

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

© Springer Science+Business Media, LLC 2011

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

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 L^AT_EX 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

Preface	vii
1 Multivariate Data and Multivariate Analysis	1
1.1 Introduction	1
1.2 A brief history of the development of multivariate analysis	3
1.3 Types of variables and the possible problem of missing values .	4
1.3.1 Missing values	5
1.4 Some multivariate data sets	7
1.5 Covariances, correlations, and distances	12
1.5.1 Covariances	12
1.5.2 Correlations	14
1.5.3 Distances	14
1.6 The multivariate normal density function	15
1.7 Summary	23
1.8 Exercises	23
2 Looking at Multivariate Data: Visualisation	25
2.1 Introduction	25
2.2 The scatterplot	26
2.2.1 The bivariate boxplot	28
2.2.2 The convex hull of bivariate data	32
2.2.3 The chi-plot	34
2.3 The bubble and other glyph plots	34
2.4 The scatterplot matrix	39
2.5 Enhancing the scatterplot with estimated bivariate densities ..	42
2.5.1 Kernel density estimators	42
2.6 Three-dimensional plots	47
2.7 Trellis graphics	50
2.8 Stalactite plots	53
2.9 Summary	56
2.10 Exercises	60

3	Principal Components Analysis	61
3.1	Introduction	61
3.2	Principal components analysis (PCA)	61
3.3	Finding the sample principal components	63
3.4	Should principal components be extracted from the covariance or the correlation matrix?	65
3.5	Principal components of bivariate data with correlation coefficient r	68
3.6	Rescaling the principal components	70
3.7	How the principal components predict the observed covariance matrix	70
3.8	Choosing the number of components	71
3.9	Calculating principal components scores	72
3.10	Some examples of the application of principal components analysis	74
3.10.1	Head lengths of first and second sons	74
3.10.2	Olympic heptathlon results	78
3.10.3	Air pollution in US cities	86
3.11	The biplot	92
3.12	Sample size for principal components analysis	93
3.13	Canonical correlation analysis	94
3.13.1	Head measurements	96
3.13.2	Health and personality	99
3.14	Summary	101
3.15	Exercises	102
4	Multidimensional Scaling	105
4.1	Introduction	105
4.2	Models for proximity data	105
4.3	Spatial models for proximities: Multidimensional scaling	106
4.4	Classical multidimensional scaling	106
4.4.1	Classical multidimensional scaling: Technical details	107
4.4.2	Examples of classical multidimensional scaling	110
4.5	Non-metric multidimensional scaling	121
4.5.1	House of Representatives voting	123
4.5.2	Judgements of World War II leaders	124
4.6	Correspondence analysis	127
4.6.1	Teenage relationships	130
4.7	Summary	131
4.8	Exercises	132
5	Exploratory Factor Analysis	135
5.1	Introduction	135
5.2	A simple example of a factor analysis model	136
5.3	The k -factor analysis model	137

5.4	Scale invariance of the k -factor model	138
5.5	Estimating the parameters in the k -factor analysis model	139
5.5.1	Principal factor analysis	141
5.5.2	Maximum likelihood factor analysis	142
5.6	Estimating the number of factors	142
5.7	Factor rotation	143
5.8	Estimating factor scores	147
5.9	Two examples of exploratory factor analysis	148
5.9.1	Expectations of life	148
5.9.2	Drug use by American college students	151
5.10	Factor analysis and principal components analysis compared	157
5.11	Summary	159
5.12	Exercises	159
6	Cluster Analysis	163
6.1	Introduction	163
6.2	Cluster analysis	165
6.3	Agglomerative hierarchical clustering	166
6.3.1	Clustering jet fighters	171
6.4	K -means clustering	175
6.4.1	Clustering the states of the USA on the basis of their crime rate profiles	176
6.4.2	Clustering Romano-British pottery	180
6.5	Model-based clustering	183
6.5.1	Finite mixture densities	186
6.5.2	Maximum likelihood estimation in a finite mixture density with multivariate normal components	187
6.6	Displaying clustering solutions graphically	191
6.7	Summary	197
6.8	Exercises	200
7	Confirmatory Factor Analysis and Structural Equation Models	201
7.1	Introduction	201
7.2	Estimation, identification, and assessing fit for confirmatory factor and structural equation models	202
7.2.1	Estimation	202
7.2.2	Identification	203
7.2.3	Assessing the fit of a model	204
7.3	Confirmatory factor analysis models	206
7.3.1	Ability and aspiration	206
7.3.2	A confirmatory factor analysis model for drug use	211
7.4	Structural equation models	216
7.4.1	Stability of alienation	216
7.5	Summary	222

7.6	Exercises	223
8	The Analysis of Repeated Measures Data	225
8.1	Introduction	225
8.2	Linear mixed-effects models for repeated measures data	232
8.2.1	Random intercept and random intercept and slope models for the timber slippage data	233
8.2.2	Applying the random intercept and the random intercept and slope models to the timber slippage data ..	235
8.2.3	Fitting random-effect models to the glucose challenge data	240
8.3	Prediction of random effects	247
8.4	Dropouts in longitudinal data	248
8.5	Summary	257
8.6	Exercises	257
	References	259
	Index	271