

# Applied Multivariate Data Analysis

# **Applied Multivariate Data Analysis**

**Second Edition**

**Brian S. Everitt**

*Institute of Psychiatry, King's College London, UK*

and

**Graham Dunn**

*School of Epidemiology and Health Sciences,  
University of Manchester, UK*



John Wiley & Sons, Ltd

First published in Great Britain in 2001 by Arnold  
This impression printed by Hodder Education,  
a part of Hachette Livre UK,  
338 Euston Road, London NW1 3BH

© 2001 Brian S. Everitt and Graham Dunn

John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West  
Sussex, PO19 8SQ, United Kingdom

For details of our global editorial offices, for customer services and  
for information about how to apply for permission to reuse the copyright  
material in this book please see our website at [www.wiley.com](http://www.wiley.com).

The right of the author to be identified as the author of this work has  
been asserted in accordance with the Copyright, Design and Patents Act  
1988.

All rights reserved. No part of this publication may be reproduced, stored in  
a retrieval system, or transmitted, in any form or by any means, electronic,  
mechanical, photocopying, recording or otherwise, except as permitted by  
the UK Copyright, Designs and Patents Act 1988, without the prior permission  
of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content  
that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often  
claimed as trademarks. All brand names and product names used in this  
book are trade names, service marks, trademarks or registered trademarks  
of their respective owners. The publisher is not associated with any product  
or vendor mentioned in this book. This publication is designed to provide  
accurate and authoritative information in regard to the subject matter covered.  
It is sold on the understanding that the publisher is not engaged in rendering  
professional services. If professional advice or other expert assistance is required,  
the services of a competent professional should be sought.

*British Library Cataloguing in Publication Data*

A catalogue record for this book is available from the British Library

*Library of Congress Cataloging-in-Publication Data*

A catalog record for this book is available from the Library of Congress

ISBN 978-0-4707-1117-0

8 9 10

Typeset in 10/12pt Times by Academic & Technical Typesetting, Bristol

---

# Contents

<b>1</b>	<b>Multivariate data and multivariate statistics</b>	<b>1</b>
1.1	Introduction	1
1.2	Types of data	2
1.3	Basic multivariate statistics	4
1.4	The aims of multivariate analysis	6
<b>2</b>	<b>Exploring multivariate data graphically</b>	<b>9</b>
2.1	Introduction	9
2.2	The scatterplot	9
2.3	The scatterplot matrix	15
2.4	Enhancing the scatterplot	17
2.5	Coplots and trellis graphics	26
2.6	Checking distributional assumptions using probability plots	41
2.7	Summary	45
	Exercises	45
<b>3</b>	<b>Principal components analysis</b>	<b>48</b>
3.1	Introduction	48
3.2	Algebraic basics of principal components	49
3.3	Rescaling principal components	52
3.4	Calculating principal component scores	53
3.5	Choosing the number of components	53
3.6	Two simple examples of principal components analysis	54
3.7	More complex examples of the application of principal components analysis	56
3.8	Using principal components analysis to select a subset of variables	63
3.9	Using the last few principal components	65
3.10	The biplot	65
3.11	Geometrical interpretation of principal components analysis	69
3.12	Projection pursuit	69

3.13	Summary	71
	Exercises	71
<b>4</b>	<b>Correspondence analysis</b>	<b>74</b>
4.1	Introduction	74
4.2	A simple example of correspondence analysis	75
4.3	Correspondence analysis for two-dimensional contingency tables	78
4.4	Three applications of correspondence analysis	80
4.5	Multiple correspondence analysis	84
4.6	Summary	91
	Exercises	91
<b>5</b>	<b>Multidimensional scaling</b>	<b>93</b>
5.1	Introduction	93
5.2	Proximity matrices and examples of multidimensional scaling	94
5.4	Metric least-squares multidimensional scaling	104
5.5	Non-metric multidimensional scaling	107
5.6	Non-Euclidean metrics	113
5.7	Three-way multidimensional scaling	114
5.8	Inference in multidimensional scaling	119
5.9	Summary	122
	Exercises	122
<b>6</b>	<b>Cluster analysis</b>	<b>125</b>
6.1	Introduction	125
6.2	Agglomerative hierarchical clustering techniques	128
6.3	Optimization methods	142
6.4	Finite mixture models for cluster analysis	148
6.5	Summary	158
	Exercises	158
<b>7</b>	<b>The generalized linear model</b>	<b>161</b>
7.1	Linear models	161
7.2	Non-linear models	165
7.3	Link functions and error distributions in the generalized linear model	168
7.4	Summary	171
	Exercises	172
<b>8</b>	<b>Regression and the analysis of variance</b>	<b>173</b>
8.1	Introduction	173
8.2	Least-squares estimation for regression and analysis of variance models	173
8.3	Direct and indirect effects	190
8.4	Summary	195
	Exercises	195
<b>9</b>	<b>Log-linear and logistic models for categorical multivariate data</b>	<b>198</b>
9.1	Introduction	198

9.2	Maximum likelihood estimation for log-linear and linear-logistic models	199
9.3	Transition models for repeated binary response measures	212
9.4	Summary	216
	Exercises	216
<b>10</b>	<b>Models for multivariate response variables</b>	<b>218</b>
10.1	Introduction	218
10.2	Repeated quantitative measures	218
10.3	Multivariate tests	222
10.4	Random effects models for longitudinal data	224
10.5	Logistic models for multivariate binary responses	237
10.6	Marginal models for repeated binary response measures	240
10.7	Marginal modelling using generalized estimating equations	242
10.8	Random effects models for multivariate repeated binary response measures	244
10.9	Summary	246
	Exercises	246
<b>11</b>	<b>Discrimination, classification and pattern recognition</b>	<b>248</b>
11.1	Introduction	248
11.2	A simple example	249
11.3	Some examples of allocation rules	250
11.4	Fisher's linear discriminant function	253
11.5	Assessing the performance of a discriminant function	254
11.6	Quadratic discriminant functions	255
11.7	More than two groups	257
11.8	Logistic discrimination	260
11.9	Selecting variables	262
11.10	Other methods for deriving classification rules	263
11.11	Pattern recognition and neural networks	264
11.12	Summary	268
	Exercises	268
<b>12</b>	<b>Exploratory factor analysis</b>	<b>271</b>
12.1	Introduction	271
12.2	The basic factor analysis model	272
12.3	Estimating the parameters in the factor analysis model	274
12.4	Rotation of factors	278
12.5	Some examples of the application of factor analysis	280
12.6	Estimating factor scores	283
12.7	Factor analysis with categorical variables	284
12.8	Factor analysis and principal components analysis compared	287
12.9	Summary	287
	Exercises	288
<b>13</b>	<b>Confirmatory factor analysis and covariance structure models</b>	<b>291</b>
13.1	Introduction	291
13.2	Path analysis and path diagrams	292

viii *Contents*

13.3	Estimation of the parameters in structural equation models	295
13.4	A simple covariance structure model and identification	295
13.5	Assessing the fit of a model	297
13.6	Some examples of fitting confirmatory factor analysis models	298
13.7	Structural equation models	302
13.8	Causal models and latent variables: myths and realities	304
13.9	Summary	306
	Exercises	306
<b>Appendices</b>		
A	Software packages	308
A.1	General-purpose packages	308
A.2	More specialized packages	309
B	Missing values	311
C	Answers to selected exercises	314
<b>References</b>		
<b>324</b>		
<b>Index</b>		
<b>337</b>		

---

# Preface

The majority of data sets collected by researchers in all disciplines are multivariate. 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 other method of multivariate analysis may be helpful, and it is with such methods that this book is largely concerned.

Multivariate analysis includes methods both for describing and exploring data and for more formal inferential procedures. 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.

We have made many changes from the first edition of our book, including a separate chapter on correspondence analysis, a section on neural networks for classification, and discussion of extensions of the generalized linear model to situations involving multiple response variables – for example, repeated measures studies. In addition, the graphical techniques chapter has been completely rewritten and many new graphical methods described. Finally, more – and, we hope – better, examples illustrating techniques are to be found in all chapters. As with the first edition, we have aimed the book both at students on statistics courses and at applied researchers dealing with multivariate data. Readers need to have some background in statistics, perhaps of the kind delivered by an introductory course covering estimation, inference, regression, analysis of variance and so on. The main mathematical requirement is a degree of familiarity with matrix algebra, although much of the more technical material is confined to tables so that the less mathematical reader will often be able to follow the discussion to some extent.

In the first part of the book (Chapters 2 to 6) we concentrate largely on what might loosely be described as the exploratory multivariate techniques; often these are primarily graphical in nature, and the graphical display of multivariate