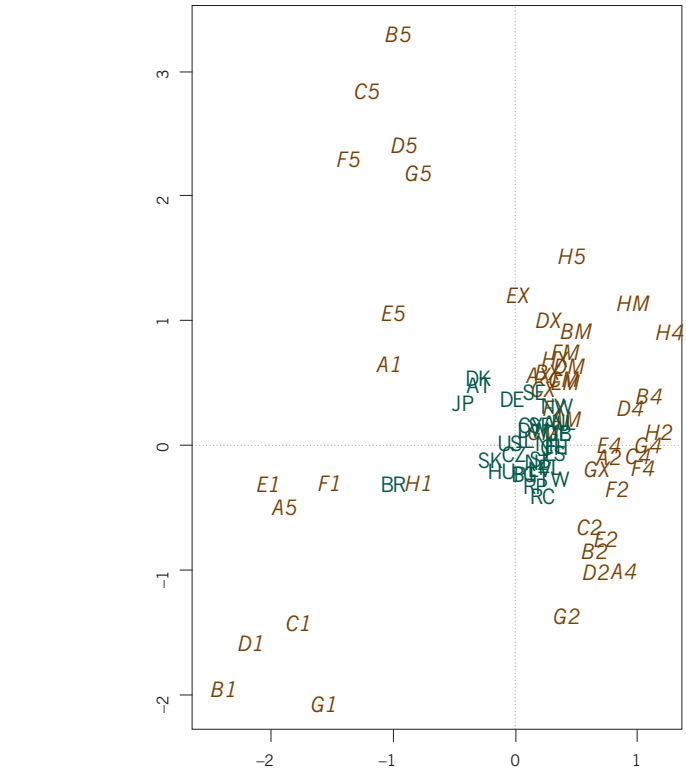
Cross-national comparison using CA biplot of concatenated tables

To get a broad overview of the differences between countries, a concatenated matrix (see Chapter 9) is assembled, cross-tabulating the countries with each of the eight questions.

Since each question has five substantive response categories plus a missing category, there are six categories per question, and the concatenated matrix has 32 rows and 48 columns. The CA asymmetric map/biplot is shown in Exhibit 14.2, with a separate amplification of the row points, which as usual are bunched up near the middle of the biplot. The missing categories of response, labelled *AX* to *HX*, are all in a group near the origin of the biplot. So too are the middle response categories (the category 3's), which we have labelled *AM* to *HM* here. Generally all the extreme response categories (1's and 5's) are to the left of centre, while the moderate response categories (2's and 4's) are to the right. The conservative-to-liberal attitude scale runs from bottom to top, with strong agreement to statements B, C, D and G at bottom left (see chapter 9 for the statement wording). Brazil is in an isolated position, showing that its respondents tend to use the extreme conservative response categories. China is also at the conservative extreme of these countries, but using more of the moderate response categories. At the top Denmark is at the most liberal position, followed by Austria and Sweden, with the Swedish using the more moderate responses.

Multiple correspondence analysis of respondent-level data

This aggregate-level picture of the countries does not reflect associations between response categories at the level of the individual respondent. Applying MCA to the original 46,638 × 8 matrix gives the biplot in Exhibit 14.3. The result is typical of social-science applications, with all the non-substantive missing response categories separating out (bottom left) and opposing the substantive responses which themselves split into the moderate and middle categories (upper left) and the extreme categories (upper right). The fact that these three types of response



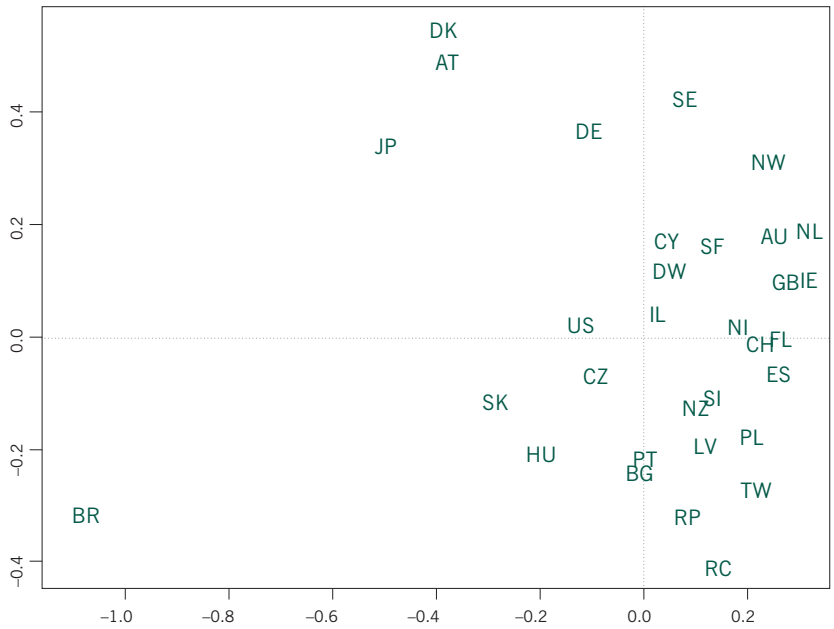
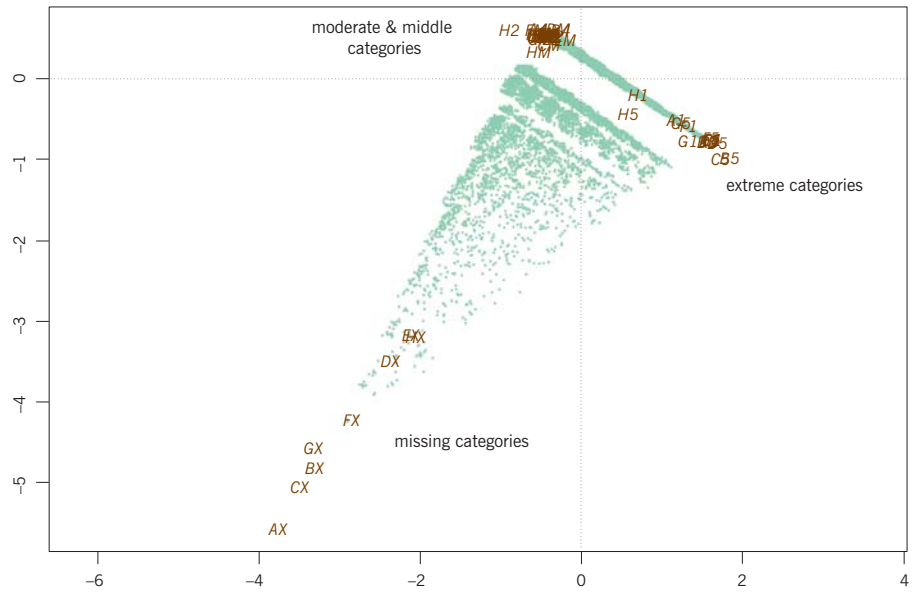


Exhibit 14.2:
CA biplot of the concatenated countries by categories matrix, and a separate plot of the countries alone

Exhibit 14.3:
MCA biplot of the respondent-level data: each dot represents one of the 46,638 respondents at the average position of his or her eight response categories



(missing, moderate and extreme) group together means that the categories within each group associate strongly: for example, someone who gives a missing response on one question is likely to give more missing responses on others, and similarly for the moderate and extreme responses. The respective groupings of the extreme and moderate responses, irrespective of the wording of the question, are stronger than the underlying attitude scale. Each respondent is displayed by a dot at the average position of his or her eight response categories. Thus the thick band of points at the top correspond to those with no missing values, followed by bands of respondents with increasing numbers of missing values as we move towards the bottom left.

Subset multiple correspondence analysis eliminating missing categories

Since we are more interested in the substantive responses, we can eliminate the strong effect of the missing categories by performing a so-called *subset analysis*. This is variant of correspondence analysis which visualizes a subset of points while maintaining the original geometry of the complete set: the centre, masses and metric of the space remain the same, what changes is the elimination of certain points from the dimension reduction process (this is not the same as declaring the missing categories as supplementary points, which would change the geometry of the active data set). Exhibit 14.4 shows the result of omitting the missing categories in the subset analysis—as in Exhibit 14.2, the first (horizontal) axis opposes extreme response categories on the left and moderate categories on the right. But now the attitude scale itself appears vertically with conservative response categories lower down and liberal responses at the top. Exceptions are

questions *F* ("work is best for a woman's independence") and *H* ("working women should get paid maternity leave"), where opinions appear unrelated to the general scale of attitude towards whether women should work or not.

In Exhibit 14.4 the middle ("*M*") categories appear grouped between the moderate ones, as one might expect. If we bring in the third dimension of the subset analysis, however, these categories are seen to separate as a group, and do not lie between the "*2*"s and the "*4*"s. Exhibit 14.5 shows the planar view of the third dimension (horizontal) and second dimension (vertical), and the "*M*"s no longer lie in their expected positions on the scale. For question responses that are consistent with an underlying ordinal attitudinal scale the MCA configuration should take the approximate form of polynomials of increasing order, as shown in Exhibit 14.6 (usually the scale appears on the first dimension, and the polynomials are with respect to the first dimension—in this example, the scale is found on the second dimension because of the strong extreme versus moderate response effect). Exhibit 14.4 fits the pattern on the left in Exhibit 14.6, while all the response categories except the middle ones in Exhibit 14.5 fit the pattern on the

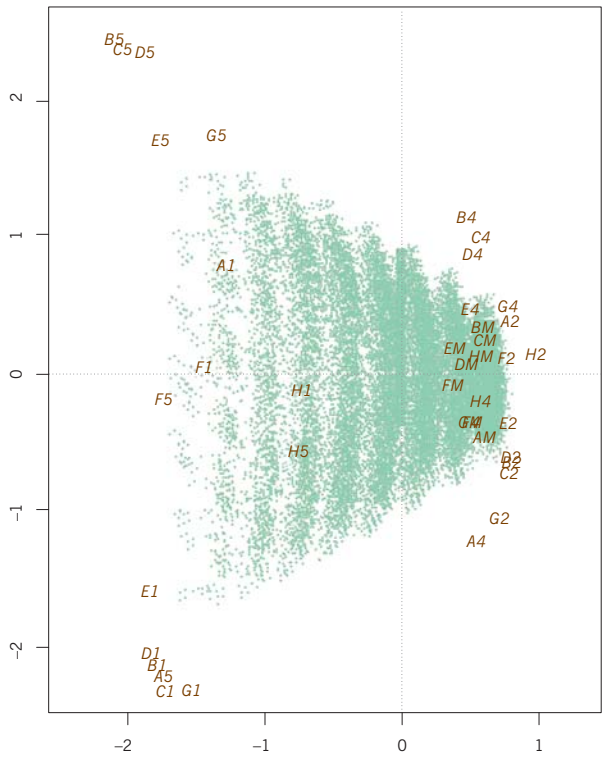


Exhibit 14.4:
Subset MCA biplot of the respondent-level data: each dot represents one of the 46,638 respondents at the average position of his or her eight response categories

Exhibit 14.5:
Subset MCA biplot of the respondent-level data, showing dimension 2 vertically, as in Exhibit 14.4, but dimension 3 horizontally. The separation of the middle categories (encircled) is now apparent

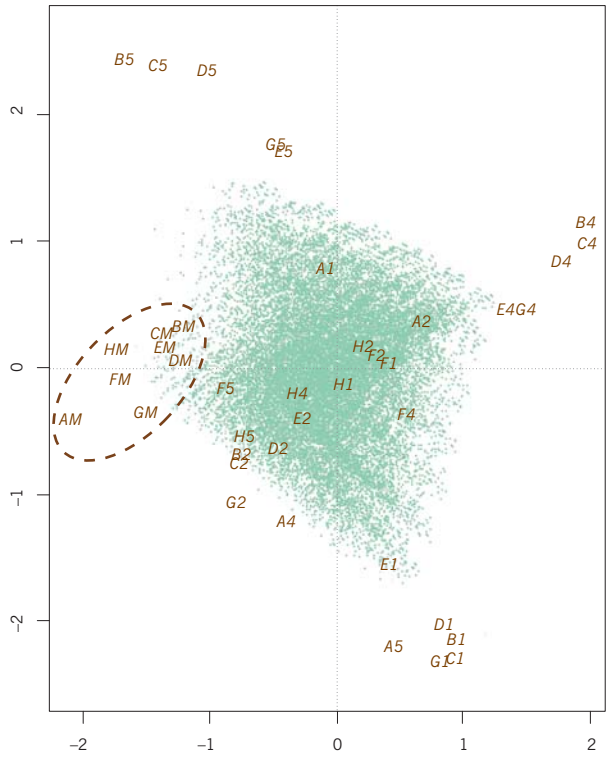
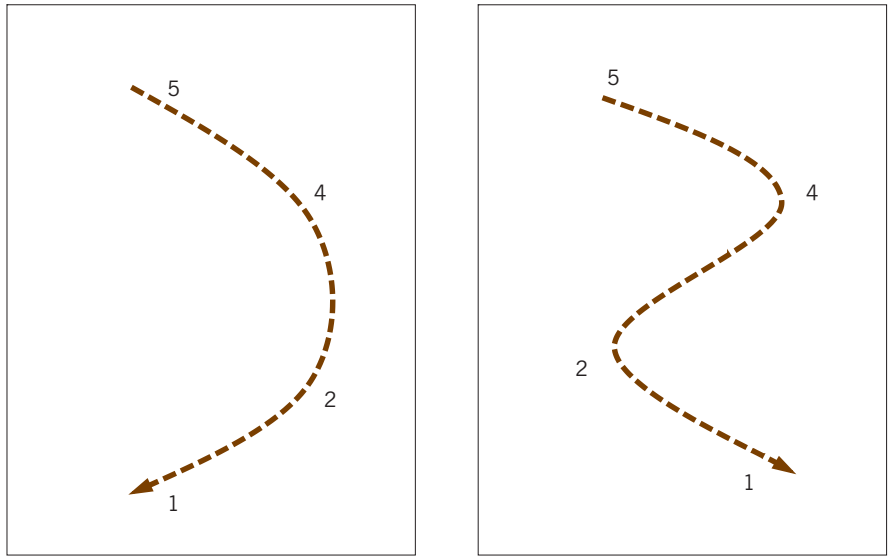


Exhibit 14.6:
General patterns in Exhibits 14.4 and 14.5 (for questions B, C, D and G, for example, all worded negatively towards working women), showing their respective quadratic and cubic patterns



right in Exhibit 14.6. What we are discovering is that the middle response categories are partially fitting into the attitudinal scale but show a distinct separation as a different response type, which might be a type of non-response or so-called "satisficing" effect, which is the respondents' way of giving an acceptable answer without having to spend time and effort forming an opinion.

The dimensions of "middleness"

If we look at the numerical diagnostics of the subset MCA, it turns out—not unexpectedly—that the middle response categories are somewhat correlated with dimension 3 (as seen in Exhibit 14.5), but these categories will have parts of their variance on many other dimensions too. If we really want to study the effect of the middle response, we need to isolate exactly where the middle categories are. Then we can see if any demographic characteristic is linked to those dimensions. There are two ways we can do this: using subset analysis again, but just on the middle categories, or using canonical correspondence analysis (CCA), where we define "middleness" as an explanatory variable. A difference between the two approaches will be that the subset analysis will focus on all the dimensions of "middleness", of which there are 8 since there are 8 questions, so that dimension reduction will be necessary, whereas the way we implement the CCA just one dimension of "middleness" will be imposed as a constraining variable on the solution, which makes it easier to relate to the demographics.

Canonical correspondence analysis to focus on middles and missings

To implement the CCA we create the explanatory variable "number of middle responses" by counting, for each respondent, how many middle responses are in his or her response set. This variable can vary from 0 to 8. Since this is the addition of the 8 columns of the 46,638 × 48 indicator matrix, the variable we are creating is actually the centroid of the 8 middle response points. It is this centroid on which the CCA will focus. In fact, we can do exactly the same for the missing values: instead of performing a subset MCA, we can add a variable "number of missing responses" and then use both of these as explanatory variables, thus giving a two-dimensional restricted space of middles and missings. This will allow a convenient investigation of possible associations with the demographics. The set-up for what amounts to a canonical MCA is shown in Exhibit 14.7.

The canonical MCA with the two constraining variables is shown in Exhibit 14.8. The missing count variable is almost exactly aligned with the horizontal first dimension, and the middle count variable slightly more than 90 degrees away in a vertical direction. The 46,638 respondents occur in only $1 + 2 + \cdots + 9 = 45$ different combinations of the two constraining variables, which are indicated by circles with an area proportional to the corresponding number of respondents. Thus the largest circle at bottom left corresponds to 0 middles and 0 missings (3107 respondents), and the next circle vertically corresponds to 1 middle and 0 missings (2354 respondents) up to the topmost circle for all 8 middles (254 respondents).

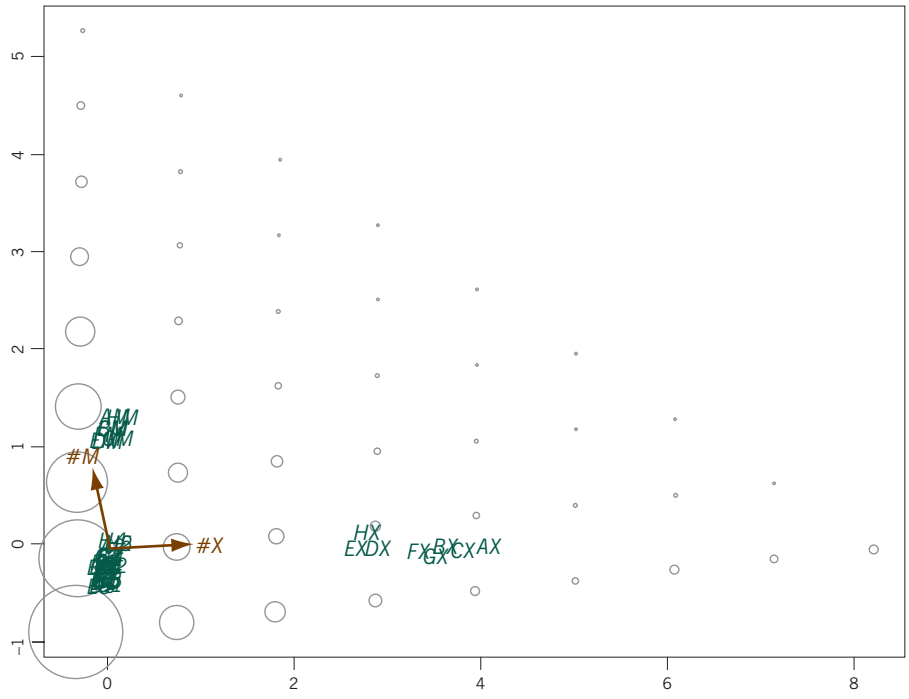
Exhibit 14.7: *Data set-up for canonical MCA biplot, showing first 10 rows of the original data on the left and recoded data on the right used for the analysis. The columns #M and #X are the sums of the M and X columns of the indicator matrix, i.e. the counts of middle and missing responses respectively*

A	B	C	D	E	F	G	H	A1	A2	AM	A4	A5	AX	B1	B2	BM	B4	B5	BX	· · ·	#M	#X
4	2	2	3	3	2	3	4	0	0	0	1	0	0	0	1	0	0	0	0	· · ·	3	0
1	5	5	5	1	1	5	2	1	0	0	0	0	0	0	0	0	0	1	0	· · ·	0	0
2	3	2	3	2	2	4	2	0	1	0	0	0	0	0	0	1	0	0	0	· · ·	2	0
4	2	1	4	4	4	4	4	0	0	0	1	0	0	0	1	0	0	0	0	· · ·	0	0
3	2	2	3	2	4	3	2	0	0	1	0	0	0	0	1	0	0	0	0	· · ·	3	0
4	9	2	3	2	3	3	5	0	0	0	1	0	0	0	0	0	0	0	1	· · ·	3	1
2	3	3	3	3	2	3	3	0	1	0	0	0	0	0	0	1	0	0	0	· · ·	6	0
1	5	4	4	3	2	4	2	1	0	0	0	0	0	0	0	0	0	1	0	· · ·	1	0
4	2	2	2	4	3	3	1	0	0	0	1	0	0	0	1	0	0	0	0	· · ·	2	0
5	1	1	1	1	4	2	2	0	0	0	0	1	0	1	0	0	0	0	0	· · ·	0	0
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	· · ·	·	·
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	· · ·	·	·
·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	·	· · ·	·	·

Moving horizontally we have 1 missing, 2 missings, and so on, with a triangular matrix structure of the respondents due to the near orthogonality of the two variables.

To visualize the demographic groups, centroids are calculated of respondent points in Exhibit 14.8 for each country (Exhibit 14.9) and for each of the age and education groups (Exhibit 14.10). The biggest dispersion is seen in Exhib-

Exhibit 14.8: *Canonical MCA of the indicator matrix with constraining variables the counts of middles and missings (#M and #X). The respondents pile up at discrete positions at the centres of the circles, the areas of which indicate the frequencies*



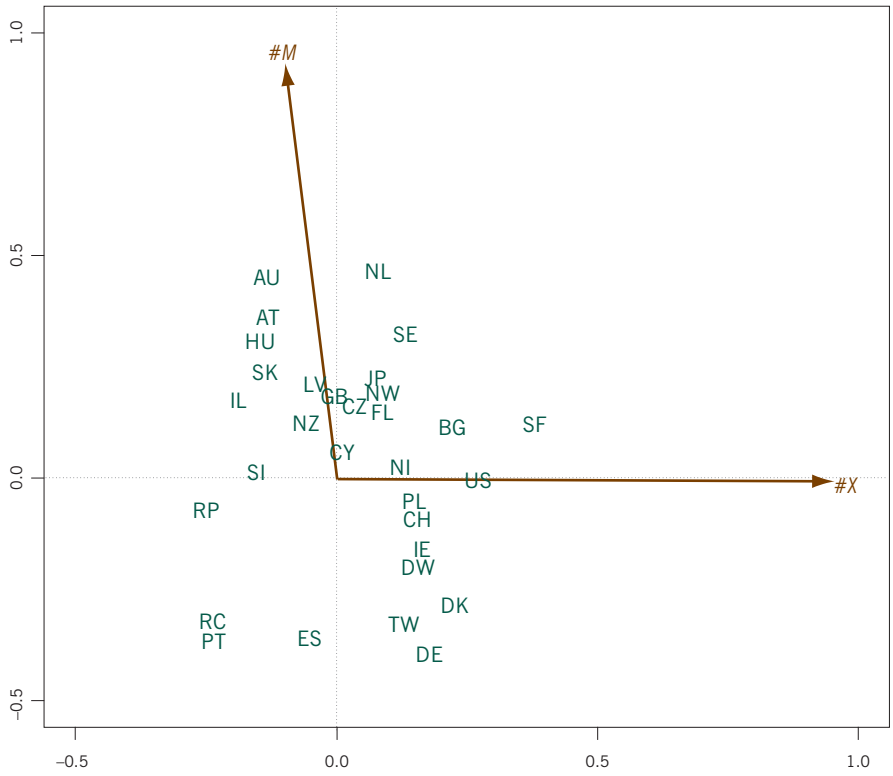


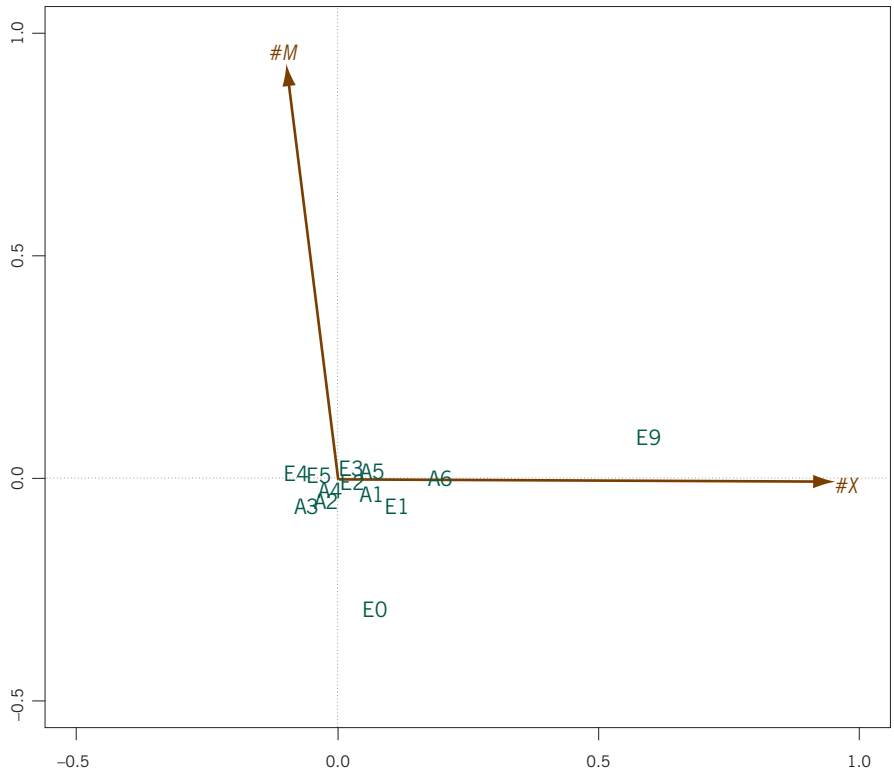
Exhibit 14.9:
Country centroids of the respondents in Exhibit 14.8

it 14.9 for the countries, with Australia and the Netherlands showing the highest use of the middle responses, and Finland the highest on missing responses. Portugal and China give missings and middles the least amongst this group of countries. In Exhibit 14.10 respondents which have their education group missing (E9, 520 cases) are far in the direction of missing responses, while the lowest education group E0 is less than average on middles; the highest education groups E4 and E5 are slightly less than average on missings. The age groups A2, A3 and A4 (26–55 years) are also slightly less than average on missings, while the youngest and oldest groups are slightly higher on average, especially A6 (66+ years). There is no age or education group with a tendency to use the middle responses.

Subset analysis of middle categories

In contrast to the canonical analysis which focuses on a single dimension of middle responses, a subset analysis focuses on the dimensions of all 8 middle categories. Exhibit 14.11 shows two views of the subset MCA of the middle categories, in standard coordinates. The first dimension (horizontal dimension in left hand map) puts all middle categories on the right, so this dimension will coincide with the biplot arrow "*#M*" in Exhibits 14.8–14.10 which simply counts the middles. The second and third dimensions, in the right hand map of Exhibit 14.11, shows

Exhibit 14.10:
Education and age group centroids of the respondents in Exhibit 14.8



that there is a clustering of the middle categories of the first three questions *AM*, *BM*, *CM* (top left), then those of the next four questions *DM*, *EM*, *FM*, *GM* (top right), and quite separately the last question's middle category *HM* at the bottom. The right hand map of Exhibit 14.11 contains information about grouping of middle responses that was not evident in the canonical MCA. The fact that the middle categories group together according to the sequences of questions might indicate a certain type of behaviour on the part of the respondents where they give middle responses to sequences of questions. We can investigate if there is any demographic variable that coincides with this phenomenon.

As in all MCA analyses, each respondent has a position in the map, so any demographic grouping can be represented by a set of centroids. Exhibit 14.12 shows the average positions of the 32 countries. Corresponding to the slightly diagonal orientation of the two clusters at the top of the right hand map of Exhibit 14.11, there are two sets of countries extending from bottom left to top right, indicated by two dashed lines. These correspond to countries that have more than average middle responses on the two clusters of questions, whereas their vertical

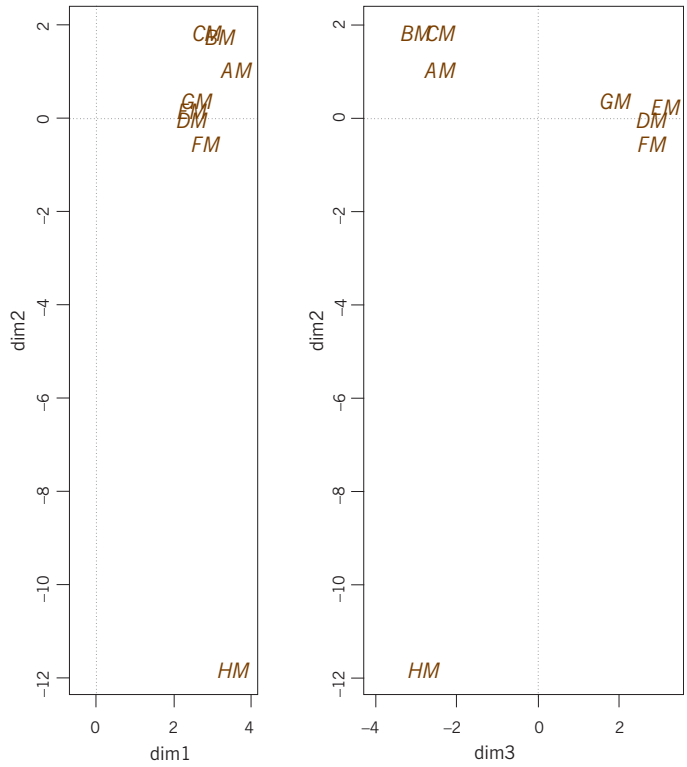


Exhibit 14.11:
Subset MCA of the 8 middle response categories, dimensions 1 by 2 (left) and dimensions 3 by 2 (right). Three clusters are evident in the right hand map

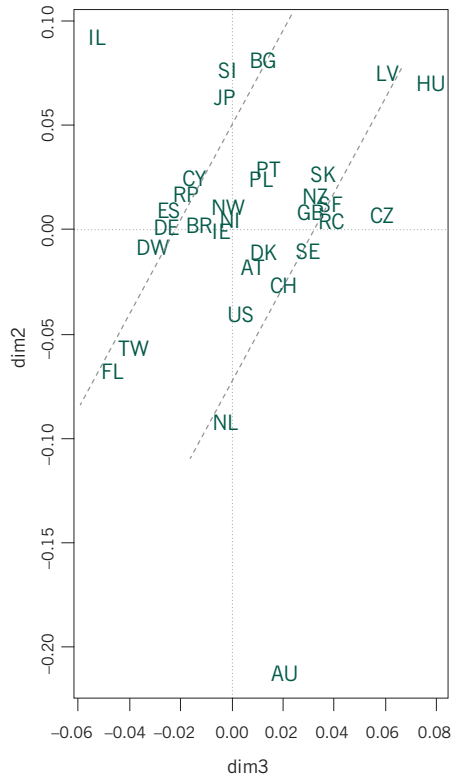
position depends on the number of middle responses on the isolated question *H*. Australia and the Netherlands, for example, which we previously saw had a high level of middle responses, both have a particularly high level on question *H*: going back to the original data, 20.5% of Australians and 13.8% Dutch responded *HM*, whereas for the other 30 countries the average response rate for this category was only 3.9%.

SUMMARY

We have shown how multiple correspondence analysis and its subset and canonical variants can allow a detailed investigation of the patterns of response in a large data set from a social survey. Based on similar studies that we have conducted on several different survey data sets of this type, from the ISSP and Eurobarometer, several results that emanate here appear to be typical:

1. In the analysis of individual responses, the missing response categories dominate. This is partially due to responses sets (all missings) by many respondents, which inflate the associations between the missing categories, but this association is strong even when the response sets are eliminated.

Exhibit 14.12:
Centroids of the countries in the right hand map of Exhibit 14.11. dashed lines indicate a set of countries with more than average middle responses on the first three questions (on left) *and on the next four questions* (on right), *with vertical spreads depending on the incidence of middle response on the last question*



2. Although the middle response categories might appear to be positioned "correctly" amongst the moderate categories in the initial best planar view of the data, inspection of further dimensions shows them to be associated among one another, and separated from their expected positions between the moderate categories. Again, we have found that this phenomenon persists even when response sets (all middles) are removed.

3. A middle response on one question is not generally associated with non-response on other questions. It seems that, if the middle response is used as a satisficing alternative to a non-response, there are some respondents that generally give just middle responses while there are others that give just non-responses.

4. Canonical MCA can be used to isolate a single dimension corresponding to the middle responses, similarly for the missing responses (or whatever responses the researcher is interested in). The positions of the respondents on these dimensions can be averaged within demographic groups to investigate their possible dispersions on these dimensions.

5. Subset MCA can be used to study the middle responses in more detail. As many dimensions as there are middle responses are analysed, so that dimension reduction is necessary to create a map. Generally, the first dimension in this analysis corresponds to the single constraining variable of "middleness" in the canonical MCA, so that the following dimensions reveal the more detailed patterns in middle response.
6. The same approach can be used to investigate patterns of any particular response category or categories. For example, the dimensions of the set of extreme response categories (1's and 5's) could be studied on their own and related to the demographic characteristics.

CHAPTER 15

Case Study 3: The Relationship between Fish Morphology and Diet

The multivariate nature of ecological data is illustrated very well in the morphological data on *Arctic charr* fish, described in Chapters 7 and 12. Apart from the fish morphology, an analysis of the stomach contents of each fish was performed to characterize the fish's diet. Because the diet is measured as a set of percentages of the stomach contents, correspondence analysis is an appropriate way of visualizing the diet variables. Now there are two multivariate observations on each fish: the set of morphological measurements as well as the set of dietary estimates. Our aim will be to decide if there is any non-random relationship between the morphology and the diet, and—if there is—to try to characterize and interpret it. Because we used log-ratio analysis in Chapter 7 to visualize the morphological data, we will maintain this approach while focusing the visualization of the morphology on its relationship to the dietary composition.

Contents

Data set "fishdiet"

The data set "morphology" was introduced in Chapter 7, consisting of 26 measurements on each of 75 *Arctic charr* fish, as well as two dichotomous variables indicating the sex (female/male) of each fish and their habitat (littoral/pelagic). In addition, another set of data is available based on an analysis of the stomach contents of each fish—these are estimated percentages of the contents, by volume, that have been classified into 6 food sources:

PlankCop	plankton – copepods	*InsectLarv*	insects – larvae
PlankClad	plankton – cladocerans	*BenthCrust*	benthos – crustaceans
InsectAir	insects – adults	*BenthMussl*	benthos – mussels

A seventh category *Others* includes small percentages of other food sources. The data for the first 10 fish are given in Exhibit 15.1.This data set, called "fishdiet" constitutes a second matrix of data on the same individuals, and can be considered as explanatory variables that possibly explain the morphological data.

Correspondence analysis of "fishdiet" data

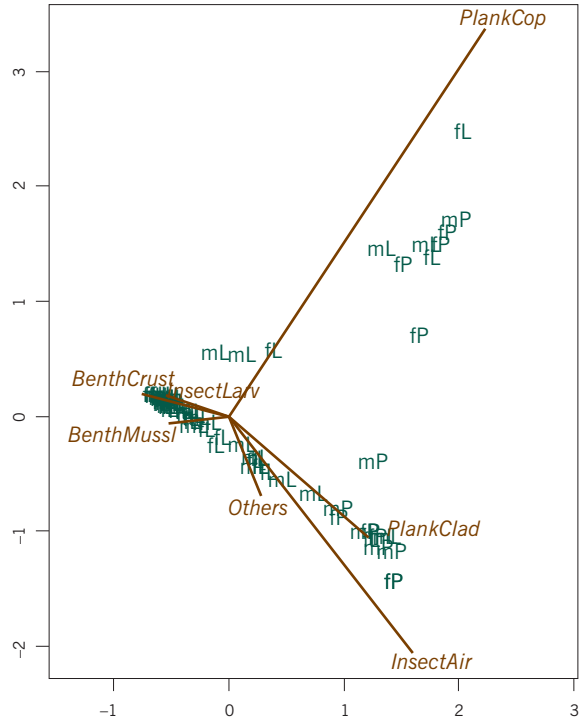
Seeing that the data are compositional, one would immediately think of using log-ratio analysis (LRA), but the large number of zeros makes this approach impractical. Correspondence analysis (CA) is a good alternative but there are two possible approaches here. The first would be to consider just the seven measured percentages—since CA converts the data to profiles, this would re-express the values relative to the actual stomach contents; for example, the first fish in Exhibit 15.1 has only 40% stomach full, and *PlankClad* is 25/40 of this total and 15/40 *InsectAir*, which would be the profile values in CA. The second is to add a column, called "*Empty*" in Exhibit 15.1, which quantifies the emptiness of the stomach, that is 100 minus the sum of the seven measured values. By including the empty component, the sums of each of the rows is now a constant 100% and CA will treat the data in their original form.

Exhibit 15.2 shows the two alternative CAs together for comparison, where we have excluded one fish with zero stomach contents, which would not be a valid observation for the first analysis (all margins have to be strictly positive for CA).

Exhibit 15.1: *Part of data set "fishdiet", showing the first 10 of the* Arctic charr *fish. Data are percentages of stomach contents of different food sources. A column "Empty" has been added as 100 minus the sum of the percentage values in the first seven columns—for example, fish 28 had the whole stomach full, so "Empty" is 0. The supplementary variables sex (1 = female, 2 = male) and habitat (1 = littoral, 2 = pelagic) are also shown*

Fish no.	*PlankCop*	*PlankClad*	*InsectAir*	*InsectLarv*	*BenthCrust*	*BenthMussl*	*Others*	*Empty*	*Sex*	*Habitat*
19	0	25	15	0	0	0	0	60	1	2
23	0	0	0	20	47	8	0	25	2	1
24	0	0	0	8	32	0	0	60	2	1
25	0	0	0	10	22	18	0	50	2	1
27	0	0	0	2	4	4	0	90	1	1
28	0	0	0	10	55	35	0	0	2	1
30	0	0	0	20	44	6	0	30	2	1
31	0	0	0	15	25	40	0	20	1	1
33	0	65	0	0	0	0	0	35	1	2
34	0	48	0	2	0	0	0	50	1	2
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.

a)



b)

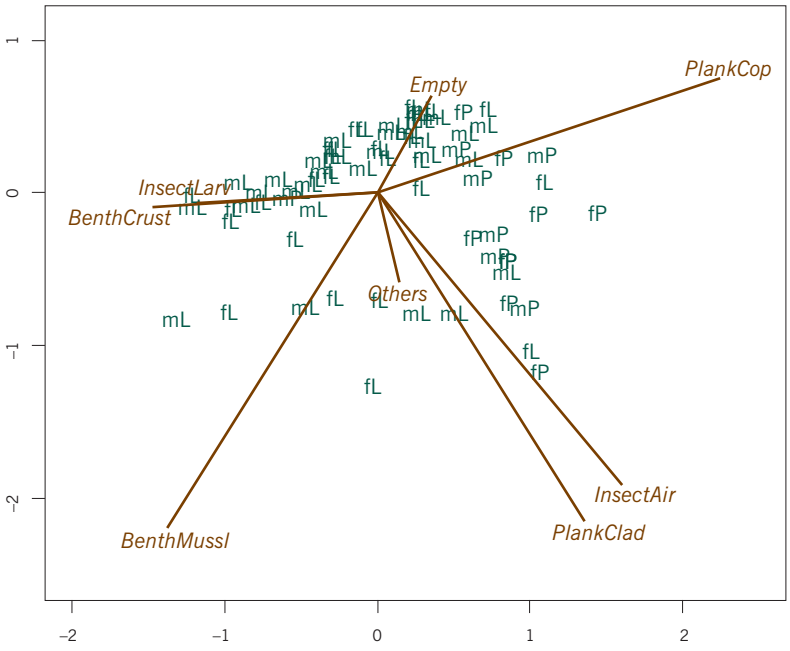
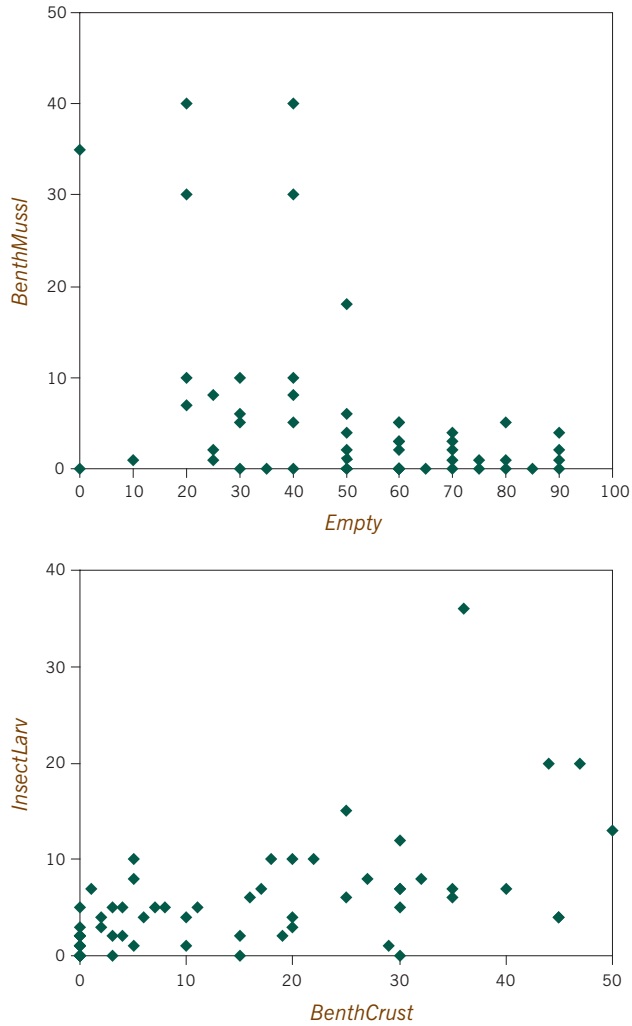


Exhibit 15.2:

CA biplots of the "fishdiet" data, asymmetric scaling with fish in principal coordinates and food sources in standard coordinates: (a) the biplot is the regular CA of the first seven columns of Exhibit 15.1, while (b) includes column 8 (Empty). Fish are labelled by their sex-habitat groups. Total inertias in the two analyses are 1.751 and 1.118 respectively

Exhibit 15.3:
Scatterplots of two pairs of variables, showing the negative relationship between BenthMussl and Empty and positive relationship between InsectLarv and BenthCrust



There are only a few fish with some *PlankCop*, generally at low percentages, but these tend to be associated with less full stomachs so that in relative terms the presence of *PlankCop* is accentuated in the CA in Exhibit 15.2a. Otherwise, there is an opposition between those with relatively more *PlankClad* and *InsectAir* (bottom right of Exhibit 15.2a) compared to those with relatively more *BenthCrust* and *BenthMussl* (to the left). When *Empty* is included (Exhibit 15.2b), it has a higher mean than the other variables, and the centroid of the display moves close to it. Projecting the fish onto the biplot axis defined by *BenthMussl* and *Empty* implies that there is an inverse relationship between the two columns, shown in the upper scatterplot of Exhibit 15.3. Expressed another way, the proportion of benthic mussels increases with stomach fullness. The coincident directions of *InsectLarv*

and *BenthCrust* imply a positive relationship between these two food sources, as shown in the lower scatterplot of Exhibit 15.3.

These two CA biplots display the data in different ways and the biologist needs to decide if either or both are worthwhile. The question is whether the percentages are of interest relative to actual stomach contents, or on their original percentage scale relative to the whole stomach. For example, the separation of the group of fish in the direction of *PlankCop* in Exhibit 15.2a is non-existent in Exhibit 15.2b—relative to what is in the stomach, this group of fish distinguishes itself from the others, but not so much when seen in the context of the stomach as a whole.

How is the diet related to habitat and sex?

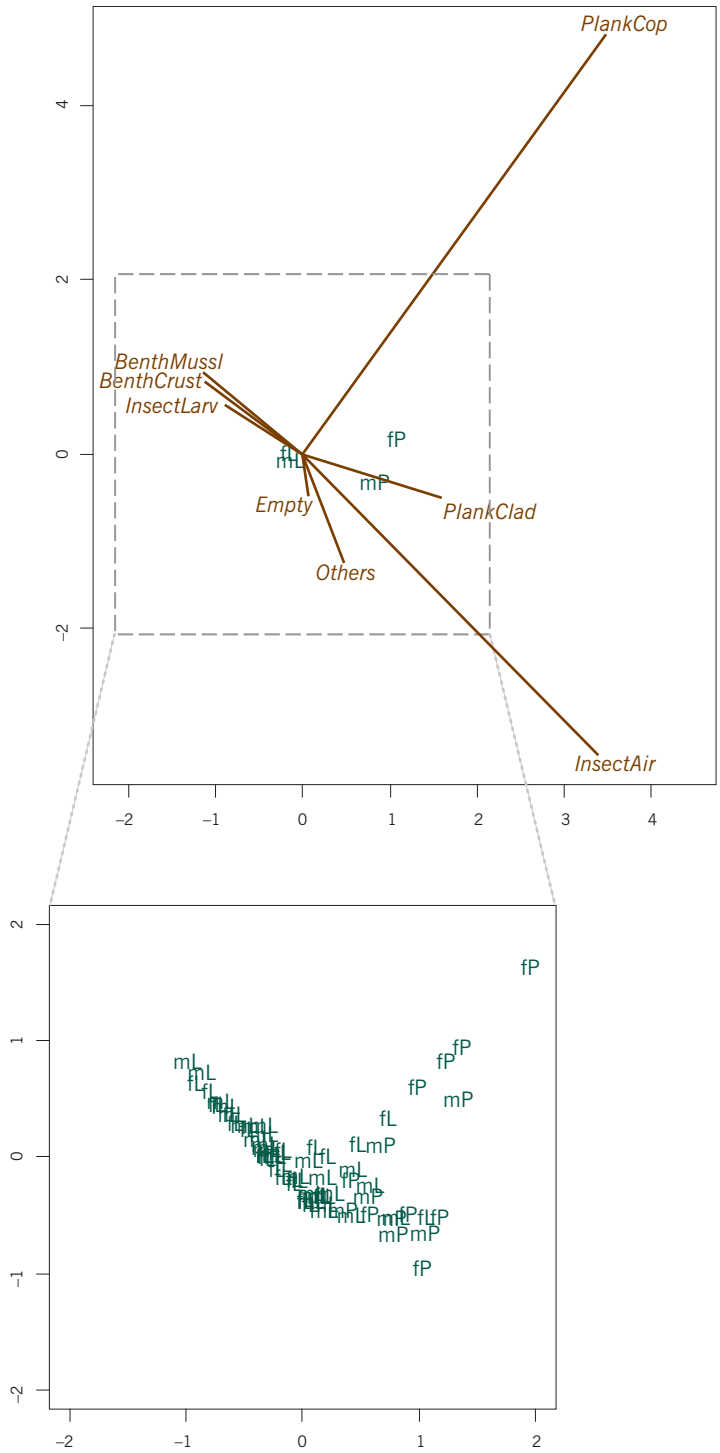
As described in Chapter 12, constraining by a categorical variable is equivalent to performing a type of centroid discriminant analysis, illustrated in Exhibit 11.2 for the morphological data. We repeat that analysis on the "fishdiet" data, with the groups defined by the interactively coded sex-habitat variable with four categories: fL, mL, fP and mP. Exhibit 15.4 shows the resulting biplot. As in the morphological analysis of Exhibit 11.2, the habitat differences are more important than the sex differences. Contrary to the morphological analysis, the diet difference between sexes in the pelagic group is bigger than that in the littoral group. The lack of difference between female and male littoral fish (fL and mL) is seen clearly by the single line of individual points to the top left of the biplot, in the direction of *BenthCrust*, *BenthMussl* and *InsectLarv*; while on the right there is a separation into two "streams", mainly female pelagic (fP) to upper right, in the direction of *PlankCop*, and mainly male pelagic (mP) to lower right, in the direction of *PlankClad* and *InsectAir*. There are some exceptions: for example, some fP points are in the lower right group, and there are few male and female littoral fish on the right.

The unconstrained CA in Exhibit 15.2b has total inertia equal to 1.118 and the part of this inertia explained by the sex-habitat variable is equal to 0.213, or 19.0% of the total. A permutation test shows that the relationship between diet and sex-habitat is highly significant: $p < 0.0001$. This part of inertia forms the total inertia in the analysis of Exhibit 15.4, which explains almost all of it (99.6%) in two dimensions (the analysis of the four centroids is three-dimensional, so the 0.4% unexplained is in the third dimension).

Canonical log-ratio analysis

In Chapter 7 we analyzed the morphological data set on its own using the log-ratio approach. We now want to relate the morphological data to the diet data, in other words constrain the dimensions of the log-ratio analysis to be related to the diet variables, which we could call *canonical*, or constrained, *log-ratio analysis* (CLRA). It is useful here to give the equations of the analysis, putting together the theories of Chapters 7 and 12.

Exhibit 15.4:
CA discriminant analysis of the sex-habitat groups (equivalent to CCA with categorical sex-habitat variable as the constraining variable). The centroids of the four groups are shown in the upper plot. The individual fish, which are contained in the box shown in the biplot, have been separated out in the plot, with enlarged scale for sake of legibility. Total inertia of the four centroids is equal to 0.213



In Chapter 7 log-ratio analysis was defined as the weighted SVD: $\mathbf{S} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{Y}\mathbf{D}_c^{\frac{1}{2}} = \mathbf{U}\mathbf{D}_\varphi\mathbf{V}^\mathsf{T}$ of the double-centred matrix: $\mathbf{Y} = (\mathbf{I} - \mathbf{1}\mathbf{r}^\mathsf{T})\mathbf{L}(\mathbf{I} - \mathbf{1}\mathbf{c}^\mathsf{T})^\mathsf{T}$ of logarithms: $\mathbf{L} = \log(\mathbf{N})$ of the data $\mathbf{N}$ (see (7.1)–(7.4)). The dimensionality of this analysis is equal to 25, one less than the number of morphometric measurements. The constraining variables are the 7 diet variables, without the "*Empty*" column (here it makes no difference whether it is included or not as an explanatory variable). The matrix $\mathbf{X}$ consists of the 7 diet variables after they have been standardized. Then (12.2) defines the projection matrix as $\mathbf{Q} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{D}_r^{\frac{1}{2}}$ and the matrix $\mathbf{S}$ is projected onto the space of the diet variables by $\mathbf{S}^* = \mathbf{Q}\mathbf{S}$. The unconstrained component of $\mathbf{S}$ in the space uncorrelated with the diet variables is $\mathbf{S}^\perp = (\mathbf{I} - \mathbf{Q})\mathbf{S}$ (see (12.3) and (12.7) respectively). The dimensionality of $\mathbf{S}^*$ is 7 in this case. The SVD of $\mathbf{S}^*$ is performed in the usual way, with subsequent computation of principal and standard coordinates.

Relationship of morphology to diet

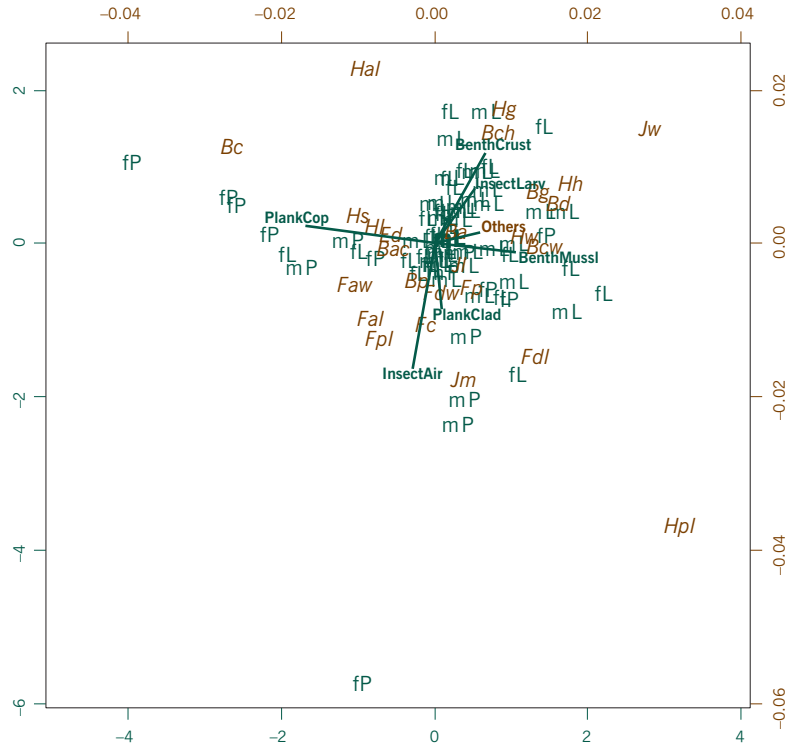
The first interesting statistic from this constrained log-ratio analysis is the part of the morphological log-ratio variance that is explained by the diet variables: it turns out to be 14.5%, so that 85.5% is not related—at least, linearly—to the diet. Our interest now turns to just that 14.5% of the explained variance, 0.0002835, compared to the total variance of 0.001961 of the morphological data. This variance is now contained in a 7-dimensional space, and our view of this space is, as always, in terms of the best-fitting plane. The principal axes of this plane account for small percentages of the total (original) morphological variance (5.6% and 4.0% respectively), but for the moment we focus on how the constrained variance (0.0002835) is decomposed, and the axes account for 38.9% and 27.6% of that amount, which is the part of the variance that interests us. Exhibit 15.5 shows the biplot of these first two constrained axes.

In Exhibit 15.5 the fish points are in standard coordinates and the morphological variables in principal coordinates. As in Exhibit 7.3, the dispersion of the fish points is so low that the coordinates have to be scaled up to appreciate their relative positions. The diet variables are displayed according to their correlation coefficients with the axes, and have also been scaled up (by 2) to facilitate their display. As explained in Chapter 12, there are two ways to show the diet variables: using the correlation coefficients as in Exhibit 15.5, or in terms of the coefficients of the linear combinations of the variables that define the axes. For example, axes 1 and 2 are in fact the linear combinations:

$$\begin{aligned}\textit{Axis 1} = {} & -0.761 \times \textit{PlankCop} + 0.272 \times \textit{PlankClad} - 0.159 \times \textit{InsectAir} + 0.103 \times \\ & \textit{InsectLarv} + 0.071 \times \textit{BenthCrust} + 0.388 \times \textit{BenthMussl} + 0.217 \times \textit{Others}\end{aligned}$$

$$\begin{aligned}\textit{Axis 2} = {} & 0.280 \times \textit{PlankCop} - 0.076 \times \textit{PlankClad} - 0.689 \times \textit{InsectAir} + 0.140 \times \\ & \textit{InsectLarv} + 0.505 \times \textit{BenthCrust} - 0.188 \times \textit{BenthMussl} + 0.116 \times \textit{Others}\end{aligned}$$

Exhibit 15.5:
Weighted LRA biplot constrained by the fish diet variables, with rows (fish) in standard coordinates and columns (morphological variables) in principal coordinates. The coordinates of the diet variables have been multiplied by 2 to make them more legible (use the green scale for these points). 66.5% of the constrained variance is accounted for (but only 9.6% of the original total variance)



where the axes (that is, the coordinates of the fish on the axes) as well as the variables are all in standard units, that is with standard deviations equal to 1.

Because the diet variables are correlated, the variable-axis correlations are not the same as the above coefficients, which are regression coefficients if the axes are regressed on the variables.

As explained in Chapter 12, the above equations are exact (that is, $R^2 = 1$ if one were to perform the regression), but thinking of Exhibit 15.5 from the biplot viewpoint, the R^2 of each diet variable can be computed as the sum of squared correlations to measure how accurately each variable is displayed:

PlankCop:	$(-0.860)^2$	+	$(0.113)^2$	=	0.752
PlankClad:	$(0.055)^2$	+	$(-0.447)^2$	=	0.203
InsectAir:	$(-0.142)^2$	+	$(-0.806)^2$	=	0.669
InsectLarv:	$(0.260)^2$	+	$(0.370)^2$	=	0.205
BenthCrust:	$(0.336)^2$	+	$(0.610)^2$	=	0.485
BenthMussl:	$(0.496)^2$	+	$(-0.052)^2$	=	0.249
Others:	$(0.299)^2$	+	$(0.083)^2$	=	0.096

PlankCop and *InsectAir* are explained more than 50%—this means that we could recover their values with an error of less than 50% by projecting the fish points onto the biplot axes that they define in Exhibit 15.5. Variables such as *PlankClad*, *InsectLarv* and *BenthMussl* are poorly reconstructed in the biplot. But remember that it was not the intention of this biplot to recover these values—in fact, this was the aim of the correspondence analysis of Exhibit 15.2. The aim here is rather to recover the values of the morphological variables that are directly related to diet, in a linear sense.

Permutation test of morphology–diet relationship

In order to test for significance of the morphology–diet relationships we are detecting, a permutation test can be performed as described previously: use the inertia explained by the diet variables as a test statistic, and then randomly permute the sets of diet values so that many (9999 in this case) additional data sets are constructed under the null hypothesis that there is no morphology–diet correlation. The result is the null permutation distribution in Exhibit 15.6. If there were no (linear) relationship between morphology and diet, we would expect a propor-

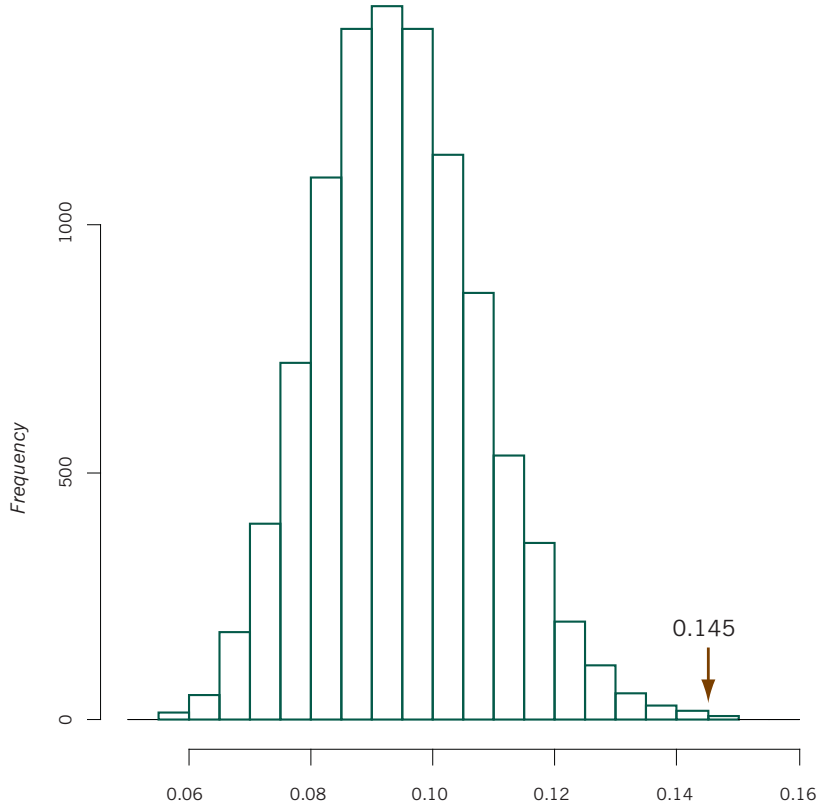


Exhibit 15.6:
Permutation distribution of the proportion of variance explained in the morphological variables by the diet variables, under the null hypothesis of no relationship between these two sets of variables. The p-value associated with the observed proportion of 0.145 is 0.0007

tion of explained variance of 0.093 (9.3%), with the estimated distribution shown. Our observed value of 0.145 (14.5%) is in the far right tail of the distribution, and only 6 of the permuted data sets gives a proportion higher than this value—hence the *p*-value is 7/10,000 = 0.0007 (the observed value is included with the 26 higher ones to make this calculation).

Up to now we included all of the diet variables, but it may be that only a subset of them explain a significant part of the variance. The individual contributions of the variables to this explained variance can not be calculated, but a stepwise search can be conducted similar to that of stepwise regression. First, the single variable that explains the most variance is computed, by trying each one at a time. The amounts of explained variance for each variable are:

PlankCop:	0.0412
PlankClad:	0.0201
InsectAir:	0.0294
InsectLarv:	0.0163
BenthCrust:	0.0241
BenthMussl	0.0285
Others:	0.0139

so that *PlankCop* explains the most. We now perform a permutation test on this explained variance, by permuting the values of *PlankCop* and recomputing the explained variance each time. The *p*-value is estimated at 0.0008, so this is highly significant (see Exhibit 15.7).

The next step is to determine which second variable, when added to *PlankCop*, explains the most variance. The results are:

PlankCop + *PlankClad:*	0.0637
PlankCop + *InsectAir:*	0.0707
PlankCop + *InsectLarv:*	0.0557
PlankCop + *BenthCrust:*	0.0631
PlankCop + *BenthMussl:*	0.0638
PlankCop + *Others:*	0.0535

so that *InsectAir* explains the most additional variance. The permutation test now involves fixing the *PlankCop* variable and permuting the values of *InsectAir*, leading to an estimated *p*-value of 0.0097 (see Exhibit 15.7).

We now continue the stepwise process by looking for a third dietary variable which adds the most explained variance to *PlankCop* and *InsectAir*.

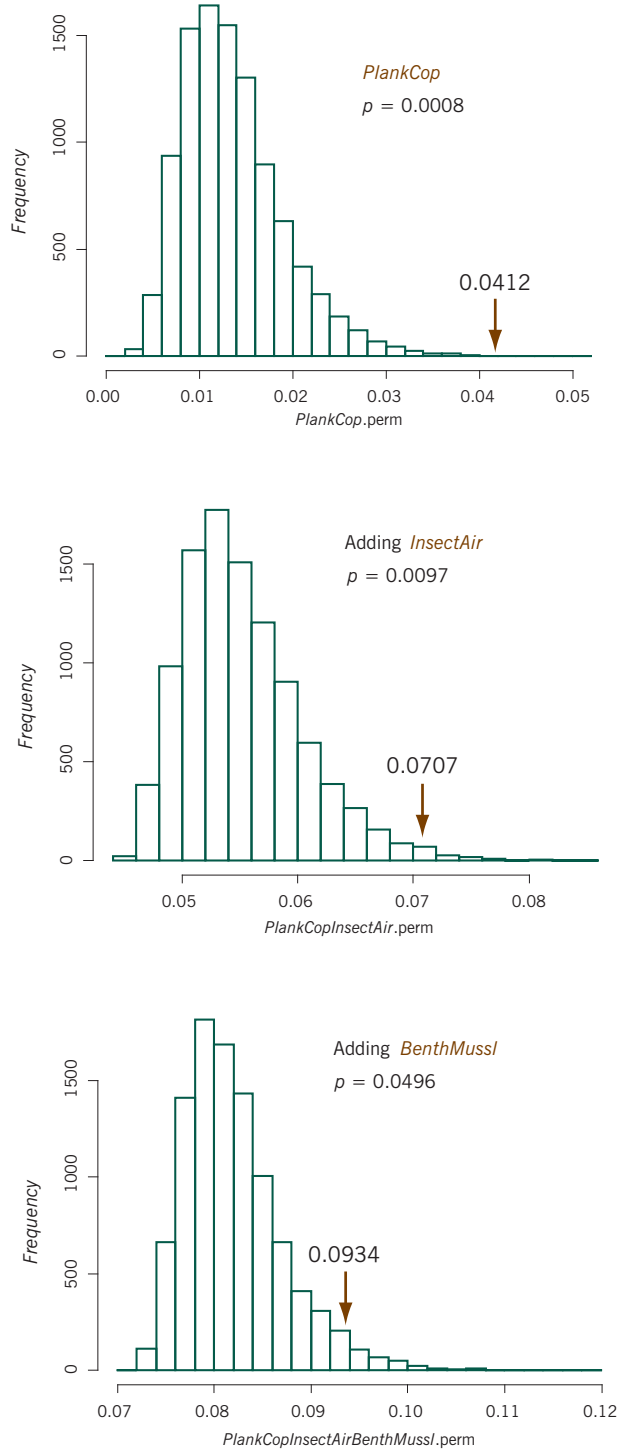
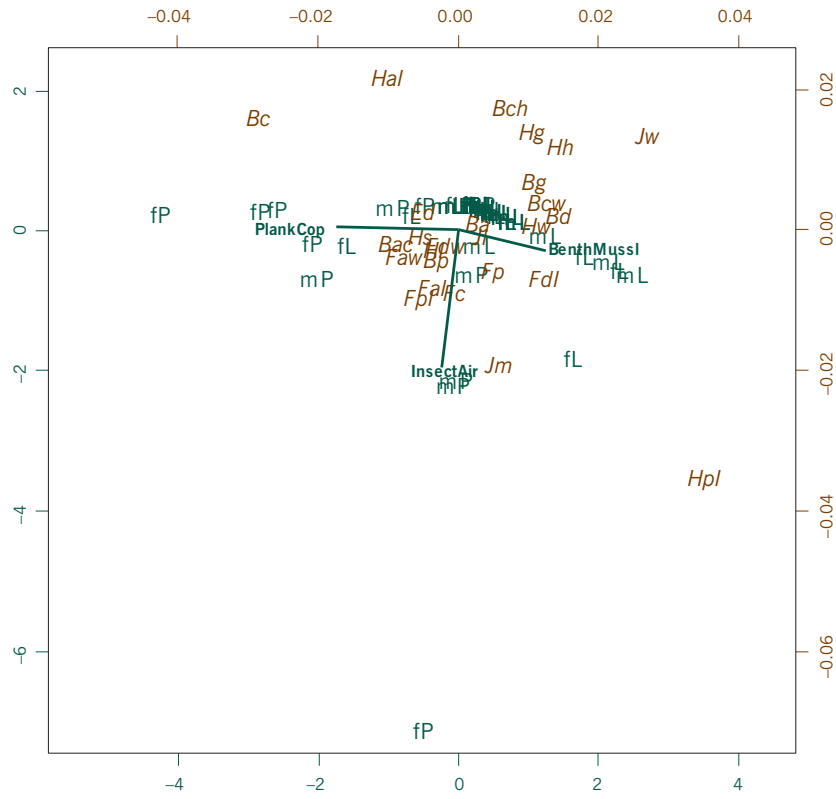


Exhibit 15.7:
Permutation distributions and observed values (explained variances) for the three stages of the stepwise process, introducing successively, from left to right, PlankCop, InsectAir and BenthMussl. The p-values given by the three tests are 0.0008, 0.0097 and 0.0496 respectively

Exhibit 15.8:
Weighted LRA biplot constrained by the three significant fish diet variables, using the same scalings as Exhibit 15.5. 85.8% of the constrained variance is accounted for



PlankCop + *InsectAir* + *PlankClad*:	0.0890
PlankCop + *InsectAir* + *InsectLarv*:	0.0826
PlankCop + *InsectAir* + *BenthCrust*:	0.0862
PlankCop + *InsectAir* + *BenthMussl*:	0.0934
PlankCop + *InsectAir* + *Others*:	0.0827

So the winner is *BenthMussl*. The permutations test fixes *PlankCop* and *InsectAir* and permutes *BenthMussl*, leading to an estimated *p*-value of 0.0496 (see Exhibit 15.7). No other variables enter below the "classical" level of 0.05 and the final canonical LRA, using the three variables *PlankCop*, *InsectAir* and *BenthMussl*, explains a total of 9.15% of the variance of the morphological data. The canonical LRA of the morphological data with just these three significant diet variables is shown in Exhibit 15.8.

Finally, the most highly contributing morphometric variables were identified—there are eight of them, out of the 26—contributing a total of 66% of the iner-

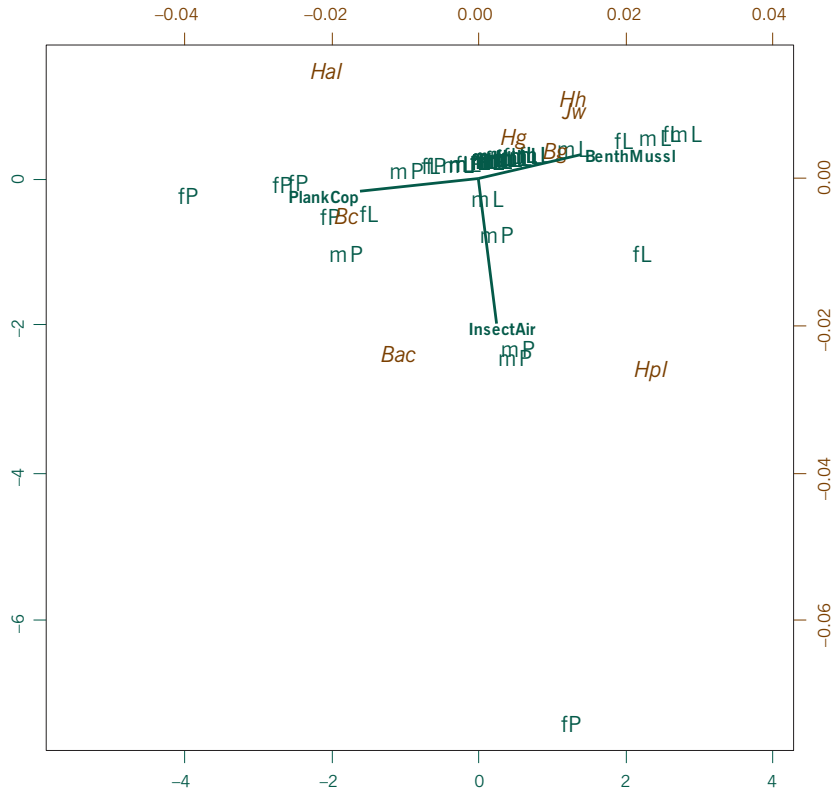


Exhibit 15.9: *Weighted LRA biplot constrained by the three significant fish diet variables, and using only the most highly contributing morphometric variables. The same scalings as Exhibits 15.5 and 15.8 are used for all three sets of points. 94.6% of the constrained variance is accounted for*

tia in the constrained biplot. The analysis was repeated from the start, using just these eight variables, and the constrained biplot is shown in Exhibit 15.9. This biplot shows the essential structure in the morphological–diet relationship. The first dimension opposes *PlankCop* against *BenthMussl*, which we already saw in the CA-DA of Exhibit 15.4 was important in the separation of pelagic from littoral groups. The band of fish seen from left to right have zero *InsectAir*, with mostly littoral fish on the right with higher than average benthic mussels in the stomach, also with larger jaw widths and head heights, and mostly pelagic fish on the left feeding on more planktonic cladocerans, and with relatively larger tails (one female littoral fish is also on the left, as in previous biplots, and seems to be an exception in this otherwise pelagic group). *InsectAir* (flying insects) defines a separate perpendicular direction of spread, pulling out a few fish, especially one female pelagic (fP) which was seen to be isolated in previous biplots—this fish has 15% *InsectAir* in its stomach, much higher than any other fish in this data set, and also happens to have one of the highest values of posterior head length (*Hpl*).

SUMMARY

This case study shows how a biplot, specifically the log-ratio biplot in this case, can allow investigation of the patterns in a multivariate data set that are directly related to a set of external variables. The use of permutation tests permits distinguishing the external variables that explain significant variation from the others. Some biological conclusions about the relationship between fish morphology and fish diet are as follows:

1. The fish included in this study are characterized by two distinct forms feeding in different habitats (pelagic *vs* littoral) and on different prey (benthos *vs* plankton).
2. The feeding habits are associated with distinctive morphologies, with fish feeding on benthic crustaceans being more bulky and with greater jaws relative to the more slender plankton eating fish.
3. In the littoral zone males and females display similar diets, whereas in the pelagic males are more oriented towards planktonic cladocerans (waterfleas) and surface insects but females prefer deep dwelling copepods.

APPENDIX A

Computation of Biplots

In this appendix the computation of biplots is illustrated using the object-orientated programming language R, which can be freely downloaded from the R-project website:

```
http://www.r-project.org
```

It is assumed that the reader has some basic knowledge of R—if not, consult some of the web resources and tutorials given on this website and in the bibliography. An R script file is given on the website:

```
http://www.multivariatestatistics.org
```

as well as the data sets, so that readers can reproduce the biplots described in this book. The idea in this Appendix is to explain some of these commands, and so serves as an R tutorial in the context of the material presented in this book.

R commands will be indicated in slanted typewriter script in *brown*, while the results are given in non-slanted green. A + at the start of a line indicates the continuation of the command (in the script file the command is given as a single line). Notice that the idea in this appendix is to educate the user in the use of R by showing alternative ways of arriving at biplot solutions, including different ways of plotting the final results. In some cases the R code might not be the most efficient way of arriving at the final goal, but will illustrate different functions in the R toolbox that can be learnt by example.

Reading data into R

There are two recommended ways of reading data into R. Suppose that you want to read in the data in Exhibit 1.1. These data are either in a text file or an Excel file, for example. Suppose that the file `EU2008.txt` is in your R working directory and contains the following:

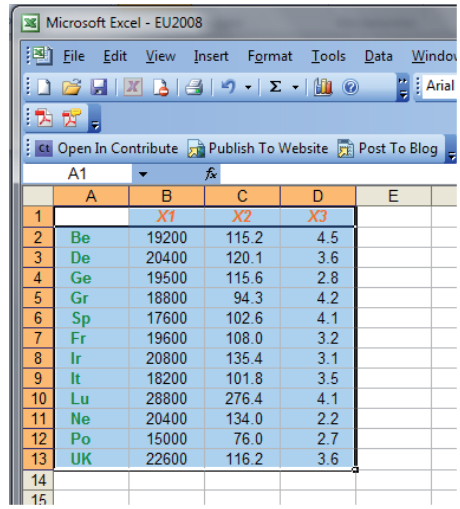
```
      X1     X2    X3
Be 19200  115.2  4.5
De 20400  120.1  3.6
Ge 19500  115.6  2.8
Gr 18800   94.3  4.2
Sp 17600  102.6  4.1
```

```
Fr   19600   108.0   3.2
Ir   20800   135.4   3.1
It   18200   101.8   3.5
Lu   28800   276.4   4.1
Ne   20400   134.0   2.2
Po   15000    76.0   2.7
UK   22600   116.2   3.6
```

Then the following command will read the data file into the data frame `EU2008`:

```
EU2008 <- read.table("EU2008.txt")
```

The alternative way (for Windows users) is to simply copy the file and then read it from the file called "clipboard". The copying can be done in the text file or in an Excel file (by painting out the data file and then using either the pull-down Edit menu, or Ctrl-C, or right-clicking on the mouse and selecting Copy), for example:



and then read the file from the clipboard using

```
EU2008 <- read.table("clipboard")
```

Notice that the function `read.table` successfully reads the table because of the blank cell in the upper left corner of the spreadsheet, which effectively signals to

the function that the first row contains the columns labels and the first column the row labels. Once the data file has been read, computations and graphical displays can commence.

The following commands reproduce Exhibit 1.2:

Chapter 1: Biplots—the Basic Idea

```
windows(width=11, height=6)
par(mfrow=c(1,2), cex.axis=0.7)
plot(EU2008[,2:1], type="n", xlab="GDP/capita",
+                            ylab="Purchasing power/capita")
text(EU2008[,2:1], labels=rownames(EU2008), col="green", font=2)
plot(EU2008[,2:3], type="n", xlab="GDP/capita",
+                            ylab="Inflation rate")
text(EU2008[,2:3], labels=rownames(EU2008), col="green", font=2)
```

The first command above sets the window size in inches—by default it would be 7 inches square—and the second command sets the plot layout with two plots side by side, and axis scale labelling in a font size 0.7 times the default. These settings remain in this window until it is closed.

Three-dimensional plotting is possible using the R package **rgl**, which should be downloaded separately—for example, using the pull-down menu in R, select Packages and Install packages, then choose a mirror site and finally choose "rgl" from the long alphabetical list of available packages. The three-dimensional display on which Exhibit 1.3 is based can then be obtained as follows:

```
library(rgl)
plot3d(EU2008[,c(2,1,3)], xlab="GDP", ylab="Purchasing power",
+                         zlab="Inflation", font=2, col="brown",
+                                           type="n")
text3d(EU2008[,c(2,1,3)], text=rownames(EU2008), font=2,
+                                                col="green")
```

The data set "bioenv" is assumed to have been read into the data frame `bioenv`, with 8 columns: the species *a* to *e*, and the three continuous variables *pollution*, *depth* and *temperature*. To calculate the linear regression of species *d* on *pollution* and *depth*:

Chapter 2: Regression Biplots

```
d <- bioenv[,4]
y <- bioenv[,6]
x <- bioenv[,7]
summary(lm(d~y+x))
```

```
(...)
Coefficients:
             Estimate  Std. Error   t value   Pr(>|t|)
(Intercept)  6.13518     6.25721     0.980    0.33554
y           -1.38766     0.48745    -2.847    0.00834 **
x            0.14822     0.06684     2.217    0.03520 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.162 on 27 degrees of freedom
Multiple R-squared: 0.4416,      Adjusted R-squared: 0.4003
F-statistic: 10.68 on 2 and 27 DF, p-value: 0.0003831
```

There are two ways to calculate the standardized regression coefficients: first by standardizing all the variables and repeating the regression:

```
ds <- (d-mean(d))/sd(d)
ys <- (y-mean(y))/sd(y)
xs <- (x-mean(x))/sd(x)
summary(lm(ds~ys+xs))

(...)
Coefficients:
               Estimate Std. Error   t value  Pr(>|t|)
(Intercept)   2.487e-17  1.414e-01 1.76e-16   1.00000
ys           -4.457e-01  1.566e-01   -2.847   0.00834 **
xs            3.472e-01  1.566e-01    2.217   0.03520 *
---
Residual standard error: 0.7744 on 27 degrees of freedom
Multiple R-squared: 0.4416,      Adjusted R-squared: 0.4003
F-statistic: 10.68 on 2 and 27 DF, p-value: 0.0003831
```

or by direct calculation using the unstandardized coefficients and the standard deviations of the variables:

```
lm(d~y+x)$coefficients[2]*sd(y)/sd(d)
          y
-0.4457286

lm(d~y+x)$coefficients[3]*sd(x)/sd(d)
        x
0.3471993
```

The standardized regression coefficients for all five variables can be calculated in a loop and stored in a matrix **B**, as in (2.2):

```
B <- lm(bioenv[,1]~y+x)$coefficients[2:3]*c(sd(y),
+                                    sd(x))/sd(bioenv[,1])
for(j in 2:5) B <- cbind(B,lm(bioenv[,j]~ y+x)$coefficients[2:3]
+                   *c(sd(y),sd(x))/sd(bioenv[j]))

B
```

```
           y           x
B -0.7171713  0.02465266
  -0.4986038  0.22885450
   0.4910580  0.07424574
  -0.4457286  0.34719935
  -0.4750841 -0.39952072
```

A regression biplot similar to the one in Exhibit 2.5 can be drawn as follows:[8]

```
plot(xs, ys, xlab="x*(depth)", ylab="y*(pollution)", type="n",
+  asp=1, cex.axis=0.7)
text(xs, ys, labels=rownames(bioenv))
text(B[,2:1], labels=colnames(bioenv[,1:5]), col="red", font=4)
arrows(0,0,0.95*B[,2],0.95*B[,1], col="red", angle=15,
+  length=0.1)
```

So far, in the plotting instructions, several graphical parameters have appeared to enhance the final figure, for example:

- `col`: sets the colour of a label or a line different from the default black, e.g. `col="red"`.
- `cex`: changes the font size of the label, e.g. `cex=0.8` scales the label to 80% of its default size.
- `cex.axis`: changes the font size of the scale on the axes.
- `font`: changes the font style, e.g. `font=4` is bold italic.

These options, and many more, are listed and explained as part of the `par` function in R—for help on this function, enter the command:

```
?par
```

To avoid repetition and commands that are full of these aesthetic enhancements of the plots, they will generally be omitted in this computational appendix from

8. Notice that any slight formatting change or improvement in Exhibit 2.5 compared to the R output, for example, the font sizes or positions of axis labels, has been done external to R to produce the final figure.

now on; but they nevertheless appear in the online script file. In addition, axis labelling will be generally omitted as well—this can be suppressed by including in the plot function the options `xlab="", ylab=""`, otherwise the default is to label the axes with the names of the variables being plotted.

Chapter 3: Generalized Linear Model Biplots

The species *d* is nonlinearly transformed by the fourth-root, and then regressed on standardized pollution and depth:

```
d0 <- d^0.25
summary(lm(d0~ys+xs))

(...)
Coefficients:
             Estimate  Std. Error  t value   Pr(>|t|)
(Intercept)   1.63908     0.09686   16.923   6.71e-16 ***
ys           -0.28810     0.10726   -2.686     0.0122 *
xs            0.05959     0.10726    0.556     0.5831
---
Residual standard error: 0.5305 on 27 degrees of freedom
Multiple R-squared: 0.2765,      Adjusted R-squared: 0.2229
F-statistic: 5.159 on 2 and 27 DF, p-value: 0.01266
```

Additional scripts are given in the online R script file for saving the coefficients for all the regressions (Exhibits 3.1, 3.4, 3.5). We give further examples just for species *d*:

Fitting a Poisson regression model for species *d*:

```
summary(glm(d~ys+xs, family=poisson))

(...)
Coefficients:
             Estimate   Std. Error  z value    Pr(>|z|)
(Intercept)   2.29617      0.06068   37.838     < 2e-16 ***
ys           -0.33682      0.07357   -4.578    4.69e-06 ***
xs            0.19963      0.06278    3.180     0.00147 **
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 144.450 on 29 degrees of freedom
Residual deviance:  88.671 on 27 degrees of freedom
AIC: 208.55
```

To get the "error" deviance for this Poisson regression:

```
poisson.glm <- glm(d~ys+xs, family=poisson)
poisson.glm$deviance/poisson.glm$null.deviance
[1] 0.6138564
```

Fitting a logistic regression model, for example for species *d*, after converting its values to presence/absence (1/0):

```
d01 <- d>0
summary(glm(d01~ys+xs, family=binomial))

(...)
Coefficients:
             Estimate   Std. Error  z value    Pr(>|z|)
(Intercept)    2.7124       0.8533    3.179    0.00148 **
ys            -1.1773       0.6522   -1.805    0.07105 *
xs            -0.1369       0.7097   -0.193    0.84708
---
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19.505 on 29 degrees of freedom
Residual deviance: 15.563 on 27 degrees of freedom
AIC: 21.563
```

To get the "error" deviance for this logistic regression:

```
logistic.glm <- glm(d01~ys+xs, family=binomial)
logistic.glm$deviance/logistic.glm$null.deviance
[1] 0.7979165
```


The data set "countries" (Exhibit 4.1) is assumed to have been read into the data frame `MT_matrix`—this is the 13 × 13 dissimilarity matrix between 13 countries given by student "MT". The R function `cmdscale` performs classical multidimensional scaling. Exhibit 4.2 is obtained as follows (notice the option `asp=1` which sets the aspect ratio equal to 1 so that the scales have identical unit intervals horizontally and vertically):

```
plot(cmdscale(MT_matrix), type="n", asp=1)
text(cmdscale(MT_matrix), labels=colnames(MT_matrix))
```

The data set "attributes" (first six columns of Exhibit 4.3) is assumed to have been read into the data frame `MT_ratings`, with 13 rows and 6 columns. To add the regression coefficients of each attribute to Exhibit 4.2 to eventually obtain Exhibit 4.5, first store the coordinates of the countries (rightmost pair of columns in Exhibit 4.3) in `MT_dims`:

```
MT_dims <- cmdscale(MT_matrix, eig=T, k=2)$points
colnames(MT_dims) <- c("dim1","dim2")
```

then calculate the regression coefficients and strore them in `MT_coefs` (Exhibit 4.4)

```
MT_coefs <- lm(MT_ratings[,1]~MT_dims[,1]+MT_dims[,2])
+  $coefficients
for(j in 2:ncol(MT_ratings)) MT_coefs<-rbind(MT_coefs,
+  lm(MT_ratings[,j]~MT_dims[,1]+MT_dims[,2])$coefficients)
```

Finally, plot the regression coefficients on the MDS plot (Exhibit 4.5)

```
plot(cmdscale(MT_matrix), type="n", asp=1)
text(cmdscale(MT_matrix), labels=colnames(MT_matrix))
arrows(0,0,MT_coefs[,2], MT_coefs[,3], length=0.1, angle=10)
text(1.2*MT_coefs[,2:3], labels=colnames(MT_ratings))
```

As an example of the definition of a function, the following is a function called `chidist` to compute the chi-square distances between the rows or columns of a supplied rectangular matrix.

```
chidist <- function(mat,rowcol=1) {
  if(rowcol= =1) {
    prof <- mat/apply(mat,1,sum)
    rootaveprof <- sqrt(apply(mat,2,sum)/sum(mat))
    }
  if(rowcol= =2) {
    prof <- t(mat)/apply(mat,2,sum)
    rootaveprof <- sqrt(apply(mat,1,sum)/sum(mat))
    }
    dist(scale(prof,FALSE,rootaveprof))
}
```

So `chidist(N,1)` calculates chi-square distances between row profiles (this is the default, so for row profiles, `chidist(N)` is sufficient), `chidist(N,2)` calculates chi-square distances between column profiles. The following code performs and saves the MDS (there are four dimensions in this problem—this is explained in Chapter 8 on correspondence analysis), and prints the percentages of variance explained on each dimension:

```
abcde <- bioenv[,1:5]
abcde_mds <- cmdscale(chidist(abcde), eig=T, k=4)
100*abcde_mds$eig/sum(abcde_mds$eig)
```

Then the site points are plotted, as in Exhibit 4.6—notice the extra parameters in the plot function for setting limits on the plot, anticipating additional points to be added to the plot, and also notice that if a matrix argument is given to the functions `plot` and `text`, then by default the first two columns are used:

```
plot(abcde.mds$points, type="n", asp=1, xlim=c(-1.2,1.6),
+  ylim=c(-1.1,1.8))
text(abcde.mds$points)
```

In this case, to add the species points, the species data are first converted to profiles, standardized by dividing them by the square roots of their marginal ("expected") values, as is the case when calculating chi-square distances:

```
abcde_prof <- abcde/apply(abcde,1,sum)
abcde_prof_stand <- t(t(abcde_prof)/sqrt(apply(abcde,2,sum)/
+  sum(abcde)))
```

The regressions are then performed on the dimensions, the coefficients saved and then added as arrows to the map plotted above:

```
mds_coefs <- lm(abcde_prof_stand[,1]~
+  abcde.mds$points[,1]+abcde.mds$points[,2])$coefficients
for(j in 2:5) mds_coefs<-rbind(mds_coefs,
+  lm(abcde_prof_stand[,j]~
+  abcde.mds$points[,1]+abcde.mds$points[,2])$coefficients)
arrows(0,0,mds_coefs[,2],mds_coefs[,3], length=0.1, angle=10)
text(1.1*mds_coefs[,2:3], labels=colnames(abcde))
```

Assuming the sediment variable has been read (in character form) into the vector `sediment`, convert it to a factor.

```
sediment <- as.factor(sediment)
```

Plot positions of sediment categories in two different ways. The first way is to average the positions of the site points for each category, to show average clay, gravel and sand site:

```
sediment.means <- cbind(tapply(abcde.mds$points[,1],
+  sediment, mean),tapply(abcde.mds$points[,2], sediment, mean))
text(sediment.means, labels=c("C","G","S"))
```

The second way is to think of them as dummy variables to be predicted by the biplot dimensions, for example by logistic regression as in Chapter 3. They are first

converted to zero/one dummies and then their logistic coefficients are used to plot them as biplot axes:

```
clay01   <- sediment=="C"
gravel01 <- sediment=="G"
sand01   <- sediment=="S"

sediment_coefs <-
+  glm(as.numeric(clay01)~abcde.mds$points[,1]+
+  abcde.mds$points[,2],family="binomial")$coefficients
sediment_coefs <- rbind(sediment_coefs,
+  glm(as.numeric(gravel01)~abcde.mds$points[,1]+
+  abcde.mds$points[,2], family="binomial")$coefficients)
sediment_coefs <- rbind(sediment_coefs,
+  glm(as.numeric(sand01)~abcde.mds$points[,1]+
+  abcde.mds$points[,2], family="binomial")$coefficients)
segments(0, 0, sediment_coefs[,2], sediment_coefs[,3])
text(sediment_coefs[,2:3], labels=c("C","G","S"))
```

Chapter 5: Reduced-dimension Biplots

This is the code to produce the biplot of the 5 × 4 matrix of rank 2 which was used as an introductory example in Chapter 1 and which is plotted here using the SVD:

```
Y <- matrix(c(8,5,-2,2,4,2,0,-3,3,6,2,3,3,-3,-6,-6,-4,1,-1,-2),
+   nrow=5)
colnames(Y) <- c("A","B","C","D")
rowcoord <- svd(Y)$u %*% diag(sqrt(svd(Y)$d))
colcoord <- svd(Y)$v %*% diag(sqrt(svd(Y)$d))
plot(rbind(rowcoord,colcoord), type="n", asp=1)
abline(h=0, v=0, lty="dotted")
text(rowcoord, labels=1:5)
text(colcoord, labels=colnames(Y))
```

Chapter 6: Principal Component Analysis Biplots

The "attributes" data set in the data frame `MT_ratings` is centred—notice the `sweep` function to subtract the column means from each column:

```
MT_means <- apply(MT_ratings,2,mean)
MT_Y <- sweep(MT_ratings, 2, MT_means)
```

The equal row and column weights are applied and the singular value decomposition (SVD) calculated:

```
MT_Y <- MT_Y/sqrt(nrow(MT_Y)*ncol(MT_Y))
MT_SVD <- svd(MT_Y)
```

The form biplot (Exhibit 6.1), showing the rows in principal coordinates and the columns in standard coordinates, is computed and plotted as follows:

```
MT_F <- sqrt(nrow(MT_Y))*MT_SVD$u%*%diag(MT_SVD$d)
MT_G <- sqrt(ncol(MT_Y))*MT_SVD$v
plot(rbind(MT_F,MT_G), type="n", asp=1, xlim=c(-3.6,2.3))
text(MT_F, labels=rownames(MT_ratings))
arrows(0, 0, MT_G[,1], MT_G[,2], length=0.1, angle=10)
text(c(1.07,1.3,1.07,1.35,1.2,1.4)*MT_G[,1],
+   c(1.07,1.07,1.05,1,1.16,1.1)*MT_G[,2],
+   labels=colnames(MT_ratings))
```

Notice two aspects of the above code: first, the plot command again contains an explicit `xlim` option to extend the horizontal axis limits slightly, to accommodate the labels of the extreme points Germany and Morocco; and second, in the text command there are explicit scaling factors—obtained by trial and error—to position the attribute labels in the plot so that they do not overlap (this is generally done externally to R, by hand, to clean up the final version of the figure).

The covariance biplot (Exhibit 6.2), showing the rows in standard coordinates and the columns in principal coordinates, is similarly computed and plotted as follows:

```
MT_F <- sqrt(nrow(MT_Y))*MT_SVD$u
MT_G <- sqrt(ncol(MT_Y))*MT_SVD$v%*%diag(MT_SVD$d)
plot(rbind(MT_F,MT_G), type="n", asp=1, xlim=c(-3.6, 2.3))
text(MT_F, labels=rownames(MT_ratings))
arrows(0, 0, MT_G[,1], MT_G[,2], length=0.1, angle=10)
text(c(1.07,1.20,1.07,1.25,1.07,1.3)*MT_G[,1],
+    c(1.07,1.07,1.04,1.02,1.16,1.07)*MT_G[,2],
+    labels=colnames(MT_ratings))
```

The basic graphical part of the scree plot of Exhibit 6.3 is drawn as follows:

```
MT_percents<-100*MT_SVD$d^2/sum(MT_SVD$d^2)
MT_percents<-MT_percents[seq(6,1)]
barplot(MT_percents, horiz=T, cex.axis=0.7)
```


The data set "`USArrests`" is in the R base package so is obtained simply with the `data` command:

```
data(USArrests)
```

Columns 1, 2 and 4 of the data frame will be used. The weighted log-ratio biplot, (7.1) to (7.4) is performed as follows, where `rm` and `cm` are the row and column margins **r** and **c**, and `mr` and `mc` are weighted means used in the double-centring:

```
N    <- USArrests[,c(1,2,4)]
P    <- N/sum(N)
rm   <- apply(P, 1, sum)
cm   <- apply(P, 2, sum)
Y    <- as.matrix(log(P))
mc   <- t(Y) %*% as.vector(rm)
Y    <- Y - rep(1,nrow(P)) %*% t(mc)
mr   <- Y %*% as.vector(cm)
Y    <- Y - mr %*% t(rep(1,ncol(P)))
Z    <- diag(sqrt(rm)) %*% Y %*% diag(sqrt(cm))
svdZ <- svd(Z)
```

The biplot of the row principal and column standard coordinates (Exhibit 7.1) is obtained as follows, where the column coordinates are scaled down by 20 to make the plot more legible. As a consequence there are two scales on the plot, indicated on the left and at the bottom for the row points, and at the top and on the right for the column points—in this case we show how the two sets of scales can be colour coded to agree with their respective points:

```
# compute form biplot coordinates from results of SVD
USA_F <- diag(1/sqrt(rm)) %*% svdZ$u[,1:2] %*% diag(svdZ$d[1:2])
USA_G <- diag(1/sqrt(cm)) %*% svdZ$v[,1:2]
# biplot – axes with different scales plotted individually
plot(rbind(USA_F, USA_G/20), xlim=c(-0.35,0.45),
+  ylim=c(-0.18,0.23), asp=1, type = "n", xaxt="n", yaxt="n")
axis(1, col.axis="green", col.ticks="green")
axis(2, col.axis="green", col.ticks="green", at=seq(-0.2,0.2,0.2))
axis(3, col.axis="brown", col.ticks="brown", at=seq(-0.4,0.4,0.2),
+  labels=seq(-8,8,4))
axis(4, col.axis="brown", col.ticks="brown", at=seq(-0.2,0.2,0.2),
+  labels=seq(-4,4,4))
text(USA_F, labels = rownames(N), col = "green")
text(USA_G/20, labels = colnames(N), col = "brown")
```

The total variance of the data can be calculated either as the sum of squares of the elements of the decomposed matrix (`Z` in the above code) or as the sum of its squared singular values:

```
sum(Z*Z)
[1] 0.01790182
```

```
sum(svdZ$d^2)
[1] 0.01790182
```

The fish morphology example goes through in a similar way, assuming that the data frame `fish` contains the data, with the first two columns being the sex and habitat (see the description later of this analysis in the computations for Chapter 11). The remaining columns are the morphometric data, stored in `fish.morph`.

```
fish.morph <- fish[,3:ncol(fish)]
```

The only difference in the plotting in Exhibit 7.3 compared to the previous example is that the column standard coordinates are divided by 50, not 20, since these data have even less variance—the sum of squares of the corresponding `Z` matrix is:

```
sum(Z*Z)
[1] 0.001960883
```

Then, instead of fish identity codes, their sex × habitat are used as labels, stored in `fish.labels`—the first statement below computes numerical codes for the four sex × habitat groups:

```
fish.sexhab <- 2*(fish[,2]-1)+fish[,1]
fish.labels <- rep("fL", nrow(fish))
fish.labels[fish.sexhab=="2"] <- "mL"
fish.labels[fish.sexhab=="3"] <- "fP"
fish.labels[fish.sexhab=="4"] <- "mP"
```

The plot of the two log-ratios in Exhibit 7.4 is obtained as follows (notice how variables can be picked out of the data.frame `fish.morph` by name):

```
logFdlFal <- log(fish.morph[,"Fdl"] / fish.morph[,"Fal"])
logFdwFal <- log(fish.morph[,"Fdw"] / fish.morph[,"Fal"])
plot(logFdlFal,logFdwFal, asp=1, pch=24, xlab="log(Fdl/Fal)",
+    ylab="log(Fdw/Fal)")
abline(a=0.0107, b=0.707, lty=2)
```

The predicted values of variable *Fdw* (dorsal fin width) are computed and then compared graphically to their actual values as follows:

```
Fdw_pred <-
+   1.0108 * fish.morph[,"Fdl"]^0.707 * fish.morph[,"Fal"]^0.293
plot(Fdw_pred, fish.morph[,"Fdw"], xlim=c(18,30), ylim=c(18,30),
+   pch=24, xlab="predicted Fdw", ylab="actual Fdw")
```

```
abline(a=0, b=1, lty=2, col="brown")
# correlation between predicted and observed
cor(Fdw_pred, fish.morph[,"Fdw"])
[1] 0.7496034
```


For the calculations of CA, and later MCA, we shall tend to use the package **ca** in R. This has to be installed from the CRAN package library first, and then loaded into an R session:

```
library(ca)
```

The "smoking" data set is included in the ca package:

```
data(smoke)
```

The commands for performing CA from first principles, as described in (8.1) and (8.2), are:

```
N     <- smoke
P     <- N/sum(N)
rm    <- apply(P, 1, sum)
cm    <- apply(P, 2, sum)
Dr    <- diag(rm)
Dc    <- diag(cm)
Z     <- diag(sqrt(1/rm))%*%(as.matrix(P)-rm%*%t(cm))
+        %*%diag(sqrt(1/cm))
svdZ <- svd(Z)
```

For the asymmetric map of Exhibit 8.1 the row principal and column standard coordinates are:

```
smoke_F <- diag(1/sqrt(rm))%*%svdZ$u %*%diag(svdZ$d)
smoke_G <- diag(1/sqrt(cm))%*%svdZ$v
```

and can be plotted in the usual way.

However, using the **ca** package Exhibit 8.2 can be obtained in just one instruction:

```
plot(ca(smoke),map="rowprincipal", col=c("green","brown"))
```

The `plot` function here is actually the `plot.ca` function, automatically recognizing the `ca` object, and the `col` option now defines the colours of the row and column symbols.

The numerical results, including the contributions to inertia, are listed using the `summary` function:[9]

```
summary(ca(smoke))

Principal inertias (eigenvalues):

dim    value       %      cum%    scree plot
1      0.074759    87.8    87.8   ************************
2      0.010017    11.8    99.5   ***
3      0.000414     0.5   100.0
       --------   -----
Total: 0.085190   100.0

Rows:
     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |    SM |   57  893   31 |  -66  92   3 | -194 800 214 |
2 |    JM |   93  991  139 |  259 526  84 | -243 465 551 |
3 |    SE |  264 1000  450 | -381 999 512 |  -11   1   3 |
4 |    JE |  456 1000  308 |  233 942 331 |   58  58 152 |
5 |    SC |  130  999   71 | -201 865  70 |   79 133  81 |

Columns:
     name   mass  qlt  inr    k=1 cor ctr    k=2 cor ctr
1 |   non |  316 1000  577 | -393 994 654 |  -30   6  29 |
2 |   lgh |  233  984   83 |   99 327  31 |  141 657 463 |
3 |   mdm |  321  983  148 |  196 982 166 |    7   1   2 |
4 |   hvy |  130  995  192 |  294 684 150 | -198 310 506 |
```

Suppose that the "benthos" data set has been read into the data frame `benthos`. First perform the CA and calculate the row contributions to the two-dimensional biplot (note that the standard coordinates are stored in the `ca` object).

```
benthos.ca <- ca(benthos)
benthos.F <- benthos.ca$rowcoord %*% diag(benthos.ca$sv)
benthos.rowcon <- benthos.ca$rowmass * (benthos.F[,1]^2 +
+              benthos.F[,2]^2) / sum(benthos.ca$sv[1:2]^2)
```

Then set up a vector of species labels where those with contributions less than 1% are labelled "·".

9. See the paper online about the `ca` function, given in the bibliography, which describes the `ca` output.

```
benthos.names <- rownames(benthos)
benthos.names[benthos.rowcon<0.01] <- "·"
```

A nonlinear transformation is performed on the contributions above 1%, to be used for the character size of the labels.

```
benthos.rowsize <- log(1+exp(1)*benthos.rowcon^0.3)
benthos.rowsize[benthos.rowcon<0.01] <- 1
```

Exhibit 8.3 is plotted, with rows (species) in standard coordinates and columns (sites) in principal coordinates, with varying label sizes for the species.

```
FF <- benthos.ca$rowcoord
GG <- benthos.ca$colcoord %*% diag(benthos.ca$sv)
plot(rbind(FF,GG), type = "n", xlab "", ylab = "", asp=1)
text(FF[,1:2], labels = benthos.names, cex=benthos.rowsize)
text(GG[,1:2], labels = colnames(benthos))
```

The biplot showing point contributions is available in the **ca** package, using map option `"rowgreen"` or `"colgreen"` depending on which set is required in principal coordinates (in our case it would be the sites, or columns). The species labels are first substituted with those where the low contributing ones are replaced by "·".

```
benthos.ca$rownames <- benthos.names
```

Because rows are species (variables) and columns are sites (samples) the symbols need to be reversed—see `help(plot.ca)` in R for information about the plot option `pch` and execute the command `pchlist()` to get a list of plotting symbols (an alternative would be to transpose the data matrix from the start). For this biplot we also use the plot option `mass` to get symbols with sizes related to the species masses. The contribution biplot of Exhibit 8.4 is obtained using plot option `map="colgreen"`, which plots the columns in principal coordinates, as before, but the species (rows) in their contribution positions:

```
plot(benthos.ca, map="colgreen", mass=c(1,0), pch=c(17,24,16,1))
```

Lines are added connecting the origin with the species points, showing how their positions are computed as standard coordinates multiplied by the square roots of the species masses:

```
for(j in 1:nrow(benthos)) lines(
+ c(0,benthos.ca$rowcoord[j,1]*sqrt(benthos.ca$rowmass[j])),
+ c(0,benthos.ca$rowcoord[j,2]*sqrt(benthos.ca$rowmass[j])))
```

Alternatively, the `segments` function can be used, which automatically recycles the coordinates 0 and 0 in the following command:

```
segments(0, 0, benthos.ca$rowcoord[,1]*sqrt(benthos.ca$rowmass),
+              benthos.ca$rowcoord[,2]*sqrt(benthos.ca$rowmass))
```

Chapter 9: Multiple Correspondence Analysis Biplots I

The original data set "women" consists of the 2471 Spanish respondents in the 2002 ISSP survey on Family and Changing Gender Roles III, including their responses to the 8 substantive and 4 demographic variables listed in Chapter 9. In the supporting website it is shown how the concatenated matrix can be extracted from the original data. The simplest way is using a concept from Chapter 10 called the *Burt matrix*, and the most complicated is by converting all question responses first into zero—one dummy variables—both of these are explained in the next section on Chapter 10 . For the moment we assume that this matrix (part of which is shown in Exhibit 9.1), called `women.concat`, has already been computed, or input directly from an external source (the concatenated matrix itself is also provided on the website). The two categories *H4* and *H5* have also been combined into a category, labelled *H4,5*, so the matrix has 23 rows and 39 columns.

The symmetric CA of the concatenated matrix can be obtained using the default plot option in the **ca** package:

```
plot(ca(women.concat))
```

Depending on the version of the **ca** package (or any other software for correspondence analysis), an inversion of the axes might be obtained. The following code shows how the sign of the second axis is reversed, once the correspondence analysis object is saved:

```
women.ca <- ca(women.concat)
women.ca$rowcoord[,2] <- -women.ca$rowcoord[,2]
women.ca$colcoord[,2] <- -women.ca$colcoord[,2]
plot(women.ca)
```

Exhibit 9.2 was obtained by plotting the symbols first, and then adding the two sets of labels in different styles, also adding the inertias and their percentages on the axes:

```
plot(women.ca, labels=0)
women.F <- women.ca$rowcoord %*% diag(women.ca$sv)
women.G <- women.ca$colcoord %*% diag(women.ca$sv)
text(women.F, labels=women.ca$rownames, pos=4, offset=0.3)
text(women.G, labels=women.ca$colnames, pos=4, offset=0.3)
```

```
text(max(women.G[,1]), 0, "0.0571 (82.1%)", adj=c(0.6,-0.6))
text(0, max(women.G[,2]), "0.0030 (4.4%)", adj=c(-0.1,-3))
```

The map is plotted in a square window, which should then be pulled into the flatter shape of Exhibit 9.2 (the aspect ratio is not affected by this action). Having pulled the window into the desired shape, repeat the plotting so that the labels are properly positioned. Then the only difference between this result and Exhibit 9.2 is an adjustment of some of the overlapping labels, performed externally to R.

Exhibit 9.3 is plotted in a similar way, but using the plot option `map="rowprincipal"`

```
plot(women.ca, map="rowprincipal", labels=c(0,2))
```

Similarly, the contribution biplot in Exhibit 9.6, where the the standard coordinates are shrunk by the square roots of the category masses, is obtained with the plot option `map="rowgreen"` (here the `mass` option is also illustrated, to make the size of the column category symbols be related to their masses:

```
plot(women.ca, map="rowgreen", mass=c(F,T))
```

To add the supplementary points for sex (`m` = male, `f` = female) and age (`a1` to `a6`):

```
women.sex <- c(rep("m",6),rep("f",6))
women.age <- rep(c("a1","a2","a3","a4","a5","a6"),2)
women.sex.F <- cbind(tapply(women.F[12:23,1],women.sex,mean),
+                    tapply(women.F[12:23,2],women.sex,mean))
women.age.F <- cbind(tapply(women.F[12:23,1],women.age,mean),
+                    tapply(women.F[12:23,2],women.age,mean))
points(rbind(women.sex.F, women.age.F), pch=21)
text(rbind(women.sex.F, women.age.F),
+  labels=c("f","m","a1","a2","a3","a4","a5","a6"),
+  pos=4, offset=0.3)
```

Chapter 10: Multiple Correspondence Analysis Biplots II

The *Burt matrix* is a by-product of the `mjca` function in the **ca** package. If `women` contains the original response data, with the first 8 columns corresponding to the eight substantive questions *A* to *H*, then the Burt matrix is obtained as follows:

```
women.Burt <- mjca(women[,1:8])$Burt
```

If the categories *H4* and *H5* have not already been combined, then this can be done by combining the corresponding rows and columns of the Burt matrix:

```
women.Burt[,39] <- women.Burt[,39]+women.Burt[,40]
women.Burt[39,] <- women.Burt[39,]+women.Burt[40,]
women.Burt <- women.Burt[-40,-40]
rownames(women.Burt)[39] <- colnames(women.Burt)[39] <- "H4,5"
```

An alternative way to compute the Burt matrix, assuming that the "women" data set has been read as dummy variables into the data frame `women.Z` containing the indicator matrix of dummy variables for the 8 questions *A* to *H* (40 dummies if *H5* included, otherwise 39 if *H4* and *H5* have been combined). Then, as given in (10.2), the Burt matrix can be computed by premultiplying the indicator matrix by its transpose (notice the `as.matrix` commands if `women.Z` is a data frame, necessary for the multiplication):

```
women.Burt<- t(as.matrix(women.Z))%*%as.matrix(women.Z)
```

In the same way, the concatenated matrix `women.concat` can be obtained from the Burt matrix of all the variables, including the demographics, or via the indicator matrices. Suppose that `womenS.Z` contains the 2107×31 indicator matrix of dummy variables for the demographic variables (2 for sex, 5 for marital status, 6 for education, 6 for age, 12 sex-age combinations), then `women.concat` can be computed by premultiplying `women.Z` by the indicator matrix of dummy variables corresponding to sex, marital status, education and the sex-age combinations:

```
women.concat<- t(as.matrix(womenS.Z[,c(3:13,20:31)])) %*%
+                as.matrix(women.Z)
```

Alternatively, select just that part of the Burt matrix of all the variables, including the demographics, corresponding to the concatenated matrix:

```
women.concat <- mjca(women)$Burt[c(3:13,20:31),]
```

To obtain the total inertia of the Burt matrix, sum the squares of its singular values in the CA:

```
sum(ca(women.Burt)$sv^2)
[1] 0.677625
```

Then use (10.1) to calculate the adjusted total inertia:

```
(8/7)*(sum(ca(women.Burt)$sv^2)-(31/64))
[1] 0.2208571
```

Total inertia of the indicator matrix, calculated from the CA (from theory we know it will be equal to (39 – 8)/8 = 3.875:

```
sum(ca(women.Z)$sv^2)
[1] 3.875
```

Exhibit 10.3 can be obtained as follows (notice the change in the point symbols using the `pch` option, to get a smaller dot symbol for the cases, and the mass option to get triangle symbols related to the relative frequency of the category):

```
plot(ca(women.Z), map="rowprincipal", labels=c(0,2),
+    pch=c(149,1,17,24), mass=c(FALSE,TRUE))
```

To see how many of the singular values (i.e., square roots of principal inertias) in the analysis of the Burt matrix are larger than 1/8:

```
which(ca(women.Burt)$sv>(1/8))
[1] 1 2 3 4 5 6 7 8 9
```

So we apply the adjustment of (10.4) to the first 9 singular values:

```
(8/7)*(ca(women.Burt)$sv[1:9]-(1/8))
[1] 0.34219 0.23260 0.12415 0.11500 0.03451 0.02575 0.01489
+    0.00897 0.00681
```

To get parts of inertia explained on the axes, square these adjusted singular values and then express relative to the adjusted total inertia calculated previously:

```
(64/49)*(ca(women.Burt)$sv[1:9]-(1/8))^2/0.2208571
[1] 0.53017 0.24496 0.06979 0.05987 0.00539 0.00300 0.00100
+    0.00036 0.00021
```

Notice that the above parts do not add up to 1, since these 9 MCA axes cannot perfectly explain the total inertia in the off-diagonal tables: we would need to use joint correspondence analysis (JCA) to achieve this. Exhibit 10.4 is obtained by substituting the square roots of the adjusted inertias for the original ones in the CA of the Burt matrix:

```
women.Burt.ca     <- ca(women.Burt)
women.Burt.ca$sv  <- diag((8/7)*(ca(women.Burt)$sv[1:9]-(1/8)))
```

The website `http://www.multivariatestatistics.org` gives the full sets of instructions for plotting Exhibits 10.4 and 10.5. Exhibit 10.6 illustrates the computation of the contribution coordinates of the categories and including supplementary points (again, notice that axes may be inverted compared to Exhibit 10.6):

```
women.BurtS.ca    <- ca(rbind(women.Burt, women.concat),
+                    suprow=40:62)
women.BurtS.Gctr <- sqrt(women.BurtS.ca$colmass) *
+                    women.BurtS.ca$colcoord
women.BurtS.ca$colcoord <- women.BurtS.Gctr
women.BurtS.ca$sv[1:9] <- (8/7)*(women.BurtS.ca$sv[1:9]-(1/8))
plot(women.BurtS.ca, map="rowprincipal", what=c("none","all"),
+    labels=0, pch=c(20,1,24,24))
text(women.BurtS.Gctr, labels=women.Burt.ca$colnames, pos=2)
women.BurtS.Fsup <- women.BurtS.ca$rowcoord[40:62,] %*%
+    diag(women.BurtS.ca$sv)
points(women.BurtS.Fsup, pch=21)
text(women.BurtS.Fsup, labels=women.BurtS.ca$rownames[40:62],
+    pos=2)
```


From the log-ratio biplot of the morphometric data of the "morphology" data set in chapter 7 we know that the total variance is equal to 0.001961. We now want to aggregate the data into the four sex × habitat groups and measure how much variance is lost. The original data should not be aggregated since we are working on a logarithmic scale. It is equivalent, however, to aggregate the log-transformed data, or aggregate the rows of the double-centred matrix of log-transformed values. We choose the second way as an illustration, first repeating the initial steps of the log-ratio analysis algorithm (see computations for Chapter 7) on the matrix `fish.morph`:

```
N  <- fish.morph
P  <- N/sum(N)
rm <- apply(P, 1, sum)
cm <- apply(P, 2, sum)
Y  <- as.matrix(log(P))
mc <- t(Y) %*% as.vector(rm)
Y  <- Y - rep(1,nrow(P)) %*% t(mc)
mr <- Y %*% as.vector(cm)
Y  <- Y - mr %*% t(rep(1,ncol(P)))
```

The group masses are calculated:

```
fish.centroids.rm <- tapply(rm, fish[,3], sum)
```

and the four centroids, by weighted averaging of the corresponding rows of **Y**:

```
fish.centroids <- tapply(rm * Y[,1], fish[,3], sum)
for(j in 2:ncol(fish.morph)) fish.centroids
+      <- cbind(fish.centroids, tapply(rm * Y[,j], fish[,3], sum))
fish.centroids <- fish.centroids / fish.centroids.rm
```

Then the LRA algorithm continues for the four centroids, using the weighted SVD:

```
Z <- diag(sqrt(fish.centroids.rm)) %*% fish.centroids %*%
+  diag(sqrt(cm))
svdZ <- svd(Z)
# principal coordinates of centroids, standard coordinates of
+  variables
FF <- diag(1/sqrt(fish.centroids.rm)) %*% svdZ$u %*% diag(svdZ$d)
GG <- diag(1/sqrt(cm)) %*% svdZ$v
```

The inertia of the centroids:

```
inertia.centroids <- sum(Z*Z)
inertia.centroids
[1] 0.000128325
```

which is 6.5% of the total variance 0.001961 of the individual fish, computed in Chapter 7.

The biplot of the centroids and variables Exhibit 11.2 has a scaling factor difference of 50 between the two sets of points, as in Exhibit 7.3.

CA-DA, which is a CA of a concatenated table where a set of variables is cross-tabulated against a single grouping variable, is illustrated by Exhibit 11.3 for the "women" data set, using the marital status categories in the first five lines of the matrix `women.concat` (see Chapter 9). In this case we chose the contribution biplot:

```
women.ca_da <- ca(women.concat[1;5,])
women.ca_da$rownames <-  c("married", "widowed", "divorced",
+                          "separated", "single")
plot(women.ca_da, map="rowgreen")
```

(again, as explained in the computations of Chapter 9, if the axes are reversed compared to Exhibit 11.3, their coordinates can be multiplied by –1).

To see some basic results of the CA object: principal inertias, their percentages, the row and column masses, chi-square distances to the centroid, inertias, standard coordinates, etc, just type the object name:

```
women.ca_da
Principal inertias (eigenvalues):
            1         2         3         4
Value       0.029316  0.002915  0.002321  0.000993
Percentage  82.48%    8.2%      6.53%     2.79%

 Rows:
           married    widowed    divorced   separated      single
Mass      0.554817   0.081633   0.021357    0.032748    0.309445
ChiDist   0.080668   0.413108   0.320868    0.226113    0.213684
Inertia   0.003610   0.013931   0.002199    0.001674    0.014129
Dim. 1    0.403377   2.291677  -0.619362   -0.617888   -1.219648
Dim. 2   -0.656420   2.210670  -2.610218   -0.462015    0.822790

 Columns:
```

etc. … More detailed results can be obtained using the `summary` function, as explained before.

The "iris" data set is available in R:

```
data(iris)
```

The first four columns contain the variables and the fifth column contains the classification into the three groups: `"setosa"`, `"versicolor"` and `"virginica"` (there are 50 in each group). Read the data into `X` and calculate the means in `G`:

```
X <- iris[,1:4]
n <-nrow(X)
p <-ncol(X)
G <- apply(X[iris[,5]=="setosa",],2,mean)
G <- rbind(G,apply(X[iris[,5]=="versicolor",],2,mean))
G <- rbind(G,apply(X[iris[,5]=="virginica",],2,mean))
g <- nrow(G)
rownames(G) <- c("setosa", "versicolor", "virginica")
colnames(G) <- c("SepL","SepW","PetL","PetW")
colnames(X) <- c("SepL","SepW","PetL","PetW")
```

Calculate the three within-group covariance matrices (notice that we prefer the definition where the sum of squares is divided by n and not $n-1$, hence the slight adjustment by $(n-1)/n = 49/50$.

```
C1 <- (49/50)*cov(X[iris[,5]=="setosa",])
C2 <- (49/50)*cov(X[iris[,5]=="versicolor",])
C3 <- (49/50)*cov(X[iris[,5]=="virginica",])
```

The average within-grouped covariance matrix **C** is just the arithmetic average since the groups are the same size (otherwise it should be the weighted average —see (11.3)):

```
C <- (C1+C2+C3)/3
```

To calculate the inverse square root of **C**, calculate its SVD (or eigenvalue-eigenvector decomposition) and then use the inverse square roots of the singular values:

```
C.svd <- svd(C)
Cminushalf <- C.svd$u %*% diag(1/sqrt(C.svd$d)) %*% t(C.svd$v)
```

Calculate the matrix **S** of (11.4), its SVD and coordinates for the contribution biplot:

```
oneg <- rep(1,g)
Ig   <- diag(oneg)
S    <- diag(rep(sqrt(1/g),g)) %*% (Ig - (1/g)* oneg %*% t(oneg))
+           %*% G %*% Sminushalf * sqrt(1/ncol(G))
S.svd <- svd(S)
S.rpc <- sqrt(g) * S.svd$u %*% diag(S.svd$d)
S.cbp <- S.svd$v
```

Calculate the coordinates of the individual $n = 150$ irises as supplementary points according to (11.5):

```
onen <- rep(1,n)
In <- diag(onen)
S.rsup <- (In - (1/n)* onen %*% t(onen)) %*% as.matrix(X)
+         %*% Cminushalf %*% S.svd$v * sqrt(1/p)
```

Plot the groups and individual points in three colours:

```
plot(S.rsup, type = "n", asp=1)
text(S.rsup, labels = "·", col = c(rep("green",50),,
+  rep("violet",50) rep("brown",50)), cex=2, font = 2)
text(S.rpc, labels = rownames(G),
+  col = c("green","violet","brown"), font = 2, adj=c(0.5,0.5))
text(S.cbp, labels = colnames(G), col = "brown", cex=0.8, font = 2)
segments(0,0,S.cbp[,1],S.cbp[,2],col="brown")
```

Variance of the group means (i.e., between-group variance) is the sum of squares of the elements of the **S** matrix:

```
sum(S*S)
[1] 8.11933
```

The total variance of the points is obtained by calculating the equivalent matrix for the individuals in the Mahalanobis metric):

```
S <- sqrt(1/n) * (I - (1/n)* onen %*% t(onen)) %*% as.matrix(X)
+    %*% Cminushalf * sqrt(1/p)
sum(S*S)
[1] 9.11933
```

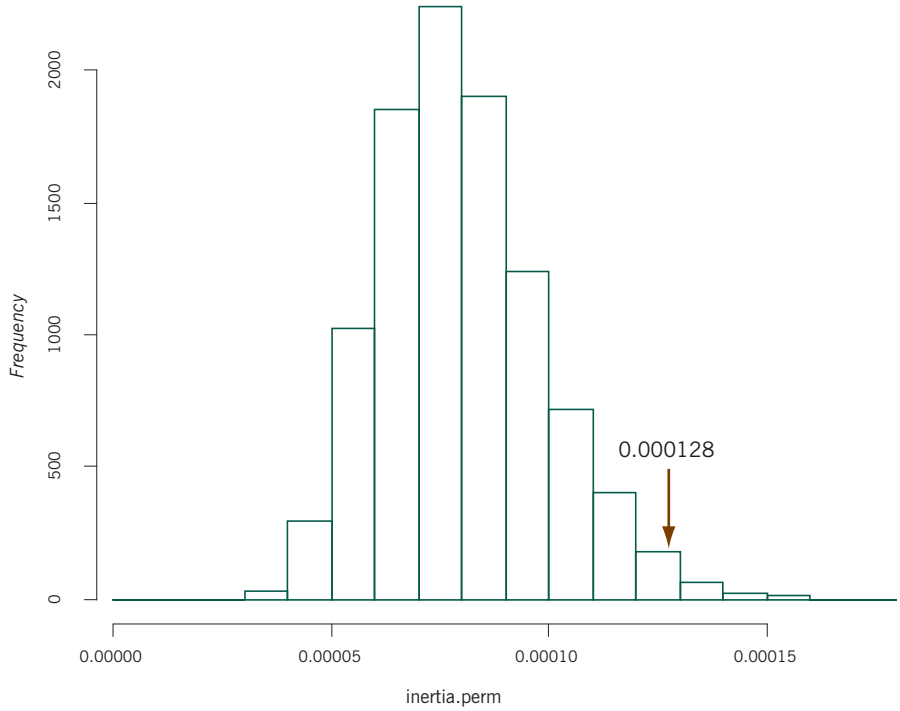
The difference between these two variance measures is exactly 1, which is the value of the within-group variance, by construction.

A Permutation Test

To test whether the between-group variance of 0.000128 in the above example of the four fish groups is significant, a permutation test consists in randomly shuffling the set of group labels assigned to the fish, recomputing the between-group variance each time (we reallocate the labels randomly 9,999 times) and seeing where the actual figure (which serves as the 10,000th permutation) lies in the permutation distribution. The following R code does the job, assuming that the same initial steps of the LRA algorithm are performed (see the 9 commands at the start of the computations for Chapter 11), ending with `Y` being the double-centred matrix of log-transformed data. Notice that the commands in the loop are just a repeat of the code previously used, but starting each iteration with a random sampling of the group labels, using the `sample` function. The initial `set.seed(317)` operation can be changed (or omitted) if you want a different set of random permutations.

```
set.seed(317)
inertia.perm <- rep(0,10000)
inertia.perm[1] <- inertia.centroids
for(iperm in 2:10000) {
  fish.perm<-sample(fish[,3])
  fish.centroids.rm <- tapply(rm, fish.perm, sum)
  fish.centroids <- tapply(rm * Y[,1], fish.perm, sum)
  for(j in 2:ncol(fish.morph)) fish.centroids
+      <- cbind(fish.centroids, tapply(rm * Y[,j], fish.perm, sum))
  fish.centroids <- fish.centroids / as.numeric(fish.centroids.rm)
  Z <- diag(sqrt(fish.centroids.rm)) %*% fish.centroids
+      %*% diag(sqrt(cm))
  inertia.perm[iperm] <- sum(Z*Z)
  }
```

Exhibit A.1:
Histogram of permutation distribution showing observed test statistic. The p*-value is the relative area of the distribution from the test statistic to the right*



To see where the value of 0.000128 lies in the permutation distribution:

```
which(sort(inertia.perm)==inertia.perm[1])
[1] 9847
```

so the number of permutations in the tail, including our observed value, is 154, which shows that it lies in the far upper tail of the distribution, with a *p*-value of 154/10,000 = 0.0154. A histogram of the permutation distribution, indicating the position of the observed value is shown in Exhibit A.1.

Chapter 12: Constrained Biplots

For the analysis of the fish morphometric data, we first add the body weight of the fish as a supplementary variable to the unconstrained log-ratio analysis. Again the nine commands at the start of the computations in Chapter 11 are repeated, up to the computation of the double-centred `Y`. Here we show the "column-principal" or "covariance" biplot, computing the SVD the usual way and then row standard and column principal coordinates:

```
Z    <- diag(sqrt(rm)) %*% Y %*% diag(sqrt(cm))
svdZ <- svd(Z)
```

```
FF    <- diag(1/sqrt(rm)) %*% svdZ$u
GG    <- diag(1/sqrt(cm)) %*% svdZ$v %*% diag(svdZ$d)
```

The body weight variable is standardized and regressed on the coordinates of the fish on the two dimensions of the morphometric log-ratio analysis. Since the fish are weighted according to their marginal totals, a weighted regression is performed, using the row masses in `rm` . Suppose that the body weight variable has been read into the vector `fish.weight`, then the R function `cov.wt` computes weighted means and variances:

```
fish.weight.mean  <- cov.wt(as.matrix(fish.weight),wt=rm)$center
fish.weight.var   <- cov.wt(as.matrix(fish.weight),wt=rm)$cov
fish.weight.stand <- (fish.weight-fish.weight.mean)/
+                     sqrt(fish.weight.var)
lm(fish.weight.stand~FF[,1]+FF[,2], weights=rm)$coefficients
  (Intercept)        FF[, 1]        FF[, 2]
-5.764505e-15   1.162973e-01   2.025847e-01
```

The coefficients 0.116 and 0.203 would define an arrow in the LRA biplot but only 5.5% of the variance of the body weight variable is explained, as can be seen by executing the `summary()` of the regression model above:

```
summary(lm(fish.weight.stand~FF[,1]+FF[,2], weights=rm))

(...)
Coefficients:
               Estimate   Std. Error     t value   Pr(>|t|)
(Intercept)  -5.765e-15    1.138e-01   -5.07e-14      1.0000
FF[, 1]       1.163e-01    1.138e-01       1.022      0.3101
FF[, 2]       2.026e-01    1.138e-01       1.781      0.0792
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1138 on 72 degrees of freedom
Multiple R-squared: 0.05531,    Adjusted R-squared: 0.02907
F-statistic: 2.108 on 2 and 72 DF,  p-value: 0.1290
```

To constrain the first axis to be perfectly correlated with body weight the definition of the projection matrix $\mathbf{Q}$ in (12.2) is particularly simple, because there is only a scalar to invert, not a matrix:

```
Q      <- diag(sqrt(rm)) %*% fish.weight.stand
+         %*% (1/((t(fish.weight.stand)%*%diag(rm)
```

```
+         %*%fish.weight.stand))) %*% t(fish.weight.stand)
+         %*% diag(sqrt(rm))
QZ     <- Q %*% Z
svdQZ  <- svd(QZ)
```

The orthogonal projection and corresponding SVD is:

```
QpZ <- Z - QZ
svdQpZ <- svd(QpZ)
```

The coordinates of the points are obtained from the first axis of the constrained analysis, and the first axis of the unconstrained one, and the body weight vector is obtained by weighted linear regression as before:

```
FF[,1] <- diag(1/sqrt(rm)) %*% svdQZ$u[,1]
GG[,1] <- diag(1/sqrt(cm)) %*% svdQZ$v[,1] * svdQZ$d[1]
FF[,2] <- diag(1/sqrt(rm)) %*% svdQpZ$u[,1]
GG[,2] <- diag(1/sqrt(cm)) %*% svdQpZ$v[,1] * svdQpZ$d[1]
fish.weight.coefs <- lm(fish.weight.stand~FF[,1]+FF[,2],
+                        weights=rm)$coefficients
```

After plotting the row and column points as before (again with two scales), body weight can be indicated by an arrow, as shown in Exhibit 12.1, as follows:

```
arrows(0, 0, 0.9*fish.weight.coefs[1], 0.9*fish.weight.coefs[2],
+      lwd=1.5, length=0.10, angle=15)
text(fish.weight.coefs[1], fish.weight.coefs[2], "weight")
```

The decomposition of the total variance into constrained and unconstrained parts:

```
sum(Z*Z)
[1] 0.001960883
sum(QZ*QZ)
[1] 7.860202e-05
sum(QpZ*QpZ)
[1] 0.001882281
100*sum(QZ*QZ)/sum(Z*Z)
[1] 4.008501
```

To perform a permutation test on the percentage of variance of the morphometric data explained by body weight, simply loop over the calculation of this percentage for random permutations of the body weight vector. The position of the

observed percentage of 4.01% in the sorted list of 10,000 permutations (9,999 plus the observed one) estimates the *p*-value:

```
set.seed(157)
bodyperm<-rep(0,10000)
total <- sum(Z*Z)
Q <- diag(sqrt(rm)) %*% fish.weight.stand %*%
+    (1/((t(fish.weight.stand) %*% diag(rm) %*% fish.weight.stand)))
+    %*% t(fish.weight.stand) %*% diag(sqrt(rm))
QZ <- Q %*% Z
bodyperm[1]<-100*sum(QZ*QZ)/total
# start permutations
for(iperm in 2:10000){
  fish.weight.stand.perm<-sample(fish.weight.stand)
  Q <- diag(sqrt(rm)) %*% fish.weight.stand.perm
  +  %*% (1/((t(fish.weight.stand.perm) %*% diag(rm)
  +  %*% fish.weight.stand.perm))) %*% t(fish.weight.stand.perm)
  +  %*% diag(sqrt(rm))
  QZ <- Q %*% Z
  bodyperm[iperm]<-100*sum(QZ*QZ)/total
  }
# find where the observed percentage is in the sorted list
which(sort(bodyperm)==bodyperm[1])
[1] 9991
```

The observed value is 10th from the top of the 10,000 values and the estimated *p*-value is thus 10/10,000 = 0.001.

For the final CCA of the data set "benthos", we illustrate the use of the function `cca` in the **vegan** package in R (this package needs to be installed separately). First read in the six environmental variables into the data frame `benthos_env`. Notice that this data set has the sites in the rows whereas benthos has the sites in the columns. For the `cca` function the sites need to be in the rows in both matrices, hence the use of `t(benthos)` in the code below. After log-transforming the environmental variables, the CCA is performed and the points and environmental variable arrows are plotted:

```
benthos_env <- log(benthos_env)
library(vegan)
benthos.cca <- cca(t(benthos), benthos_env)
plot(benthos.cca, display=c("lc","bp","sp"), type="n")
text(benthos.cca, display="bp", labels=colnames(benthos_env))
text(benthos.cca, display="sp", labels=rownames(benthos))
text(benthos.cca, display="lc", labels=colnames(benthos))
```

The plotting options chosen give the asymmetric biplot with sites in standard coordinates and species in principal coordinates (i.e., at weighted averages of the site points), and the environmental variables as biplot vectors using their regression coordinates on the CCA axes—these biplot coordinates are identical to the (weighted) correlation coefficients with the axes, since the site coordinates are standardized.

This plot is highly cluttered by all the species labels and so we can prune them down to the set of species that is most contributing to the display, as we have shown before in Chapter 8 for the same data set "benthos". The following code assigns "." labels to all species with less than a 1% contribution to the triplot:

```
benthos.cca.sp <- benthos.cca$CCA$v.eig
benthos.cca.spcon <- benthos.cca$colsum * (benthos.cca.sp[,1]^2 +
+  benthos.cca.sp[,2]^2) / sum(benthos.cca$CCA$eig[1:2])
benthos.names <- rownames(benthos)
benthos.names[benthos.cca.spcon<0.01] <- "·"
```

Plotting is then repeated as before, but substituting `benthos.names` for the original `rownames(benthos)` when species labels are plotted:

```
text(benthos.cca, display="sp", labels=benthos.names)
```

Further R scripts for the three case studies in Chapters 13 to 15 are given in the supporting website.

Biplot Software in R

The objective of this computational appendix is to educate readers in the use of R to construct biplots. Seeing and understanding the commands associated with specific figures in this book will assist users to perform their own analyses and biplots in the same way, as well as make them more proficient in R. Apart from these scripts, there are additional functions and one software package available for biplots.

The R functions `princomp` and `prcomp` for principal component analysis (PCA) both have a plotting function `biplot`. For example, `biplot(princomp(X))` draws the biplot of the PCA of the data matrix `X`. There are always two scales on the axes, one for the rows and one for the columns (similar to Exhibits 11.2, 12.1 and 12.5, for example). The `biplot` function (which is actually `biplot.princomp` or `biplot.prcomp` depending on which PCA function is used) has some scaling options for the axes:

`scaling=1`	form biplot (rows principal, columns standard)
`scaling=0`	covariance biplot (columns principal, rows standard)

(in fact, `scaling=alpha` produces a biplot where rows are scaled by the singular values to power `alpha` and the columns to power `1–alpha`, so `scaling=0.5` gives the symmetric biplot).

In addition there is a general biplot function `biplot` for plotting two given sets of points simultaneously.

The **ca** package described in Chapters 8 to 10 has several biplot options in the `plot.ca` function, although they are referred to as "maps". These are summarized below:

`map="rowprincipal"`	plots rows in principal, columns in standard coordinates
`map="colprincipal"`	plots columns in principal, rows in standard coordinates
`map="symbiplot"`	symmetric biplot, with row and column coordinates scaled by the square roots of the singular values on respective axes
`map="rowgreen"`	plots rows in principal, columns in contribution coordinates
`map="colgreen"`	plots columns in principal, rows in contribution coordinates

In addition, there are two options `"rowgab"` and `"colgab"`, due to Ruben Gabriel, who proposed multiplying the standard coordinates by the respective masses, whereas in the contribution biplots `"rowgreen"` and `"colgreen"` the square roots of the masses are used, which gives the specific interpretation in terms of contributions.

An R package **caGUI** is available as an interactive front end to the **ca** package. Finally, there is an interactive package **BiplotGUI** for R, mostly aimed at calibrated biplots, which were illustrated in Chapters 2 and 3 when introducing the biplot idea through regression and generalized linear models (see comments about calibrated biplots in the Epilogue).

Appendix B

Bibliography

This bibliography is not intended to be complete but rather gives the main literature and web resources about biplots so that the reader can continue to learn about this method.

The term "biplot" originates in Ruben Gabriel's *Biometrika* paper in 1971:

- Gabriel, K.R. (1971). The biplot graphic display of matrices with application to principal component analysis. *Biometrika* 58, 453–467.

This paper, which at the time of writing has 1008 citations on Google Scholar and 682 on the Science Citation Index (ISI Web of Knowledge), is widely regarded as the origin of the idea. It is worthwhile to repeat its abstract:

> "Any matrix of rank two can be displayed as a biplot which consists of a vector for each row and a vector for each column, chosen so that any element of the matrix is exactly the inner product of the vectors corresponding to its row and its column. If a matrix is of higher rank, one may display it approximately by a biplot of a matrix of rank two which approximates the original matrix. The biplot provides a useful tool of data analysis and allows the visual appraisal of the structure of large data matrices. It is especially revealing in principal component analysis, where the biplot can show inter-unit distances and indicate clustering of units as well as display variances and correlations of the variables."

A less cited paper by Ruben Gabriel, but nevertheless one of my favourite ones on the biplot, appeared the following year in the *Journal of Applied Meteorology* (Ruben was also well-known for his work as a statistician in weather modification projects):

- Gabriel, K.R. (1972). Analysis of meteorological data by means of canonical decompositions and biplots. *Journal of Applied Meteorology* 11, 1071–1077.

In this paper he gives the biplot associated with linear discriminant analysis, also known as canonical variate analysis. He also talks about the vectors linking pairs of variables in a biplot (like the "links" in log-ratio analysis).

Another gem is by Dan Bradu and Ruben Gabriel in *Technometrics* in 1978:

- Bradu, D. and Gabriel, K.R. (1972). The biplot as a diagnostic tool for models of two-way tables. *Technometrics* 20, 47–68.

In this paper they show how certain models lead to points lying in straight lines in the full space of the data, and thus approximately in a biplot that has a good fit to the data. Thus a subset of row points and/or column points lying in a straight line in a biplot suggest models in that submatrix of the data. In addition, orthogonality of the lines suggests a simpler model.

All the above papers are required reading for those interested in the origins of the technique.

Other authors also had the idea of adding variables to an existing configuration of points to make joint displays, although they did not call them biplots. For example, Doug Carroll's vector model for preferences is a biplot:

- Carroll, J.D. (1972). Individual differences and multidimensional scaling. In R.N. Shepard, A.K. Romney, and S.B. Nerlove, eds, *Multidimensional Scaling: Theory and Applications in the Behavioral Sciences (Vol. 1)*, 105–155. Seminar Press, New York.

Only one book exists to date specifically on the topic of biplots, by John Gower and David Hand:

- Gower, J.C. and Hand, D.J (1996). *Biplots*. Chapman & Hall, London, UK.

This book is very complete, both on linear and nonlinear biplots, giving a rigorous theoretical treatment of the subject. Another book by John Gower is with coauthors Sugnet Gardner-Lubbe and Niel le Roux:

- Gower, J.C., Gardner-Lubbe, S. and le Roux, N. (2010). *Understanding Biplots*. Wiley, Chichester, UK.

As far as the vast literature on the singular value decomposition (SVD) is concerned, I mention only two sources, by the author of one of the landmark algorithms for the SVD, Gene Golub in 1971, which seems to be an important year for the biplot:

- Golub, G.H. and Reinsch, C. (1971). The singular value decomposition and least squares solutions. In J.H. Wilkinson and C. Reinsch, eds, *Handbook for Automatic Computation*, 134–151. Springer-Verlag, Berlin.

and the other a classic book by Paul Green and Doug Carroll, originally published in 1976, which was the first time I saw the geometric interpretation of the SVD (called "basic structure" by these authors)—this book is invaluable as a practical introduction to matrix and vector geometry in multivariate analysis:

- Green, P.E. and Carroll, J.D. (1997). *Mathematical Tools for Applied Multivariate Analysis, Revised Edition.* Academic Press, New York.

Most books or articles that treat the methods presented in this book will have a section or chapter on biplots and their interpretation in the context of that method. This is just a tiny selection of some of the literature that can be consulted, and by no means the primary references:

Principal component analysis

- Joliffe, I.T. (2002). *Principal Component Analysis* (2nd edition). Springer, New York.

Log-ratio analysis (unweighted form)

- Aitchison, J. and Greenacre, M. (2002). Biplots of compositional data. *Applied Statistics* 51, 375–392.

Log-ratio analysis (weighted form)

- Greenacre, M. and Lewi, P.J. (2009). Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio scale measurements. *Journal of Classification* 26, 29–54.

Correspondence analysis

- Greenacre, M. (2007). *Correspondence Analysis in Practice* (2nd edition). Chapman & Hall/CRC, London. Spanish translation may be freely downloaded at: `http://www.fbbva.es` and `http://www.multivariatestatistics.org`

Multiple correspondence analysis

- Greenacre, M. and Blasius, J., eds (2006). *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC Press, London.
- Michalidis, G. and de Leeuw, J. (1998). The Gifi system for descriptive multivariate analysis. *Statistical Science* 13, 307–336.

Discriminant analysis/centroid biplots

- Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning* (2nd edition). Springer, New York. This book may be freely downloaded at `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

Constrained biplots

- Legendre, P. and Legendre, L. (1998). *Numerical Ecology* (2nd edition). Elsevier, Amsterdam.

Finally we give some resources on the internet, on R packages relevant to this book (in alphabetic order of package names).

- Thioulouse, J. and Dray, S. (2007). Interactive multivariate data analysis in R with the **ade4** and **ade4TkGUI** packages. *Journal of Statistical Software*. Download from `http://www.jstatsoft.org/v22/i05/paper`

- De Leeuw, J. and Mair, P. (2009). Simple and canonical correspondence analysis using the R package **anacor**. *Journal of Statistical Software*. Download from `http://www.jstatsoft.org/v31/i05/paper`

- De Leeuw, J. and Mair, P. (2009). Gifi methods for optimal scaling in R: the package **homals**. *Journal of Statistical Software*. Download from: `http://www.jstatsoft.org/v31/i04/paper`

- La Grange, A., le Roux, N. and Gardner-Lubbe, S. (2000). **BiplotGUI**: Interactive biplots in R. *Journal of Statistical Software*. Download from: `http://www.jstatsoft.org/v30/i12/paper`

- Nenadić, O. and Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The **ca** package. *Journal of Statistical Software*. Download from `http://www.jstatsoft.org/v20/a03/paper`

- Markos, A. (2010). **caGUI**: a Tcl/Tk GUI for the functions in the **ca** package. Download from `http://cran.r-project.org/web/packages/caGUI/index.html`

- Graffelman, J. (2010). **calibrate**: Calibration of scatterplot and biplot axes. Download from `http://cran.r-project.org/web/packages/calibrate/index.html`

- Oksanen, J. (2010). **vegan**: Community Ecology Package. Download from `http://cran.r-project.org/web/packages/vegan/index.html`

And some relevant websites:

`http://www.multivariatestatistics.org`
: Supporting website for the series of statistics books published by the BBVA Foundation, including the present book *Biplots in Practice,* with glossary of terms and chapter summaries in Spanish, as well as supplementary material such as animated graphics and links to the data sets and R code.

`http://www.carme-n.org`
Correspondence Analysis and Related Methods Network, with R scripts and data from this book, from *Correspondence Analysis in Practice*, Second Edition, and from *Multiple Correspondence Analysis and Related Methods*.

`http://gifi.stat.ucla.edu`
Jan de Leeuw's website for the Gifi system (centred around multiple correspondence analysis and related methods) and R functions

`http://www.imperial.ac.uk/bio/research/crawley/statistics`
Michael Crawley's material from his book *Statistics: an Introduction Using* R

`http://www.issp.org`
Source of many data sets from the International Social Survey Program

`http://www.r-project.org`
The R project for statistical computing

`http://cc.oulu.fi/~jarioksa/softhelp/vegan.html`
Jari Oksanen's website for the **vegan** package in R, a very complete package which includes PCA, CA, CCA and many more multivariate methods, as well as permutation tests.

`http://people.few.eur.nl/groenen/mmds/datasets`
Website with data sets from book *Modern Multidimensional Scaling* by Ingwer Borg and Patrick Groenen

`http://biplot.usal.es/ClassicalBiplot/index.html`
Website of José Luis Vicente Villardon's biplot software for biplots and simple correspondence analysis

APPENDIX C

Glossary of Terms

In this appendix an alphabetical list of the most common terms used in this book is given, along with a short definition of each. Words in italics refer to terms which are contained in the glossary.

— *adjusted principal inertias*: a modification of the results of a *multiple correspondence analysis* that gives a more accurate and realistic estimate of the inertia accounted for in the solution.

— *aspect ratio:* the ratio between a unit length on the horizontal axis and a unit length on the vertical axis in a graphical representation; should be equal to 1 for any *biplot* or *map* that has a spatial interpretation.

— *asymmetric biplot/map:* a joint display of the rows and columns where the two clouds of points have different scalings (also called *normalizations*), usually one in *principal coordinates* and the other in *standard coordinates*.

— *biplot:* a joint display of points representing the rows and columns of a table such that the *scalar product* between a row point and a column point approximates the corresponding element in the table in some optimal way.

— *biplot axis:* a line in the direction of a *biplot vector* onto which points can be projected in order to estimate values in the table being analysed.

— *biplot vector:* a vector drawn from the origin to a point in a *biplot*, often representing a variable of the data matrix, or a *supplementary variable*.

— *bootstrapping:* a computer-based method of investigating the variability of a statistic, by generating a large number of replicate samples, with replacement, from the observed sample.

— *Burt matrix:* a particular matrix of *concatenated tables*, consisting of all two-way cross-tabulations of a set of categorical variables, including the cross-tabulations of each variable with itself.

— *calibration:* the process of putting a scale on a *biplot axis* with specific tic-marks and values.

— *canonical correspondence analysis (CCA):* extension of *correspondence analysis* to include external explanatory variables; the solution is constrained to have dimensions that are linearly related to these explanatory variables.

— *centroid:* weighted average point.

— *chi-square distance:* weighted *Euclidean distance* measure between *profiles*, where each squared difference between profile elements is divided by the corresponding element of the average profile; the distance function used in *correspondence analysis*.

— *classical scaling:* a version of *multidimensional scaling* which situates a set of points in multidimensional Euclidean space, based on their interpoint distances or dissimilarities, and then projects them down onto a low-dimensional space of representation.

— *concatenated table:* a number of tables, usually based on cross-tabulating the same individuals, joined together row-wise or column-wise or both.

— *contingency table:* a cross-tabulation of a set of cases or objects according to two categorical variables; hence the grand total of the table is the number of cases.

— *contribution to variance/inertia:* component of variance (or *inertia*) accounted for by a particular point on a particular *principal axis*; these are usually expressed relative to the corresponding *principal variance* (or *principal inertia*) on the axis (giving a diagnostic of how the axis is constructed) or relative to the variance (or *inertia*) of the point (giving a measure of how well the point is explained by the axis).

— *contribution (or standard) biplot: biplot* in which one set of points, usually the variables, is normalized so that their squared coordinates are the parts of variance (or *inertia*) on the respective axes; this scaling facilitates the interpretation of the solution space.

— *correspondence analysis (CA):* a method of displaying the rows and columns of a table as points in a spatial map, with a specific geometric interpretation of the positions of the points as a means of interpreting the similarities and differences between rows, the similarities and differences between columns and the association between rows and columns. Fundamental concepts in the definition of CA are those of *mass* and *chi-square distance*.

— *covariance biplot: asymmetric biplot* where the variables (usually columns) are in *principal coordinates*, thus approximating the covariance structure of the vari-

ables (i.e., lengths of *biplot vectors* approximate standard deviations, angle cosines between biplot vectors approximate correlations), and the rows (usually cases) in *standard coordinates*.

— *dimension:* in the context of biplots, a synonym for axis.

— *dimensionality:* the number of dimensions inherent in a table needed to reproduce its elements exactly in a *biplot* or *map*; in this context it is synonymous with the *rank* of the matrix being analyzed.

— *dimension reduction*: the action of finding fewer *dimensions* than the *dimensionality* of a matrix, which can reproduce the matrix optimally.

— *dissimilarity:* a measure of difference between objects which is like a *distance* but does not satisfy the *triangular inequality*.

— *distance:* a measure of difference between pairs of objects which is always positive or zero, and zero if and only if the objects are identical, and furthermore satisfies the *triangular inequality*.

— *double centring:* an operation applied to a data matrix which first subtracts the row means from each row of the matrix, and then subtracts the columns means from each column of the row-centred matrix. Often the centring includes weights on the rows and the columns (e.g., the *masses* in *correspondence analysis* and *log-ratio analysis*). The (weighted) means of the rows and the columns of a double-centred matrix are all 0.

— *dual biplots*: a pair of *asymmetric biplots* which are versions of the same *singular-value decomposition*; usually, the allocation of the singular values to the left or right matrix of singular vectors is what distinguishes the pair.

— *dummy variable:* a variable that takes on the values 0 and 1 only; used in one form of *multiple correspondence analysis* to code multivariate categorical data.

— *eigenvalue:* a quantity inherent in a square matrix, forming part of a decomposition of the matrix into the product of simpler matrices. A square matrix has as many eigenvalues and associated eigenvectors as its rank; in the context of *biplots*, eigenvalue is a synonym for the *principal variance* or *principal inertia*.

— *Euclidean distance:* distance measure between vectors where squared differences between corresponding elements are summed, followed by taking the square root of this sum.

— *form biplot: asymmetric biplot* where the cases (usually rows) are in *principal coordinates*, thus approximating distances between cases, and the columns (usually variables) in *standard coordinates*.

— *generalized linear model (GLM):* a generalization of linear regression, where there are several possible transformations of the mean of the response variable and several possible choices of the conditional probability distribution; examples are Poisson regression (transformation: logarithm; probability distribution: Poisson) and logistic regression (transformation: logit, or log-odds; probability distribution: binomial).

— *generalized linear model biplot:* similar to the *regression biplot* , except that the coefficients are obtained through a *generalized linear model*; hence any *calibrations* of the *biplot axes* are not at equal intervals as in regression biplots, but reflect the transformation of the mean of the corresponding response variable.

— *gradient:* in optimization theory, the vector of partial derivatives of a multivariable function with respect to its variables, indicating the direction of steepest ascent of the function; when the function is linear (e.g., a regression equation), then the gradient is simply the vector of coefficients of the variables.

— *indicator matrix:* the coding of a multivariate categorical data set in the form of *dummy variables*.

— *inertia:* weighted sum of squared distances of a set of points to their *centroid*; in *correspondence analysis* the points are *profiles*, weights are the *masses* of the profiles and the distances are *chi-square distances*.

— *interactive coding:* the formation of a single categorical variable from all the category combinations of two or more categorical variables.

— *joint correspondence analysis (JCA):* an adaptation of *multiple correspondence analysis* to analyse all unique two-way cross-tabulations of a set of categorical variables (contained in the *Burt matrix*) while ignoring the cross-tabulations of each variable with itself.

— *left matrix:* the first matrix in the decomposition of the *target matrix*, which provides the coordinates of the rows in a *biplot*.

— *linear discriminant analysis:* a dimension-reduction method which aims to optimally separate the *centroids* of groups of multivariate points, using the *Mahalanobis distance* to define distances between points.

— *link vector:* the vector in a *biplot* that joins two points and thus represents the difference vector between the two (e.g., the difference between two variables).

— *log-ratio:* given two elements in the same row or same column of a strictly positive data matrix, this is the logarithm of the ratio of the values.

— *log-ratio analysis:* a dimension-reduction method for a table of strictly positive data all measured on the same scale, based on log-transforming the data and *double-centring* before decomposing by the *singular value decomposition*. The rows and columns are preferably weighted, usually proportional to the margins of the table. Log-ratio analysis effectively analyzes all *log-ratios* in the rows and the columns of the table.

— *log-ratio distance:* the distance function underlying *log-ratio analysis*, based on the differences between all *log-ratios* in the rows or in the columns.

— *Mahalanobis distance:* a distance function used in *linear discriminant analysis*, which aims to de-correlate and standardize the variables within each of the groups being separated.

— *map:* a spatial representation of points with a *distance* or *scalar product* interpretation.

— *mass:* a weight assigned to a point; in *correspondence analysis* and *log-ratio analysis*, the row and column masses are the marginal totals of the table, divided by the grand total of the table.

— *monotonically increasing function:* a function that steadily increases as its argument increases; that is, its derivative (or slope) is always positive.

— *multidimensional scaling (MDS):* the graphical representation of a set of objects based on their interpoint *distances* or *dissimilarities*.

— *multiple correspondence analysis (MCA):* for more than two categorical variables, the *correspondence analysis* of the *indicator matrix* or *Burt matrix* formed from the variables.

— *nested principal axes:* a property of a *biplot* or *map* where solutions consist of a set of uncorrelated principal axes which combine in an ordered way: for example, the best three-dimensional solution consists of the two axes of the best two-dimensional solution plus the third axis.

— *normalization:* refers to the scale of a variable or a principal axis in terms of its variance; for example, a variable divided by its standard deviation is normalized to have variance 1, while the *principal coordinates* of a set of points on a particular axis have a normalization equal to the *eigenvalue* (*principal variance* or *principal inertia*) of that axis.

— *permutation test:* generation of data permutations, either all possible ones or a large random sample, assuming a null hypothesis, in order to obtain the null distribution of a test statistic and thus estimate the *p*-value associated with the observed value of the statistic.

— *principal axis:* a direction of spread of points in multidimensional space that optimizes the variance or *inertia* displayed; can be thought of equivalently as an axis which best fits the points in a least-squares sense, often weighted.

— *principal component analysis (PCA):* a method of dimension reduction which attempts to explain the maximum amount of variance in a data matrix in terms of a small number of *dimensions*, or components.

— *principal coordinates:* coordinates of a set of points projected onto a *principal axis*; the (weighted) sum of squared coordinates of the points along an axis equals the *principal inertia* on that axis.

— *principal inertia (or principal variance):* the *inertia* (or variance) displayed along a *principal axis*; often referred to as an *eigenvalue*.

— *profile:* a row or a column of a table divided by its total; the profiles are the points visualized in *correspondence analysis*.

— *projection:* given a point in a high-dimensional space, its projection onto a low-dimensional subspace refers to that point closest to the original point; the action of projection is usually perpendicular to the subspace.

— *projection matrix:* a matrix which when multiplied by a vector gives the projection of that vector on a low-dimensional subspace.

— *rank:* the rank of a matrix in a geometric context is the number of *dimensions* needed to reproduce the matrix exactly.

— *redundancy analysis (RDA):* extension of *principal component analysis* to include external explanatory variables; the solution is constrained to have *dimensions* that are linearly related to these explanatory variables.

— *regression biplot:* a *biplot* which has as its system of display axes a set of explanatory variables (in the simplest case, two variables), showing firstly a set of case points in terms of these variables and secondly a set of *biplot vectors* with coordinates defined by regression coefficients from the respective linear regressions of response variables on the explanatory variables. If the axes are standardized, then the biplot vectors are defined by the standardized regression coefficients.

— *right matrix:* the second matrix in the decomposition of the *target matrix*, which provides the coordinates of the columns in a *biplot.*

— *scalar product:* for two point vectors, the product of their lengths multiplied by the cosine of the angle between them; directly proportional to the projection of one point on the vector defined by the other.

— *scree plot:* a bar chart of the set of *eigenvalues* (*principal variances* or *inertias*) associated with a *biplot,* in descending order of magnitude.

— *simplex:* a triangle in two dimensions, a tetrahedron in three dimensions, and generalizations of these geometric figures in higher dimensions; in *correspondence analysis J*-element *profiles* lie inside a simplex defined by *J vertices* in $(J-1)$-dimensional space.

— *singular value decomposition (SVD):* the decomposition of a matrix into the product of three matrices with simple structure: the matrix of left singular vectors multiplied by the diagonal matrix of singular values (all positive and in descending order) multiplied by the transposed matrix of right singular vectors. The SVD is the natural generalization of the *eigenvalue*–eigenvector decomposition, but applicable more generally to rectangular matrices.

— *standard coordinates:* coordinates of a set of unit points projected onto *principal axes*—their (weighted) sum of squares along an axis equals 1.

— *standardized regression coefficient:* a regression coefficient that corresponds to a variable that has been normalized to have variance (or *inertia*) 1.

— *subset correspondence analysis:* a variant of *correspondence analysis* which allows subsets of rows and/or columns to be analysed, while maintaining the same *chi-square distance* function and point *masses* as for the full table.

— *supplementary point:* a point which has a position (e.g., a vector of data in principal component analysis or a *profile* in correspondence analysis) with *mass* set

equal to zero; in other words, a supplementary point is displayed on the map but has not been used in the construction of the map.

— *supplementary variable:* a variable which is positioned in a map by (weighted) least-squares regression on the *principal axes*; the variable is usually depicted as a vector with coordinates equal to the regression coefficients.

— *symmetric map:* a simultaneous display of the *principal coordinates* of the rows of a matrix and the *principal coordinates* of its columns. While the distance geometry of both rows and columns is shown, this map is not a *biplot*, but approximates one if the *eigenvalues* of the axes are not too different.

— *target matrix:* a matrix which is decomposed into the product of two matrices, the *left* and *right matrices*, which provide the coordinates for the rows and columns respectively in a *biplot*.

— *transition formula:* the relationship between the row points and column points in a *map* or a *biplot*.

— *triangular inequality:* a property of a true distance function whereby the distance between two objects is necessarily less than or equal to the sum of the distances from the two objects to a third one.

— *triplot:* a *biplot* showing, in addition, a third set of points or vectors corresponding to the explanatory variables which constrain the solution, for example in *canonical correspondence analysis* or *redundancy analysis*.

— *vertex:* a unit point in multidimensional space, with all elements zero except one with value 1, usually a unit *profile* in CA which delimits the *simplex* within which the points in CA lie.

— *weighted Euclidean distance:* similar to *Euclidean distance*, but with a positive weighting factor for each squared difference term.

APPENDIX D

Epilogue

Up to now this book has presented known facts about the theory and practice of biplots. In this final section I give my personal opinions about biplots and their use in practice. For, as the title of the book declares, this book is mainly about the practice and indeed the usefulness of this method as a research tool. I start off with a reflection of what the term "biplot" means and then treat some specific aspects which have a greater or lesser repercussion when it comes to practical applications.

What constitutes a biplot?

In my understanding of the term, a biplot is a representation of the rows and the columns of a data matrix in a joint display, with few dimensions, usually two-dimensional but nowadays possibly three-dimensional when viewed with special software such as R's **rgl** package. Because of the orthogonality and nested property of the principal axes of a biplot, further dimensions can be studied separately, for example, by considering axes 3 and 4 of the solution space in a planar display (see Exhibit 14.11, for example, where different planar projections were displayed). The essential feature of the biplot is that it displays scalar products between row and column points of a target matrix (the data matrix, appropriately centred and normalized), according to the fundamental result in Gabriel's original paper, formulated in Chapter 1 as:

$$\textit{target matrix} = \textit{left matrix} \cdot \textit{right matrix}$$

(see also the abstract from Ruben Gabriel's original article, which is reproduced in the Bibliography). The left and right matrices of low rank (dimensionality), obtained conveniently from the singular value decomposition (SVD), provide respective row and column coordinates that are used to plot the rows and columns as points or vectors as the case may be.

Let us suppose for convenience of description that the rows have been plotted as points and the columns as vectors drawn from the origin of the display. The idea of plotting columns as vectors gives the idea that each column has been regressed on the axes of the biplot and the vector actually represents the regression plane (or hyperplane for a three-dimensional biplot). This plane is defined uniquely by the vector that indicates the direction of steepest ascent of the plane, that is the gradient vector with elements equal to the regression coefficients. Since contours

(or isolines) of a plane are at right-angles to this gradient vector, estimated values of each row for that column can be obtained by projecting them onto the biplot axis through the vector.

These are thus the basic properties of a biplot. Variations exist, for example the row points could be approximating a particular interpoint distance, or they could be standardized along principal axes. Then there are nonlinear transformations of the data, which induce biplot axes with nonlinear scales, or weighting of the points when determining the solution space, but the biplot basically results in two sets of points, one of which is optionally drawn as a set of vectors.

Calibration of biplot axes

The idea of calibrating biplot axes adds understanding about how the biplot works and how to interpret it. As shown in Chapters 2 and 3, adding tic marks to a biplot axis that passes through a biplot vector and then calibrating the axis according to the original scale of the corresponding variable, gives insight to what the biplot vector actually represents. But, when a user has digested this fact once and for all, I can see no practical purpose in leaving the calibrations on the final analyses reported in research findings. Firstly, it is only possible to include calibrations when biplot axes are few, which is seldom the case apart from small educational examples; secondly, they clutter up the display and detract from its simplicity; and thirdly, they do not allow one to define a meaningful length of the biplot vector, which can be a useful aspect of the biplot's interpretation (see, for example, the contribution biplot, discussed again below).

Good biplot design

The comments above about calibrated biplots lead naturally into the subject of what constitutes good graphical design for a biplot display. The objective should be to include as little as possible, but enough for the user to make a correct interpretation of what is presented. Overloading the biplot with calibrations, for example, for every biplot axis is not necessary, since it is known that the centre of the biplot represents the (weighted) average of each variable (data matrices are almost always centred) and all one needs to see is how the points line up on a biplot axis above and below their average. Knowing the scale of the biplot is relevant, but the tic marks and calibrations on the principal axes should be few and discrete; sometimes we have added principal variances (or inertias) and their percentages to axes, or simply mentioned these in the text or caption. Including what Tufte calls "chartjunk" just increases the biplot's ink-to-information ratio unnecessarily, but things like coloured labels, symbols of different sizes and textures are all useful instruments for communicating more about the plot. Omitting labels of uninteresting points is also a useful strategy. Consider the two versions of the same correspondence analysis solution in Exhibit D.1, performed on a table of fatty acid compositions for a sample of fish. The upper biplot is the asymmetric one with row points (fish, displayed by dots) in principal coordinates and column

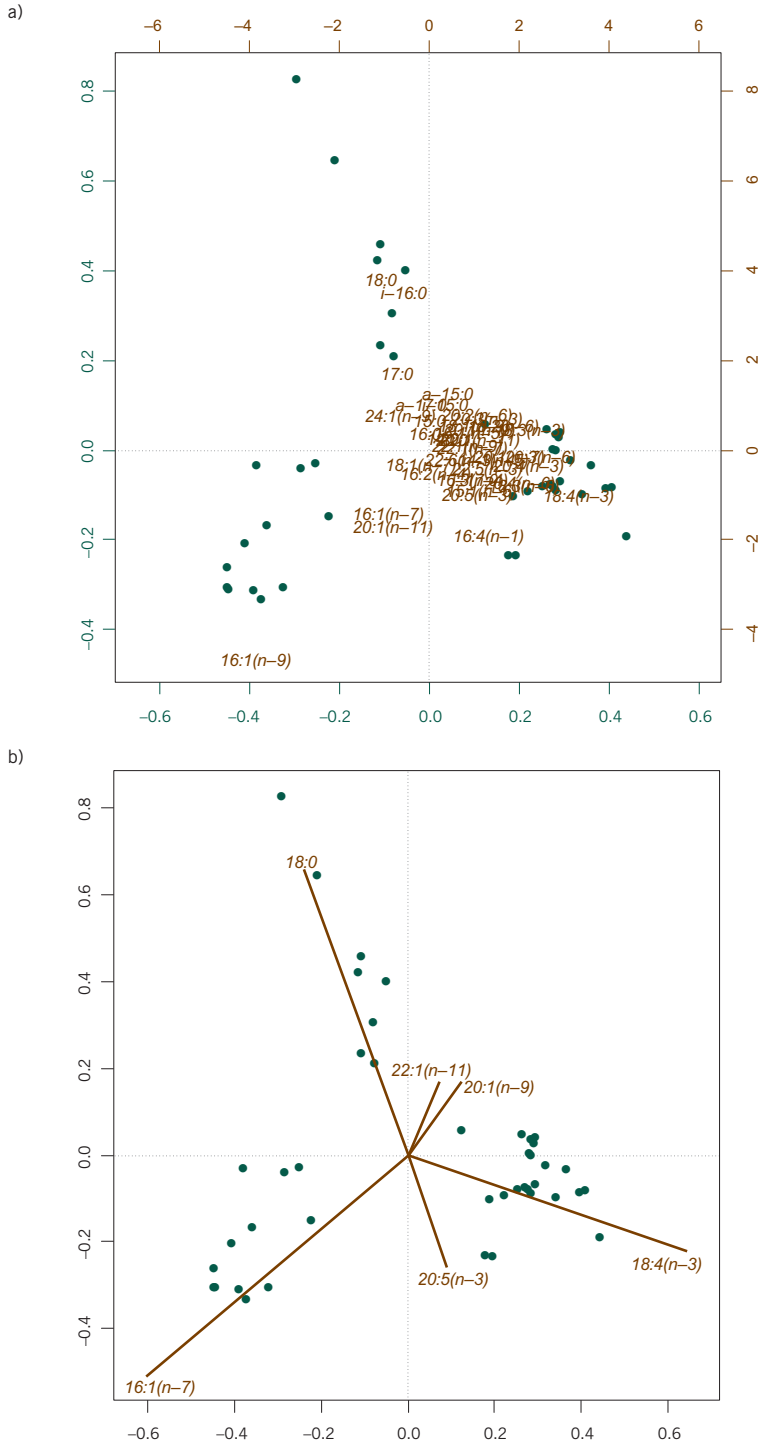


Exhibit D.1:

Two versions of the same correspondence analysis of a data set of compositions of 40 fatty acids in 42 fish: on top, the asymmetric ("rowprincipal") biplot, including all the points; at the bottom, the contribution biplot, with only the most contributing fatty acids shown

points in standard coordinates, and because of the very low inertia in these data two scales are necessary. This biplot is very cluttered with points and the fact that there are three groups of fish is partly obscured; also, it would be even worse if we added the vectors to the variable points (the fatty acids). The lower biplot is the contribution biplot, which only needs one scale and where only fatty acids are displayed that contribute more than average to the principal axes, which gives a much cleaner and easier to interpret solution. Only six out of the original 40 fatty acids remain and it is seen clearly that there are mainly three fatty acids, each associated with one of the three groups. Furthermore, the fatty acid *16:1(n–7)* which is responsible for characterizing the group of fish at bottom left is not at all clear in the upper biplot, where one might have thought that the most important one was *16:1(n–9).* Since these six fatty acids are the major contributors, the biplot would remain almost the same if the other 34 fatty acids were removed from the data set and a subset correspondence analysis performed.

Quality of a biplot

The dimension-reduction step is necessary to be able to visualize a high-dimensional data set in a few dimensions, but variance (or inertia) is lost in the process. There is a measure of variance retained and variance lost, in raw amounts or percentages, and also numerical diagnostics for how much variance of each point, row or column, is retained in the solution and how much is lost. On the other hand, the solution space is determined by different rows and columns in varying amounts. A point may be displayed accurately in a biplot, with little variance lost, but it could have played almost no role in determining the solution (the reverse is not true: points that generally determine the solution are usually quite accurately displayed). The contributions, both of the solution to the variances of individual points and of the points to the solution space, are important numerical diagnostics that support the interpretation of the biplot.

The contribution biplot

The idea of incorporating the contribution of a column (for example) into the length of its corresponding vector is, in my opinion, one of the most important variations of the biplot display. Users have difficulty in deciding which vectors are important for the interpretation of the biplot, so by rescaling individual vectors to correspond to their part contributions to the principal axes, the important vectors are immediately made more evident to the user, since their lengths along principal axes are the longest.

So why do we not always use the contribution biplot? The answer is, very simply, that in gaining this property of the interpretation, another property is inevitably lost. For example, in the correspondence analysis (CA) of the "benthos" data set, the standard coordinates of the species points indicate vertex, or extreme unit profile, positions and each sample point lies at a weighted average of the species points, using the sample's profile elements across species as weights (see Exhibit

8.3)—this is often called the *barycentric property* of CA. When the standard coordinates are multiplied by the square roots of their masses to reduce them to their contribution coordinates, this property is lost but now the main contributing species are visible. In the log-ratio analysis of the data set "morphology" in Chapter 7, the ability to detect equilibrium relationships when variables fall on straight lines (see Exhibit 7.3) would clearly be lost if each variable were rescaled into its position in terms of contribution coordinates. So, as I said in the Epilogue to *Correspondence Analysis in Practice*, you cannot "have your cake and eat it too"—all the desirable properties one might like to have in a biplot cannot be included in one display, although we can introduce additional graphical "tricks" such as omitting the labels of low contributing points and making the size of symbols related to an omitted aspect of the data, such as the point masses (see Exhibit 8.3, for example).

Computation of a biplot

The solution of a biplot is found by performing a weighted least-squares fit of the product of the left and right matrices to the target matrix, a solution that is conveniently encapsulated in the SVD. One way of computing the SVD is by a process known as *alternating least squares*. Suppose that the target matrix is $\mathbf{S}$ and the approximation $\mathbf{S} \approx \mathbf{X}\mathbf{Y}^\mathsf{T}$ is sought. Writing this approximation as an equality by including a matrix of residuals, or "errors":

$$\mathbf{S} = \mathbf{X}\mathbf{Y}^\mathsf{T} + \mathbf{E}$$

is recognizable as a regression problem if either $\mathbf{X}$ or $\mathbf{Y}^\mathsf{T}$ is fixed. For example, for a fixed $\mathbf{X}$ the least-squares solution for $\mathbf{Y}^\mathsf{T}$ is $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{S}$. If there are weights associated with the rows of the matrix, where the weights are in the diagonal matrix $\mathbf{D}_w$, then the weighted least-squares solution for $\mathbf{Y}^\mathsf{T}$ would be $(\mathbf{X}^\mathsf{T}\mathbf{D}_w\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{D}_w\mathbf{S}$. Having estimated $\mathbf{Y}^\mathsf{T}$, it is regarded as fixed and a similar regression, with or without weights, is performed to estimate a new $\mathbf{X}$. From step to step it can be proved that the residual sum of squared errors reduces, so the process is repeated until the solution converges. At each step the estimates have to be orthonormalized: for example, the left matrix $\mathbf{X}$ would be orthonormalized so that $\mathbf{X}^\mathsf{T}\mathbf{D}_w\mathbf{X} = \mathbf{I}$, which means that the regression step is just a matrix multiplication.[10] The main point is that the left and right matrices are solutions of alternative regressions, with or without weights.

10. This is the more complicated aspect of this algorithm, which we omit here. In practice the dimensions can be computed one at a time: start with any vector $\mathbf{x}$ with the same number of elements as rows of $\mathbf{S}$ and which has been normalized as $\mathbf{x}^\mathsf{T}\mathbf{x} = 1$, then a solution for $\mathbf{y}$ simplifies as $\mathbf{S}^\mathsf{T}\mathbf{x}$; then normalize $\mathbf{y}$ so that $\mathbf{y}^\mathsf{T}\mathbf{y} = 1$ and estimate $\mathbf{x}$ as $\mathbf{S}\mathbf{y}$; normalize $\mathbf{x}$ and continue this process until convergence, which gives the first pair of singular vectors, the singular value α being the norm of the final $\mathbf{x}$ or $\mathbf{y}$ before being normalized to 1. This first dimension is then subtracted out from $\mathbf{S}$,. i.e. $\mathbf{S}$ is replaced by $\mathbf{S} - \alpha\mathbf{x}\mathbf{y}^\mathsf{T}$ (using normalized $\mathbf{x}$ and $\mathbf{y}$) and the process is repeated to find the solution for the second dimension. The only difference when weighted solutions are required is to normalize $\mathbf{x}$ and $\mathbf{y}$ at each step using the weights, and use weighted least-squares regression in each step, which leads to the solution of a weighted SVD.

The optimality of a biplot

Since there are variations of the biplot display, the question arises as to how each variation approximates the original data. The answer is quite simply that the approximation is always the same, with the definition of row and column weights depending on the scaling of the biplot coordinates. The different biplots of a CA illustrate what is meant. In Chapter 8, equation (8.2) defined correspondence analysis in terms of the regular unweighted SVD of the matrix of standardized residuals $\mathbf{D}_r^{-½}(\mathbf{P} - \mathbf{r}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-½}$:

$$\mathbf{D}_r^{-½}(\mathbf{P} - \mathbf{r}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-½} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T}, \text{ where } \mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$$

(notice here that the identification conditions on $\mathbf{U}$ and $\mathbf{V}$ do not contain weights). If we were interested in biplotting the standardized residuals themselves, we would use coordinate matrices such as $\mathbf{U}\mathbf{D}_\alpha$ and $\mathbf{V}$, or $\mathbf{U}\mathbf{D}_\alpha^{½}$ and $\mathbf{V}\mathbf{D}_\alpha^{½}$ for example. But CA is not biplotting the standardized residuals—in the case of the asymmetric CA biplots, for example, one set of points is plotted in principal coordinates and the other set in standard coordinates, for example $\mathbf{F} = \mathbf{D}_r^{-½}\mathbf{U}\mathbf{D}_\alpha$ for rows and $\mathbf{\Gamma} = \mathbf{D}_c^{-½}\mathbf{V}$ for columns, or $\mathbf{\Phi} = \mathbf{D}_r^{-½}\mathbf{U}$ for rows and $\mathbf{G} = \mathbf{D}_c^{-½}\mathbf{V}\mathbf{D}_\alpha$ for columns—see the definitions in (8.3) and (8.4). In order to have the corresponding matrix of scalar products on the right hand side of the CA definition, the target matrix in the defining equation becomes:

$$\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1} - \mathbf{1}\mathbf{1}^\mathsf{T} = \mathbf{\Phi}\mathbf{D}_\alpha\mathbf{\Gamma}^\mathsf{T}, \text{ where } \mathbf{\Phi}^\mathsf{T}\mathbf{D}_r\mathbf{\Phi} = \mathbf{\Gamma}^\mathsf{T}\mathbf{D}_c\mathbf{\Gamma} = \mathbf{I}$$

i.e., in various scalar forms: $p_{ij}/(r_ic_j) - 1 = (p_{ij}/r_i - c_j)/c_j = (p_{ij}/c_j - r_i)/r_i = \sum_k \alpha_k\phi_{ik}\gamma_{jk}$ (the first form shows the ratio of the data element to its expected value, while the second and third forms show how the asymmetric map represents the profiles' deviations from their expected values relative to their expected values). This defines a *generalized* (or *weighted*) SVD where the rows and columns are weighted by the row masses $\mathbf{r}$ and column masses $\mathbf{c}$ respectively. Associating the singular values with the left or right singular vectors in this version of the definition (the standard coordinates) will give the two types of asymmetric biplot and the low-dimensional approximation of the scalar products to the target matrix is by weighted least-squares using the masses.

The form of the definition for the contribution biplot, where the column points, for example, are rescaled by the square roots of their respective masses, plots $\mathbf{F}$ and $\mathbf{V}$ jointly. This implies the generalized SVD in terms of the row profiles $\mathbf{D}_r^{-1}\mathbf{P}$:

$$(\mathbf{D}_r^{-1}\mathbf{P} - \mathbf{1}\mathbf{c}^\mathsf{T})\mathbf{D}_c^{-½} = \mathbf{\Phi}\mathbf{D}_\alpha\mathbf{V}^\mathsf{T}, \text{ where } \mathbf{\Phi}^\mathsf{T}\mathbf{D}_r\mathbf{\Phi} = \mathbf{V}^\mathsf{T}\mathbf{V} = \mathbf{I}$$

that is, with rows weighted by the row masses $\mathbf{r}$, and columns unweighted. The target matrix consists of the standardized profile elements $(p_{ij}/r_i - c_j)/c_j^{½}$, hence the

alternative name of standard biplot for the contribution biplot. The weighting makes sense since the row profiles should be weighted but not the columns, which have already been standardized.

Nonlinear biplots

Another development of the biplot is that of so-called "nonlinear" biplots, where variables are represented by curves and estimation of the values for each case on a particular variable is performed by finding the point on the curve closest to the case. While being of theoretical interest, nonlinear biplots are unlikely to find favour amongst users because of their complexity of interpretation, which detracts from the simplicity of the "linear" biplot treated in this book and its desirable properties such as decomposition of variance, nesting of dimensions and parallel contours for each biplot axis. It is extremely difficult to make any deductions about the properties of variables and their inter-relationships when they are represented by different curves. My view is that it is much better, from a practical point of view, to consider appropriate nonlinear transformations of the original variables, which users and non-specialists can understand, and then use the linear biplot, bearing in mind the nonlinear scales of the resulting biplot axes.

The future of biplots

The SVD as a mathematical result has been known for more than a century, and its property of identifying matrix approximations of any desired rank by (weighted) least squares makes it the most useful matrix result in the area of multivariate data analysis. Algorithms for computing the SVD are well-researched and provide global optima for the dimension-reducing methods that have been presented in this book. The SVD has appeared, is appearing and will appear in every area of scientific research where tabular data are collected. Wherever there is an SVD, there is a biplot. Data are often collected by a painstaking and expensive process and I have always thought it a pity that the richness of a data set is not fully exposed to the researcher who has taken so much trouble to collect it. The biplot is a tool for exploring complex data sets of all types and sizes. The future of the biplot is in its further application in many different areas of research, to make data more transparent to the researcher, and to assist in the interpretation and in the discovery of structures and patterns, both suspected and unsuspected.

List of Exhibits

INDEX

Note: E after the page number indicates a reference to an Exhibit.

About the Author

Michael Greenacre, Professor of Statistics at the Pompeu Fabra University in Barcelona and research collaborator with the BBVA Foundation, was educated initially in his country of birth, South Africa, and then obtained his doctorate in Paris at the Pierre et Marie Curie University (Paris VI) under professor Jean-Paul Benzécri, the creator of correspondence analysis as it is known today. He specialized in the visualization of large multivariate data sets, especially in the social and environmental sciences, and spent sabbatical research periods at Rothamsted Experimental Station (UK); Bell Laboratories, Rochester University and Stanford University (USA); the École des Mines (France); and the Norwegian Polar Environmental Centre in Tromsø (Norway). Besides co-editing three books on data visualization, he has written three books on correspondence analysis, the latest of which *(Correspondence Analysis in Practice*, 2nd edition) has been published in Spanish *(La práctica del análisis de correspondencias*, 2008) by the BBVA Foundation.