

# 多元数据分析

(英文版)

## ANALYZING MULTIVARIATE DATA



JAMES LATTIN - J. DOUGLAS CARROLL - PAUL E. GREEN



James M. Lattin  
斯坦福大学  
J. Douglas Carroll  
(美) 罗格斯大学 著  
Paul E. Green  
宾夕法尼亚大学



机械工业出版社  
China Machine Press

经典原版书库

O212.4

3

# 多元数据分析

(英文版)

Analyzing Multivariate Data

James M. Lattin

斯坦福大学

J. Douglas Carroll

罗格斯大学

Paul E. Green

宾夕法尼亚大学

(美)

著



机械工业出版社



Z106927

北京信息工程学院图书馆

James M. Lattin, J. Douglas Carroll, and Paul E. Green: Analyzing Multivariate Data (ISBN: 0-534-34974-9).

Original edition copyright © 2003 Brooks/Cole.

First published by Brooks/Cole, an imprint of Thomson Learning, United States of America.

All rights reserved.

Reprinted for the People's Republic of China by Thomson Asia Pte Ltd and China Machine Press under the authorization of Thomson Learning. No part of this book may be reproduced in any form without the express written permission of Thomson Learning Asia and China Machine Press.

本书英文影印版由汤姆森学习出版社与机械工业出版社合作出版。未经出版者书面许可，不得以任何方式复制或抄袭本书内容。

版权所有，侵权必究。

本书版权登记号：图字：01-2002-1061

### 图书在版编目（CIP）数据

多元数据分析（英文版）/（美）拉廷（Lattin, J. M.）等著. -北京：机械工业出版社，2003.7

（经典原版书库）

书名原文：Analyzing Multivariate Data

ISBN 7-111-12412-X

I. 多… II. 拉… III. 多元分析 - 高等学校 - 教材 - 英文 IV. O212.4

中国版本图书馆CIP数据核字（2003）第050107号

机械工业出版社（北京市西城区百万庄大街22号 邮政编码 100037）

责任编辑：杨海玲

北京昌平奔腾印刷厂印刷·新华书店北京发行所发行

2003年7月第1版第1次印刷

787mm×1092mm 1/16·36.25印张

印数：0 001-2 000册

定价：69.00元（附光盘）

凡购本书，如有倒页、脱页、缺页，由本社发行部调换

--	--	--	--

## *Dedication*

*To the people who have provided me with so much love and support throughout my life: my parents, my sister, and my wife and children (who make my every day worthwhile).*

Jim Lattin

*Lovingly dedicated to my mother, Nolie Carroll, my wife, Sylvia, and my two adult sons, Greg and Steve.*

Doug Carroll

*Many thanks to my continually expanding family, to my departed wife, Betty, and to Donna who has brightened my life.*

Paul Green

## About the Authors



**James M. Lattin** is the Robert A. Magowan Professor of Marketing at the Graduate School of Business, Stanford University. He holds an A.B. from Dartmouth College and a Ph.D. in Management Science from the Sloan School of Management at Massachusetts Institute of Technology. Dr. Lattin has been at Stanford since 1984, where he has taught courses in marketing management and data analysis in the M.B.A. and Ph.D. programs; he has also taught extensively in the executive education programs offered by the Graduate School of Business. Dr. Lattin's research interests include choice behavior, database marketing, and models of customer acquisition and retention. His research has won awards from the American Marketing Association in 1984, the *Journal of Retailing* in 1992, and the William O'Dell Award from the *Journal of Marketing Research* in 1996; he was also a finalist for the 1998 John D. C. Little Award for best paper in marketing science and a winner of the 1999 Frank M. Bass Award from the INFORMS College on Marketing. Dr. Lattin has done consulting and executive teaching for a number of companies, including Adobe Systems, Microsoft, and Eli Lilly. He is a member of the marketing advisory committee for the Stanford Federal Credit Union and an advisory board member of several start-up companies, including Bonustree.com, Design Within Reach, and MS2, Inc.



**J. Douglas Carroll** is Board of Governors Professor of Management and Psychology at the Graduate School of Management of Rutgers University. He received his Ph.D. from Princeton University in Psychology (Psychometrics and Mathematical

Psychology), with a minor in Mathematics in 1963. Formerly a Distinguished Member of Technical Staff at AT&T Bell Laboratories, Dr. Carroll's research is on multi-dimensional scaling and related techniques of data analysis, especially as applied to marketing and psychology. Dr. Carroll is an Associate Editor of *Psychometrika*, serves on the editorial boards of the *Journal of Marketing Research* and the *Journal of Classification*, has been a Consulting Editor for the *Journal of Experimental Psychology: General*, and is an editorial consultant for numerous other scientific and professional journals. He is a Fellow of the American Psychological Association (APA), the American Psychological Society (APS), the American Statistical Association (ASA), and the American Association for the Advancement of Science (AAAS). He is past President of the Classification Society of North America (CSNA), the Psychometric Society, the Society of Multivariate Experimental Psychology (SMEP), and Division 5, the Division of Evaluation, Measurement and Statistics, of the APA, as well as past Chair of the Statistics in Marketing Section of the ASA. He was a member of the Founding Council of the International Federation of Classification Societies (IFCS) and is past President of IFCS. Dr. Carroll was corecipient of the Alpha Kappa Psi Award for the most distinguished paper in the *Journal of Marketing* for 1981 and of the 1989 JAMS Best Article Award for the best paper in the *Journal of the Academy of Marketing Science*. He has received the Distinguished Scientific Contribution Award from the APA, and the William James Fellow Award from the APS.



**Paul E. Green** is Professor of Marketing at the Wharton School, University of Pennsylvania. His research emphasizes quantitative methods and new measurement techniques in market analysis and consumer research. He is author or coauthor of 15 books and 5 monographs. His past industry experience has involved such firms as E. I. DuPont de Nemours and Sun Oil Company; recent consulting clients include Air Products, IBM, Bristol-Myers Squibb, Subaru, AT&T, Pfizer, and Marriott. Dr. Green has been honored for his research by the American Marketing Association, the American Statistical Association, the Association for Consumer Research, the Marketing Science Institute, and the Market Research Society (British). He has received the Parlin Award for the Advancement of Science in Marketing, the Paul D. Converse Award in Marketing Theory, and for best articles, the AMA's Alpha Kappa Psi Award (three-time winner) and the William O'Dell Award. He is a Fellow of the American Statistical Association, the Decision Sciences Institute, the Association for Consumer Research, the Academy of Marketing Science, and the Attitude Research Hall of Fame. He received the AMA/Irwin Marketing Educator of the Year Award in 1991 and the Academy of Marketing Science Distinguished Educator of the Year Award in 1992. In 1996, the American Marketing Association established the Paul Green Award for the *Journal of Marketing Research* article that "exhibits the most potential to contribute significantly to the practice of marketing research and research in marketing." In that same year, Dr. Green became the first recipient of the Gilbert A. Churchill, Jr. Award for "lifetime achievement in marketing research."

APR 1997



on to the analysis of dependence. We find that an early discussion of data reduction techniques and measurement models is helpful before discussing canonical correlation and structural equation models. Also, when discussing techniques for the analysis of dependence, this book now considers both single-dependent-variable and multiple-dependent-variable versions of the technique together in the same chapter.

- *New topics*, including logit choice models and structural equation models with latent variables.
- *Expanded coverage* of techniques for the analysis of interdependence, especially scaling methods and cluster analysis.
- *Discussion of cross-validation*. Overfitting is a serious problem that accompanies any exploratory analysis of multivariate data (particularly with the techniques used to perform analysis of dependence). In each chapter, we present approaches that can be used to assess the statistical significance and the generalizability of the results of a given analysis.
- *Software independent*. Students from different disciplines studying different substantive problems have a tendency to adopt different statistical packages. For that reason, this textbook is designed to be “software independent”; that is, not written from the perspective of any one particular application. Instead, we have developed student workbooks specific to particular software packages (e.g., SAS and SPSS) to accompany the textbook.
- *Broader variety of sample problems and exercises*. Instead of a single data set (the Alpha TV Commercial Study from the original), we have chosen to include a wide variety of data sets to show students how multivariate methods can be used to provide insights into different types of problems. More than 100 data sets are included on the CD-ROM that accompanies the text. An Instructor’s Manual, with solutions to the exercises at the end of each chapter, is also available.
- *Selected readings*. In addition to the more comprehensive bibliography at the end of the book, we also provide a set of selected readings at the end of each chapter. These readings are not intended to be exhaustive but to give the student some idea of the origins of each method and some general resources pertaining to issues of importance related to each method.

Some sacrifices have been made to keep the scope of the book manageable. Topics no longer covered include automatic interaction detection (AID) and monotonic analysis of variance (MONANOVA). Conjoint analysis, which was not covered in the original book, is also not covered here.

As far as prerequisites go, the book assumes some familiarity with basic statistics. Most students coming to a course that uses this text will have seen regression analysis in some shape or form (although some will have a less than satisfactory grasp of the intuition underlying regression, unfortunately). The book does make use of matrix algebra, but students should not have to derive the equations to be able to understand the concepts and methods presented herein. To the extent possible,

particular statistical packages (e.g., SAS and SPSS). These workbooks explain how the concepts in the text are linked to the application software and show the student how to perform the analyses presented in each chapter. The program templates provided in the workbooks enable students to run their own analyses of the more than 100 data sets (most taken from real applications in the published literature) contained in the CD-ROM that accompanies the text.

- *Be able to interpret the results of the analysis.* In each chapter, we raise the important issues and problems that tend to come up with the application of each method. We place special emphasis on assessing the generalizability of the results of an analysis, and suggest ways in which students can test the validity of their findings.

## Similarities with Predecessor

The book shares a number of similarities with its predecessor:

- *Practical orientation.* This book is still for the data analyzer. It continues to have a pragmatic orientation designed to appeal to applications-oriented researchers. Each chapter offers at least one real-world application as well as a discussion of the issues related to the proper interpretation of the results.
- *Intuitive approach.* The goal still is to have students understand how these methods work (rather than to present them as a “black box”). We seek to build students’ intuition with a combination of geometrical reasoning (lots of pictures) and limited mathematics (i.e., some matrix algebra to support the intuition). The writing style is still informal and the tendency is still toward concrete numerical demonstration rather than mathematical proof and/or abstract argument.
- *Interdisciplinary.* The book is not written with a single audience in mind. The illustrations and sample problems are drawn from a wide range of areas, including marketing research, sociology, psychology, and economics.
- *Presentation format.* Each of the chapters still follows a fairly standard format. We begin by discussing the objectives of each technique and some areas of potential application. We then explain how each method works with words and pictures (followed by a more mathematical exposition). An example (or two) helps to make clear the application of the technique and the interpretation of the results. We also provide a discussion of the problems and questions that can arise when doing this type of multivariate analysis.

## Differences from Predecessor

This book also differs from its predecessor in several respects:

- *Organization of topics.* This book now begins with analysis of interdependence (i.e., factor analysis, multidimensional scaling, cluster analysis) before moving



we have tried to modularize the mathematics (i.e., confine them to relatively self-contained sections) so as not to deter the interested but less mathematically minded student. The instructor has the option of covering the material in Chapters 2 and 3 in class or of assigning the material as background reading.

## Acknowledgments

Each of the three authors is indebted to a number of individuals who have provided support, guidance, and inspiration. Jim Lattin acknowledges his coauthors, Paul Green and Doug Carroll, for the role they played in influencing his understanding of multivariate methods and in shaping his approach to teaching these techniques. He would also like to express his gratitude to Joel Levine, Dartmouth College, for “rescuing” him in his freshman year and inspiring a deep and abiding interest in data analysis. Many thanks to Curt Hinrichs for his extraordinary patience and gentle handling throughout the duration of this project; Curt always seemed to have the right words of encouragement to keep things moving along. Jim also acknowledges the contributions over the years from the students in his classes in multivariate data analysis. Too numerous to mention individually, these students worked problems, unearthed interesting findings, and provided feedback on course notes and teaching materials. Jim appreciates the financial support provided by the Stanford Graduate School of Business, especially from the Center for Electronic Business & Commerce and the Robert A. Magowan Chair. Finally, Jim would like to express his love and affection to his incredibly supportive family: wife Marilyn, son Alex, and daughter Lindsay.

Doug Carroll is deeply indebted to his two coauthors—Paul Green especially for having taken by far the principal role in writing the previous version of *Analyzing Multivariate Data*, whose success was the primary inspiration for this revised version, as well as for his invaluable support and encouragement, and many contributions to this much-revised version. Special thanks to Jim Lattin, who unambiguously took primary responsibility for producing this much-needed revision. Doug also owes an irrepayable debt to his mentors at Princeton during his graduate work there—Professors Harold Gulliksen (now deceased) and Ledyard Tucker (now an Emeritus Professor at the University of Illinois at Champaign-Urbana)—for teaching him the basics of multidimensional scaling, two-way and three-way factor analysis, and so many other areas of applied multivariate analysis, as well as to Professor Bob Abelson at Yale University and Roger Shepard, first at Bell Labs, then at Harvard and Stanford. In addition to Paul Green, Doug has profited enormously from interaction with a large number of highly talented collaborators over the years, including Jih-Jie Chang, Sande Pruzansky, Joseph Kruskal, Phipps Arabie, Mike Wish, Geert De Soete, Wayne DeSarbo, Anil Chaturvedi, Yoshio Takane, Willem Heiser, Jacqueline Meulman, Christoph Klauer, Sharon Weinberg, Suzanne Winsberg, Crawford Clark, and others too numerous to mention here. Doug also is eternally indebted to his marvelous secretarial assistant, Kathleen Power, without whose irreplaceable help in innumerable different ways his contribution to this book would not have been possible.

Doug also wants to express a particularly strong sense of gratitude to Ulas Akkucuk, his research assistant and Ph.D. student at Rutgers, without whose expert help with the data analysis the chapter on multidimensional scaling probably could not have been written in its current form. Finally, none of his work, on this book or in other domains, would be possible without the love and constant support of his wife, Sylvia.

Paul Green acknowledges his debt to Jim Lattin for funding (in small part) the original *Analyzing Multivariate Data* book. More importantly, Jim has expanded and modernized the original to the point where it is a brand-new book, both literally and figuratively. Paul's indebtedness to Doug Carroll goes back to the 1960s, when Paul and his students were introduced to the exciting research going on at Bell Laboratories. Roger Shepard, Joe Kruskal, and Doug Carroll were at the forefront of new developments in multidimensional scaling, clustering, and monotonic ANOVA. Doug Carroll, in particular, cheerfully and skillfully guided Paul and his colleagues through the then arcane (at least to business school professors) field of psychometrics and data analysis methods, both metric and nonmetric. Paul's indebtedness to Doug's patient and wise tutelage continues to this day. Finally, Paul thanks his family and his departed wife, Betty, for their patience and support through his many years of research and writing.

The three of us are also indebted to the individuals involved in reviewing, testing, editing, and producing the book. Profound thanks to the reviewers of the manuscript: Tammo Bigmolt (Tilburg University), Bruce Cooil (Vanderbilt University), Robert Gould (University of California, Los Angeles), Kamel Jedidi (Columbia Business School), Peter Lenk (University of Michigan Business School), Thomas E. Love (Case Western Reserve University), Zhen Luo (Pennsylvania State University), Doug MacLachlan (University of Washington), George A. Marcoulides (California State University, Fullerton), Kent Nakamoto (Virginia Tech), Julia Norton (California State University, Hayward), Jason Osborne (University of Oklahoma), Bill Rayens (University of Kentucky), Hirokuni Tamura (University of Washington), Niels Waller (University of California, Davis), and Steve Zanakakis (Florida International University). Special thanks to Julia Norton and Steve Zanakakis, who also classroom-tested the manuscript, and to Doug MacLachlan, who did an accuracy check of the final document. Collectively these individuals provided useful feedback that helped to greatly improve the book.

The authors would also like to acknowledge the contributions of the following individuals for their role in moving the book along through the production process: Curt Hinrichs, publisher; Matt Ballantyne, project manager; Katherine Brayton, editorial assistant; Hal Lockwood of Penmarin Books, production editor; Steven Summerlight, copy editor; and Tom Webster of Lineworks, illustrator. Only those who have gone through this process can truly understand how much work went into turning an ugly caterpillar of a manuscript into this beautiful butterfly of a book.

James M. Lattin  
J. Douglas Carroll  
Paul E. Green

But perhaps the best feature of the book (in the opinion of the young professor) was the way the authors used the geometry underlying the mathematics to show how the techniques really worked. Even a student with only a tentative grasp of matrix algebra can see what is happening when he or she understands that each matrix operation corresponds to a stretching (or shrinking) and rotation of the data. After the original text went out of print, the young professor continued to teach the course from the notes he had developed. Many things about the course changed (e.g., topics were added, dropped, and rearranged; new examples and larger data sets were included to keep pace with the increased computational capabilities of today's software packages), but the underlying pedagogy remained the same.

This new book is the result of the collaboration between the now not-so-young professor and the two authors of the original text. It is not so much a revision as it is a rebirth: a fresh look at multivariate techniques more than 20 years later, with new examples, new data, and some new methods, but grounded in the same pedagogical approach (applications-oriented, intuitively motivated using the underlying geometry of the method) that guided the creation of the original.

## Organization

The book is organized into three parts. By way of introduction, Part I (Chapters 1 through 3) provides a general overview of multivariate methods, some helpful background on vectors and matrices and their geometrical interpretation, and a review of multiple regression analysis. Part II (Chapters 4 through 8) focuses on the analysis of interdependence, both among variables (principal components, factor analysis) and among objects (multidimensional scaling, cluster analysis). Part III (Chapters 9 through 13) covers canonical correlation and methods used in the analysis of dependence, including structural equation models with latent variables, logit choice models, and special cases of the general linear model (analysis of variance, discriminant analysis).

## Objectives

Our objective is to make students intelligent users of these multivariate techniques and good critics of multivariate analyses performed by others. If students are to be intelligent users and good critics of the techniques discussed in this book, they must have some grasp of theory, application, and interpretation. In other words, they must

- *Have some intuition as to how the technique works.* To this end, we use a geometric interpretation to provide the students with a mental picture of how each method works. We use mathematics to support the underlying intuition (rather than as a substitute for it).
- *Be able to apply the technique.* We take a hands-on approach, providing illustrative examples in each chapter based on real-world data. To facilitate the application of these methods, we have developed student workbooks specific to

--	--	--	--

## *Preface*

Once upon a time, over two decades ago now, two gentlemen (Paul Green and Doug Carroll) collaborated on a textbook titled *Analyzing Multivariate Data*. Their objective was to produce a book with a pragmatic orientation—"a book for the data analyzer." Quoting from the preface of that book,

Most users of multivariate statistical techniques are not professional statisticians. They are applications-oriented researchers—psychologists, sociologists, marketing researchers, management scientists, and so on—who, from time to time, need the techniques to help them in their work. This text has been written for them and for students of these disciplines. . . . As implied by the title, emphasis on data analysis and the objectives of people who do data analysis has shaped the character of the whole enterprise.

Many people adopted the book, including a young professor (Jim Lattin) who was teaching a course on multivariate data analysis for the very first time. The level of the text seemed quite appropriate for the mix of graduate students taking the course (mainly first- and some second-year graduate students from different parts of the university). It was not too difficult (i.e., it did not rely too heavily on mathematics beyond the preparation of the typical student) and not too simplistic (i.e., it was not a "cookbook"). Because the book presented a variety of applications, it appealed to a relatively broad cross-section of students (not only students in marketing, organizational behavior, and accounting from the Graduate School of Business, but also students in engineering, education, economics, food research, psychology, sociology, and statistics).

# Brief Contents

## **PART I**

### *Overview 1*

- 1 Introduction 3
- 2 Vectors and Matrices 19
- 3 Regression Analysis 38

## **PART II**

### *Analysis of Interdependence 81*

- 4 Principal Components  
Analysis 83
- 5 Exploratory Factor  
Analysis 127
- 6 Confirmatory Factor  
Analysis 171

- 7 Multidimensional  
Scaling 206

- 8 Cluster Analysis 264

## **PART III**

### *Analysis of Dependence 311*

- 9 Canonical Correlation 313

- 10 Structural Equation  
Models with Latent  
Variables 352

- 11 Analysis of Variance 386

- 12 Discriminant Analysis 426

- 13 Logit Choice Models 474

--	--	--	--

# *Contents*

Preface XVII

**PART I**

---

## *Overview 1*

### **1 Introduction 3**

- 1.1 The Nature of Multivariate Data 4**
  - 1.1.1 Some Definitions 4*
  - 1.1.2 Observations and Data 4*
  - 1.1.3 Levels of Measurement 5*
- 1.2 Overview of Multivariate Methods 9**
  - 1.2.1 Principal Components 9*
  - 1.2.2 Factor Analysis 11*
  - 1.2.3 Multidimensional Scaling 11*
  - 1.2.4 Clustering 12*

- 1.2.5 *Canonical Correlation* 13
- 1.2.6 *Structural Equation Models with Latent Variables* 14
- 1.2.7 *Analysis of Variance* 15
- 1.2.8 *Discriminant Analysis* 16
- 1.2.9 *Logit Choice Models* 16

**1.3** Format of Succeeding Chapters 17

- Selected Readings 18
- General* 18
- Data and Measurement* 18

## **2 Vectors and Matrices 19**

**2.1** Introduction 19

**2.2** Definitions 20

- 2.2.1 *Vectors* 20
- 2.2.2 *Matrices* 20
- 2.2.3 *Coordinate Systems* 21

**2.3** Geometric Interpretation of Operations 25

- 2.3.1 *Scalar Multiplication: Scaling* 25
- 2.3.2 *Vector Multiplication: Projection* 26
- 2.3.3 *Matrix Multiplication* 28

**2.4** Matrix Properties 32

- 2.4.1 *Singular Value Decomposition* 32
- 2.4.2 *Cross-Product Matrix* 32

**2.5** Learning Summary 35

- Selected Readings 36
- General* 36
- Exercises 36

## **3 Regression Analysis 38**

**3.1** Introduction 38

**3.2** Regression Analysis: How It Works 39

- 3.2.1 *Intuition* 39
- 3.2.2 *Mechanics* 46



- 3.3** Sample Problem: Leslie Salt Property 47
- 3.4** Questions Regarding the Application of Regression Analysis 50
  - 3.4.1 *How Good Is the Fit?* 53
  - 3.4.2 *Is It Significant?* 54
  - 3.4.3 *Detecting Problems with the Model* 56
  - 3.4.4 *Comparing Models* 65
  - 3.4.5 *Forecasting* 68
  - 3.4.6 *Model Validation* 71
- 3.5** Learning Summary 75
  - Selected Readings 77
    - General* 77
    - Outliers and Regression Diagnostics* 77
    - Resampling and Model Validation* 78
  - Exercises 78

## PART II

---

# *Analysis of Interdependence* 81

## **4** Principal Components Analysis 83

- 4.1** Introduction 83
  - 4.1.1 *Potential Applications* 84
- 4.2** Principal Components: How It Works 91
  - 4.2.1 *Intuition* 91
  - 4.2.2 *Mechanics* 97
- 4.3** Sample Problem: Gross State Product 101
  - 4.3.1 *Data* 101
  - 4.3.2 *Results* 105
- 4.4** Questions Regarding the Application of Principal Components 109
  - 4.4.1 *When Is It Appropriate to Use Principal Components?* 109
  - 4.4.2 *How Should the Data Be Scaled?* 111
  - 4.4.3 *How Many Components Should Be Retained?* 112
  - 4.4.4 *How to Assess the Validity of the Solution?* 117

- 4.5 Learning Summary** 126
  - Selected Readings 123
    - General* 123
    - How Many Components?* 123
    - Jackknife and Bootstrapping* 123
  - Exercises 123

## **5 Exploratory Factor Analysis 127**

- 5.1 Introduction** 127
  - 5.1.1 Potential Applications* 128
- 5.2 Exploratory Factor Analysis: How It Works** 131
  - 5.2.1 Intuition* 131
  - 5.2.2 Mechanics* 141
- 5.3 Sample Problem: Perceptions of Ready-to-Eat Cereals** 147
  - 5.3.1 Data* 147
  - 5.3.2 Results* 148
- 5.4 Questions Regarding the Application of Factor Analysis** 153
  - 5.4.1 Can I Obtain a Solution with Correlated Factors?* 153
  - 5.4.2 How Can I Use the Results in Subsequent Analyses?* 156
  - 5.4.3 How Can I Assess the Validity of the Factor Structure?* 159
- 5.5 Learning Summary** 163
  - Selected Readings 165
    - General* 165
    - Rotation* 166
  - Exercises 166

## **6 Confirmatory Factor Analysis 171**

- 6.1 Introduction** 171
  - 6.1.1 Potential Applications* 172
- 6.2 Confirmatory Factor Analysis: How It Works** 175
  - 6.2.1 Intuition* 175
  - 6.2.2 Mechanics* 184