Use R!

Series Editors:
Robert Gentleman Kurt Hornik Giovanni Parmigiani

For other titles published in this series, go to
http://www.springer.com/series/6991

Brian Everitt • Torsten Hothorn

An Introduction to Applied Multivariate Analysis with R

Springer

Brian Everitt
Professor Emeritus
King's College
London, SE5 8AF
UK
brian.everitt@btopenworld.com

Torsten Hothorn
Institut für Statistik
Ludwig-Maximilians-Universität München
Ludwigstr. 33
80539 München
Germany
Torsten.Hothorn@stat.uni-muenchen.de

Series Editors:
Robert Gentleman
Program in Computational Biology
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Avenue, N. M2-B876
Seattle, Washington 98109
USA

Kurt Hornik
Department of Statistik and Mathematik
Wirtschaftsuniversität Wien
Augasse 2-6
A-1090 Wien
Austria

Giovanni Parmigiani
The Sidney Kimmel Comprehensive
Cancer Center at Johns Hopkins University
550 North Broadway
Baltimore, MD 21205-2011
USA

ISBN 978-1-4419-9649-7 e-ISBN 978-1-4419-9650-3
DOI 10.1007/978-1-4419-9650-3
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011926793

To our wives, Mary-Elizabeth and Carolin.

Preface

The majority of data sets collected by researchers in all disciplines are multivariate, meaning that several measurements, observations, or recordings are taken on each of the units in the data set. These units might be human subjects, archaeological artifacts, countries, or a vast variety of other things. In a few cases, it may be sensible to isolate each variable and study it separately, but in most instances all the variables need to be examined simultaneously in order to fully grasp the structure and key features of the data. For this purpose, one or another method of multivariate analysis might be helpful, and it is with such methods that this book is largely concerned. Multivariate analysis includes methods both for describing and exploring such data and for making formal inferences about them. The aim of all the techniques is, in a general sense, to display or extract the signal in the data in the presence of noise and to find out what the data show us in the midst of their apparent chaos.

The computations involved in applying most multivariate techniques are considerable, and their routine use requires a suitable software package. In addition, most analyses of multivariate data should involve the construction of appropriate graphs and diagrams, and this will also need to be carried out using the same package. R is a statistical computing environment that is powerful, flexible, and, in addition, has excellent graphical facilities. It is for these reasons that it is the use of R for multivariate analysis that is illustrated in this book.

In this book, we concentrate on what might be termed the "core" or "classical" multivariate methodology, although mention will be made of recent developments where these are considered relevant and useful. But there is an area of multivariate statistics that we have omitted from this book, and that is multivariate analysis of variance (MANOVA) and related techniques such as Fisher's linear discriminant function (LDF). There are a variety of reasons for this omission. First, we are not convinced that MANOVA is now of much more than historical interest; researchers may occasionally pay lip service to using the technique, but in most cases it really is no more than this. They quickly

move on to looking at the results for individual variables. And MANOVA for repeated measures has been largely superseded by the models that we shall describe in Chapter 8. Second, a classification technique such as LDF needs to be considered in the context of modern classification algorithms, and these cannot be covered in an introductory book such as this.

Some brief details of the theory behind each technique described are given, but the main concern of each chapter is the correct application of the methods so as to extract as much information as possible from the data at hand, particularly as some type of graphical representation, via the R software.

The book is aimed at students in applied statistics courses, both undergraduate and post-graduate, who have attended a good introductory course in statistics that covered hypothesis testing, confidence intervals, simple regression and correlation, analysis of variance, and basic maximum likelihood estimation. We also assume that readers will know some simple matrix algebra, including the manipulation of matrices and vectors and the concepts of the inverse and rank of a matrix. In addition, we assume that readers will have some familiarity with R at the level of, say, Dalgaard (2002). In addition to such a student readership, we hope that many applied statisticians dealing with multivariate data will find something of interest in the eight chapters of our book.

Throughout the book, we give many examples of R code used to apply the multivariate techniques to multivariate data. Samples of code that could be entered interactively at the R command line are formatted as follows:

```
R> library("MVA")
```

Here, `R>` denotes the prompt sign from the R command line, and the user enters everything else. The symbol `+` indicates additional lines, which are appropriately indented. Finally, output produced by function calls is shown below the associated code:

```
R> rnorm(10)

 [1]  1.8808  0.2572 -0.3412  0.4081  0.4344  0.7003  1.8944
 [8] -0.2993 -0.7355  0.8960
```

In this book, we use several R packages to access different example data sets (many of them contained in the package **HSAUR2**), standard functions for the general parametric analyses, and the **MVA** package to perform analyses. All of the packages used in this book are available at the Comprehensive R Archive Network (CRAN), which can be accessed from http://CRAN.R-project.org.

The source code for the analyses presented in this book is available from the **MVA** package. A demo containing the R code to reproduce the individual results is available for each chapter by invoking

```
R> library("MVA")
R> demo("Ch-MVA") ### Introduction to Multivariate Analysis
R> demo("Ch-Viz") ### Visualization
```

```
R> demo("Ch-PCA") ### Principal Components Analysis
R> demo("Ch-EFA") ### Exploratory Factor Analysis
R> demo("Ch-MDS") ### Multidimensional Scaling
R> demo("Ch-CA")  ### Cluster Analysis
R> demo("Ch-SEM") ### Structural Equation Models
R> demo("Ch-LME") ### Linear Mixed-Effects Models
```

Thanks are due to Lisa Möst, BSc., for help with data processing and LaTeX typesetting, the copy editor for many helpful corrections, and to John Kimmel, for all his support and patience during the writing of the book.

January 2011

Brian S. Everitt, London
Torsten Hothorn, München

Contents