Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Appendix C Offprint

Computational Note

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación BBVA

Computational Note

This appendix is a short summary of the software used for the analyses in this book, using packages from the R environment for statistical computing and graphics. Data sets and R code for reproducing the results are given online at the supporting website:

www.multivariatestatistics.org

As an introduction to the online code, we give here a list of some of the common R functions and packages used in the computations of this book.

Contents

Functions from the **base** package in R

The open-access R software has become the standard for statistical computing, especially for conducting research, thanks to its flexible programming environment. It is downloadable for free from the R project website

www.r-project.org

The simple installation process sets up R with what is called the **base** package, consisting of various functions that are commonly used (later we list more specialized packages that need to be downloaded and installed separately). Here we list some of the useful functions in the **base** package:

- `hist` – takes a set of data on a continuous or count variable and makes a histogram; user can choose the interval boundaries; see Exhibits 1.2 and 1.3, for example.

- `qqnorm` – takes a set of data on a continuous variable and plots the sample quantiles against the quantiles of a normal distribution; if the points follow the 45-degree diagonal line of the plot, the data can be regarded as normal, otherwise not; see Chapter 17.

- `shapiro.wilks` – takes a set of data on a continuous variable and performs the Shapiro-Wilks test for normality; if the *p*-value is small then normality is rejected; see Chapter 17.

- `pairs` – takes a rectangular data matrix as input and computes all bivariate scatterplots; see Exhibits 1.4 and 20.8.

- `boxplot` – takes a set of data on a continuous variable and makes a box-and-whisker plot, optionally with a categorical variable that makes boxplots for each category alongside one another, with a common scale; see Exhibits 1.5 and 1.8.

- `scale` – takes a set of data on a continuous variable and standardizes it by subtracting its mean (i.e., centering) and dividing by its standard deviation (i.e., normalization); centering or normalization can be switched off; see Chapter 3.

- `dist` – takes a rectangular data matrix as input and computes a distance matrix between the rows, with several choices of distance functions; for example, see Exhibit 4.5.

- `cor` – takes either two sets of data or a matrix of data with variables in columns and computes the single correlation in the former case, or the correlation matrix in the latter case; optionally computes Spearman rank correlations; see Exhibit 6.4.

- `table` – takes a single categorical and counts the frequencies in each category; if two categorical variables are given the function counts the frequencies in the cross-tabulation; see Exhibit 6.6.

- `sample` – takes a set of data and performs random sampling, without replacement (this re-arranges, or shuffles, the data set randomly) for permutation testing or with replacement for bootstrapping; see Chapter 18.

- `hclust` – takes a matrix of distance or dissimilarities (e.g., created by function `dist`) and performs hierarchical clustering; various clustering algorithms can be selected, including Ward clustering; see Chapters 7 and 8.

- `kmeans` – takes a rectangular matrix of data and the specified number of groups and performs *k*-means nonhierarchical clustering; see Chapter 8.

- `cmdscale` – takes a matrix of distances or dissimilarities (e.g., created by function `dist`) and performs classical multidimensional scaling; see Chapter 9.

- `lm` – takes data on a response variable and one or more explanatory variables (or predictors) and performs least-squares linear regression; weights can be specified for weighted least-squares regression; see Chapters 10–20.

- `glm` – takes data on a response variable and one or more explanatory variables (or predictors) and performs generalized linear modelling (GLM); several link functions and error distributions can be specified, giving linear regression, Poisson regression and logistic regression, for example; see Chapters 10 and 18.

- `prcomp` and `princomp` – alternative functions for computing a principal component analysis on a rectangular data matrix, where rows are assumed to be sampling units and columns to be variables; see Chapter 12.

- `kruskal.test` – takes a data set for a continuous variable and a grouping variable and performs the Kruskal-Wallis rank test of difference between groups (the nonparametric equivalent of a one-way ANOVA); see Chapter 17.

Package **ca**

The **ca** package performs correspondence analysis (function `ca`) and multiple correspondence analysis (function `mjca` – this generalization of CA to multivariate categorical data, more used in the social sciences, is not discussed in this book). Various graphical options are available using function `plot.ca`, including plotting with contribution coordinates and three-dimensional visualization of a CA solution with three principal axes, using function `plot3d.ca`, including interaction with 3d display such as rotation and zooming. The 3d graphics uses the R package **rgl**; see Chapter 13.

Package **vegan**

The **vegan** package performs a variety of multivariate analyses and includes most of the methods treated in this book, and aimed at biologists (specifically botanists, but the terminology can be equated to any biological application). Methods that are not included in R's base package described above are computation of Bray-Curtis and Gower dissimilarities (function `vegdist` with options `method="bray"` or `method="gower"` respectively), various diversity measures

(function `diversity`), nonmetric multidimensional scaling (function `metaMDS`), canonical correspondence analysis (function `cca`), redundancy analysis (function `rda` – like `cca` but for continuous response variables) and various permutation tests (e.g., function `permutest`); see Chapters 15 and 20.

Packages **maptools**, **mapdata** and **mapproj**

Packages **maptools** and **mapdata** provide functions that allow drawing of geographical maps, with **mapdata** containing the outlines of all the world's landmasses and several countries. The **mapproj** package performs a variety of map projections, using function `mapproject`, based on the latitude and longitude coordinates of a set of spatial locations. This is useful for obtaining coordinates on which Euclidean distances can be computed that approximate great circle distances; see Chapters 11 and 19.

Package **mgcv**

This package performs generalized additive modelling (GAM), using a function `gam` that functions very similarly to `glm` for generalized linear modelling. Explanatory variables can be defined as smooth functions using the function `s`, for example `s(x)` for predictor `x`; see Chapters 18, 19 and 20.

Packages **rpart** and **tree**

These packages are alternatives for classification and regression trees, also called *recursive partitioning* (hence **rpart**). They define tree models in the same style as functions `lm`, `glm` and `gam`, as a response variable ~ sum of explanatory variables. Plotting the result using `plot` gives the tree plot; see Chapter 18.

Additional functions in supporting material

Several additional functions that are used in our applications are given in the supporting material on www.multivariatestatistics.org.

- `fuzzy.tri` – takes a set of data on a continuous variable, with a specified number of categories, and transforms to fuzzy categories using triangular membership functions; hinges are by default defined as quantiles, but can be supplied by the user; see Exhibit 3.3.

- `chidist` – takes a rectangular matrix of same-scale nonnegative data as input and computes the matrix of chi-square distances between rows or between columns; see Exhibit 4.7.

- `jaccard` – takes a rectangular matrix of presence-absence data (ones and zeros) and computes the matrix of Jaccard dissimilarities between rows or between columns (this can also be achieved in the **vegan** package using function `vegdist` with `method="jaccard"`); see Chapter 5 and Exhibit 7.1.