



海外优秀数学类教材系列丛书

THOMSON

影印版

*Applied Multivariate Methods
for Data Analysts*

应用多元统计 分析方法

□ DALLAS E. JOHNSON



高等教育出版社
Higher Education Press



应用多元统计分析

应用多元统计分析

A population of k variables characterizes k individuals in
the k space. A study is

应用多元统计 分析方法

应用多元统计分析



应用多元统计分析

0212.4
Y9D

海外优秀数学类教材系列丛书

影印版

*Applied Multivariate Methods
for Data Analysts*

应用多元统计分析方法

Dallas E. Johnson
Kansas State University



高等教育出版社
Higher Education Press

SCH/2/111

图字: 01-2004-3218 号

Dallas E. Johnson

Applied Multivariate Methods for Data Analysts, first Edition

ISBN: 0-534-23796-7

Copyright © 1998 by Duxbury, a division of Thomson Learning

Original language published by Thomson Learning (a division of Thomson Learning Asia Pte Ltd). All Rights reserved. 本书原版由汤姆森学习出版集团出版。版权所有, 盗印必究。

Higher Education Press is authorized by Thomson Learning to publish and distribute exclusively this English language reprint edition. This edition is authorized for sale in the People's Republic of China only (excluding Hong Kong, Macao SAR and Taiwan).

Unauthorized export of this edition is a violation of the Copyright Act. No part of this publication may be reproduced or distributed by any means, or stored in a database or retrieval system, without the prior written permission of the publisher.

本书英文影印版由汤姆森学习出版集团授权高等教育出版社独家出版发行。此版本仅限在中华人民共和国境内(但不允许在中国香港、澳门特别行政区及中国台湾地区)销售。未经授权的本书出口将被视为违反版权法的行为。未经出版者预先书面许可, 不得以任何方式复制或发行本书的任何部分。

981-265-506-9

图书在版编目(CIP)数据

应用多元统计分析方法 = Applied Multivariate Methods for Data Analysts / (美) 约翰逊 (Johnson, D.

E.) 著. —北京: 高等教育出版社, 2005. 6

(海外优秀数学类教材系列丛书)

ISBN 7-04-016545-7

I. 应... II. 约... III. 多元分析—统计分析—分析方法 IV. 0212.4

中国版本图书馆 CIP 数据核字(2005)第 044048 号

策划编辑 徐可 责任编辑 徐可 封面设计 王凌波 责任印制 陈伟光

| | | | |
|------|----------------|------|---|
| 出版发行 | 高等教育出版社 | 购书热线 | 010-58581118 |
| 社 址 | 北京市西城区德外大街 4 号 | 免费咨询 | 800-810-0598 |
| 邮政编码 | 100011 | 网 址 | http://www.hep.edu.cn |
| 总 机 | 010-58581000 | | http://www.hep.com.cn |
| 经 销 | 北京蓝色畅想图书发行有限公司 | 网上订购 | http://www.landaco.com |
| 印 刷 | 北京民族印刷厂 | | http://www.landaco.com.cn |
| 开 本 | 787×1092 1/16 | 版 次 | 2005 年 6 月第 1 版 |
| 印 张 | 36.5 | 印 次 | 2005 年 6 月第 1 次印刷 |
| 字 数 | 650 000 | 定 价 | 43.30 元(含光盘) |

本书如有缺页、倒页、脱页等质量问题, 请到所购图书销售部门联系调换。

版权所有 侵权必究

物料号 16545-00

出版者的话

在我国已经加入WTO、经济全球化的今天，为适应当前我国高校各类创新人才培养的需要，大力推进教育部倡导的双语教学，配合教育部实施的“高等学校教学质量与教学改革工程”和“精品课程”建设的需要，高等教育出版社有计划、大规模地开展了海外优秀数学类系列教材的引进工作。

高等教育出版社和 Pearson Education, John Wiley & Sons, McGraw-Hill, Thomson Learning等国外出版公司进行了广泛接触，经国外出版公司的推荐并在国内专家的协助下，提交引进版权总数100余种。收到样书后，我们聘请了国内高校一线教师、专家、学者参与这些原版教材的评介工作，并参考国内相关专业的课程设置和教学实际情况，从中遴选出了这套优秀教材组织出版。

这批教材普遍具有以下特点：(1)基本上是近3年出版的，在国际上被广泛使用，在同类教材中具有相当的权威性；(2)高版次，历经多年教学实践检验，内容翔实准确、反映时代要求；(3)各种教学资源配套整齐，为师生提供了极大的便利；(4)插图精美、丰富，图文并茂，与正文相辅相成；(5)语言简练、流畅、可读性强，比较适合非英语国家的学生阅读。

本系列丛中，有 Finney、Weir 等编的《托马斯微积分》(第10版, Pearson)，其特色可用“呈传统特色、富革新精神”概括，本书自20世纪50年代第1版以来，平均每四五年就有一个新版面世，长达50余年始终盛行于西方教坛，作者既有相当高的学术水平，又热爱教学，长期工作在教学第一线，其中，年近90的G.B.Thomas教授长年在MIT工作，具有丰富的教学经验；Finney教授也在MIT工作达10年；Weir是美国数学建模竞赛委员会主任。Stewart编的《微积分》(第5版, Thomson Learning)配备了丰富的教学资源，是国际上最畅销的微积分原版教材，2003年全球销量约40余万册，在美国，占据了约50%~60%的微积分教材市场，其用户包括耶鲁等名牌院校及众多一般院校。本系列丛书还包括Anton编的经典教材《线性代数及其应用》(第8版, Wiley)；Jay L.Devore编的优秀教材《概率论与数理统计》(第5版, Thomson Learning)等。在努力降低引进教材售价方面，高等教育出版社做了大量和细致的工作，这套引进的教材体现了一定的权威性、系统性、先进性和经济性等特点。

通过影印、翻译、编译这批优秀教材，我们一方面要不断地分析、学习、消化吸收国外优秀教材的长处，吸取国外出版公司的制作经验，提升我们自编教材的教学资源配套标准，使我国高校教材建设水平上一个新的台阶；与此同时，我们还将尝试组织海外作者和国内作者合编外文版基础课数学教材，并约请国内专家改编部分国外优秀教材，以适应我国实际教

学环境。

这套教材出版后，我们将结合各高校的双语教学计划，开展大规模的宣传、培训工作，及时地将本套丛书推荐给高校使用。在使用过程中，我们衷心希望广大高校教师和同学提出宝贵的意见和建议。

高等教育出版社高等理科分社联系电话：010-58581384，E-mail: xuke@hep.com.cn。

高等教育出版社
2004年4月20日

Preface

I once attended a conference at which George Box stated that “Statistics is much too important to be left entirely to statisticians.” A bit later, Walt Federer stated that “Science is much too important to be left entirely to scientists.” Both of these famous statisticians were correct! Never before in the history of science and statistics has there been a greater need for interactions and collaborations between scientists and statisticians. This book helps to facilitate such collaborations and interactions. I have been fortunate in that I have had substantial contact with scientists during my tenure at Kansas State University. These collaborations have greatly influenced my approach to teaching multivariate methods. I believe that multivariate methods are too important to be taught only to statisticians.

Furthermore, I have been teaching public seminars and college courses in applied multivariate analyses for the last 20 years. In these seminars and courses, students have posed many important questions that multivariate methods can help answer. As data sets grow in size, multivariate methods become ever more useful. Today’s technologies make it very easy to collect large amounts of data; multivariate methods are needed to determine whether such massive amounts of data actually contain information. It has been said that while it is easy to collect data, it is much harder to collect information. Multivariate methods can help determine whether there is information in data, and they can also help to summarize that information when it exists.

To date, textbooks have emphasized only the theory of multivariate methods or only the application of the methods. Readers were given information that was either too advanced to apply or too elementary to illustrate the power of the methods. This text has broken the mold by focusing on the why, when, how, and what of multivariate analyses and answering the following questions:

Why should multivariate methods be used?

When should they be used?

How can they be used?

And what has been learned by the application of the methods?

Ideally, users of this book will have had a previous course in statistics that included multiple regression. Some familiarity with matrix algebra is desirable, but not crucial. My approach assumes familiarity with most of the statistical

concepts encountered in a first course in statistics, such as means and standard deviations, correlations, p-values, hypothesis tests, and confidence limits.

While this text is loaded with examples using real data, several of the exercises are directed at data sets that students are asked to provide from their own experiences. I find that students enjoy working with their own data. So, when I teach multivariate methods, I require each student to provide a data set for class use along with a description of the data's important features and the reasons behind its being collected. These data sets are then placed in a computer directory that every student in the class can access. I then use these data sets as much as possible when assigning exercises to the class. I strongly encourage instructors who use this book to do the same.

Other unique features of this text include:

- annotated computer output, emphasizing SAS and SPSS
- extensive use of graphics to explain concepts
- data disk that contains data files from text discussion and exercises, as well as computer commands used to create the analyses described throughout the text

I owe much of the development of this text to those who have participated in my seminars and courses. From these "students," I learned about their needs, their concerns, and their abilities. In writing this text, I have tried to address their needs and concerns, while recognizing their differing abilities.

Acknowledgments

I wish to express my appreciation to all who helped me with the development of this text. I am particularly grateful to the students at Kansas State University and students who have taken public seminars through the Institute of Professional Education. These students have provided numerous valuable suggestions that have greatly improved the content of this text. I would also like to thank Ms. Carolyn Crockett and Mr. Alexander Kugushev for their valuable suggestions. I would like to thank the following reviewers for their helpful comments: Marcia Gumpertz, North Carolina State University; John E. Hewett, University of Missouri, Columbia; Linda S. Hynan, Baylor University; Dipak Jain, Northwestern University; Lincoln Moses, Stanford University; Mack C. Shelley II, Iowa State University; Eric Smith, Virginia Polytech Institute; and Richard Sundheim, St. Cloud State University. I also thank Mrs. Jane Cox for her help in creating many of the formulas in this text. Finally, I would like to thank my parents, Chet and Dorothy Johnson, for giving me the opportunity for furthering my education and my wife, Erma, for the help and support that she provided during this endeavor.

Dallas Johnson

Contents

| | |
|--|-----------|
| 1. APPLIED MULTIVARIATE METHODS | 1 |
| 1.1 An Overview of Multivariate Methods | 1 |
| Variable- and Individual-Directed Techniques | 2 |
| Creating New Variables | 2 |
| Principal Components Analysis | 3 |
| Factor Analysis | 3 |
| Discriminant Analysis | 4 |
| Canonical Discriminant Analysis | 5 |
| Logistic Regression | 5 |
| Cluster Analysis | 5 |
| Multivariate Analysis of Variance | 6 |
| Canonical Variates Analysis | 7 |
| Canonical Correlation Analysis | 7 |
| Where to Find the Preceding Topics | 7 |
| 1.2 Two Examples | 8 |
| Independence of Experimental Units | 11 |
| 1.3 Types of Variables | 11 |
| 1.4 Data Matrices and Vectors | 12 |
| Variable Notation | 13 |
| Data Matrix | 13 |
| Data Vectors | 13 |
| Data Subscripts | 14 |
| 1.5 The Multivariate Normal Distribution | 15 |
| Some Definitions | 15 |
| Summarizing Multivariate Distributions | 16 |
| Mean Vectors and Variance–Covariance Matrices | 16 |
| Correlations and Correlation Matrices | 17 |
| The Multivariate Normal Probability Density Function | 19 |
| Bivariate Normal Distributions | 19 |

| | | | |
|-----------|---|-----------|-----------|
| 1.6 | Statistical Computing | 22 | |
| | Cautions About Computer Usage | 22 | |
| | Missing Values | 22 | |
| | Replacing Missing Values by Zeros | 23 | |
| | Replacing Missing Values by Averages | 23 | |
| | Removing Rows of the Data Matrix | 23 | |
| | Sampling Strategies | 24 | |
| | Data Entry Errors and Data Verification | 24 | |
| 1.7 | Multivariate Outliers | 25 | |
| | Locating Outliers | 25 | |
| | Dealing with Outliers | 25 | |
| | Outliers May Be Influential | 26 | |
| 1.8 | Multivariate Summary Statistics | 26 | |
| 1.9 | Standardized Data and/or Z Scores | 27 | |
| | Exercises | 28 | |
| 2. | SAMPLE CORRELATIONS | | 35 |
| 2.1 | Statistical Tests and Confidence Intervals | 35 | |
| | Are the Correlations Large Enough to Be Useful? | 36 | |
| | Confidence Intervals by the Chart Method | 36 | |
| | Confidence Intervals by Fisher's Approximation | 38 | |
| | Confidence Intervals by Ruben's Approximation | 39 | |
| | Variable Groupings Based on Correlations | 40 | |
| | Relationship to Factor Analysis | 46 | |
| 2.2 | Summary | 46 | |
| | Exercises | 47 | |
| 3. | MULTIVARIATE DATA PLOTS | | 55 |
| 3.1 | Three-Dimensional Data Plots | 55 | |
| 3.2 | Plots of Higher Dimensional Data | 59 | |
| | Chernoff Faces | 61 | |
| | Star Plots and Sun-Ray Plots | 63 | |

| | | | |
|------------|--|-----------|-----------|
| | Andrews' Plots | 65 | |
| | Side-by-Side Scatter Plots | 66 | |
| 3.3 | Plotting to Check for Multivariate Normality | 67 | |
| | Summary | 73 | |
| | Exercises | 73 | |
| 4. | EIGENVALUES AND EIGENVECTORS | | 77 |
| 4.1 | Trace and Determinant | 77 | |
| | Examples | 78 | |
| 4.2 | Eigenvalues | 78 | |
| 4.3 | Eigenvectors | 79 | |
| | Positive Definite and Positive Semidefinite Matrices | 80 | |
| 4.4 | Geometric Descriptions ($p = 2$) | 82 | |
| | Vectors | 82 | |
| | Bivariate Normal Distributions | 83 | |
| 4.5 | Geometric Descriptions ($p = 3$) | 87 | |
| | Vectors | 87 | |
| | Trivariate Normal Distributions | 87 | |
| 4.6 | Geometric Descriptions ($p > 3$) | 90 | |
| | Summary | 91 | |
| | Exercises | 91 | |
| 5. | PRINCIPAL COMPONENTS ANALYSIS | | 93 |
| 5.1 | Reasons for Using Principal Components Analysis | 93 | |
| | Data Screening | 93 | |
| | Clustering | 95 | |
| | Discriminant Analysis | 95 | |
| | Regression | 95 | |
| 5.2 | Objectives of Principal Components Analysis | 96 | |
| 5.3 | Principal Components Analysis on the Variance–Covariance Matrix Σ | 96 | |
| | Principal Component Scores | 98 | |
| | Component Loading Vectors | 98 | |

| | | |
|-----------|--|------------|
| 5.4 | Estimation of Principal Components | 99 |
| | Estimation of Principal Component Scores | 99 |
| 5.5 | Determining the Number of Principal Components | 99 |
| | Method 1 | 100 |
| | Method 2 | 100 |
| 5.6 | Caveats | 107 |
| 5.7 | PCA on the Correlation Matrix P | 109 |
| | Principal Component Scores | 110 |
| | Component Correlation Vectors | 110 |
| | Sample Correlation Matrix | 110 |
| | Determining the Number of Principal Components | 110 |
| 5.8 | Testing for Independence of the Original Variables | 111 |
| 5.9 | Structural Relationships | 111 |
| 5.10 | Statistical Computing Packages | 112 |
| | SAS ^R PRINCOMP Procedure | 112 |
| | Principal Components Analysis Using Factor Analysis Programs | 118 |
| | PCA with SPSS's FACTOR Procedure | 124 |
| | Summary | 142 |
| | Exercises | 142 |
| 6. | FACTOR ANALYSIS | 147 |
| 6.1 | Objectives of Factor Analysis | 147 |
| 6.2 | Caveats | 148 |
| 6.3 | Some History of Factor Analysis | 148 |
| 6.4 | The Factor Analysis Model | 150 |
| | Assumptions | 150 |
| | Matrix Form of the Factor Analysis Model | 151 |
| | Definitions of Factor Analysis Terminology | 151 |
| 6.5 | Factor Analysis Equations | 151 |
| | Nonuniqueness of the Factors | 152 |
| 6.6 | Solving the Factor Analysis Equations | 153 |

| | | |
|-------------|---|------------|
| 6.7 | Choosing the Appropriate Number of Factors | 155 |
| | Subjective Criteria | 156 |
| | Objective Criteria | 156 |
| 6.8 | Computer Solutions of the Factor Analysis Equations | 157 |
| | Principal Factor Method on R | 158 |
| | Principal Factor Method with Iteration | 159 |
| 6.9 | Rotating Factors | 170 |
| | Examples ($m = 2$) | 171 |
| | Rotation Methods | 172 |
| | The Varimax Rotation Method | 173 |
| 6.10 | Oblique Rotation Methods | 174 |
| 6.11 | Factor Scores | 180 |
| | Bartlett's Method or the Weighted Least-Squares Method | 181 |
| | Thompson's Method or the Regression Method | 181 |
| | Ad Hoc Methods | 181 |
| | Summary | 212 |
| | Exercises | 213 |
| 7. | DISCRIMINANT ANALYSIS | 217 |
| 7.1 | Discrimination for Two Multivariate Normal Populations | 217 |
| | A Likelihood Rule | 218 |
| | The Linear Discriminant Function Rule | 218 |
| | A Mahalanobis Distance Rule | 218 |
| | A Posterior Probability Rule | 218 |
| | Sample Discriminant Rules | 219 |
| | Estimating Probabilities of Misclassification | 220 |
| | Resubstitution Estimates | 220 |
| | Estimates from Holdout Data | 220 |
| | Cross-Validation Estimates | 221 |
| 7.2 | Cost Functions and Prior Probabilities (Two Populations) | 229 |
| 7.3 | A General Discriminant Rule (Two Populations) | 231 |
| | A Cost Function | 232 |
| | Prior Probabilities | 232 |

| | | | |
|------------|---|-----|------------|
| | Average Cost of Misclassification | 232 | |
| | A Bayes Rule | 233 | |
| | Classification Functions | 233 | |
| | Unequal Covariance Matrices | 233 | |
| | Tricking Computing Packages | 234 | |
| 7.4 | Discriminant Rules (More than Two Populations) | | 235 |
| | Basic Discrimination | 238 | |
| 7.5 | Variable Selection Procedures | | 245 |
| | Forward Selection Procedure | 245 | |
| | Backward Elimination Procedure | 246 | |
| | Stepwise Selection Procedure | 246 | |
| | Recommendations | 247 | |
| | Caveats | 247 | |
| 7.6 | Canonical Discriminant Functions | | 255 |
| | The First Canonical Function | 256 | |
| | A Second Canonical Function | 257 | |
| | Determining the Dimensionality of the Canonical Space | | 260 |
| | Discriminant Analysis with Categorical Predictor Variables | | 273 |
| 7.7 | Nearest Neighbor Discriminant Analysis | | 275 |
| 7.8 | Classification Trees | | 283 |
| | Summary | 283 | |
| | Exercises | 283 | |
| 8. | LOGISTIC REGRESSION METHODS | | 287 |
| 8.1 | Logistic Regression Model | | 287 |
| 8.2 | The Logit Transformation | | 287 |
| | Model Fitting | 288 | |
| 8.3 | Variable Selection Methods | | 296 |
| 8.4 | Logistic Discriminant Analysis (More Than Two Populations) | | 301 |
| | Logistic Regression Models | 301 | |
| | Model Fitting | 302 | |
| | Another SAS LOGISTIC Analysis | | 314 |
| | Exercises | 316 | |

| | | |
|-------------|--|------------|
| 9. | CLUSTER ANALYSIS | 319 |
| 9.1 | Measures of Similarity and Dissimilarity | 319 |
| | Ruler Distance | 319 |
| | Standardized Ruler Distance | 320 |
| | A Mahalanobis Distance | 320 |
| | Dissimilarity Measures | 320 |
| 9.2 | Graphical Aids in Clustering | 321 |
| | Scatter Plots | 321 |
| | Using Principal Components | 322 |
| | Andrews' Plots | 322 |
| | Other Methods | 322 |
| 9.3 | Clustering Methods | 322 |
| | Nonhierarchical Clustering Methods | 323 |
| | Hierarchical Clustering | 323 |
| | Nearest Neighbor Method | 323 |
| | A Hierarchical Tree Diagram | 325 |
| | Other Hierarchical Clustering Methods | 326 |
| | Comparisons of Clustering Methods | 327 |
| | Verification of Clustering Methods | 327 |
| | How Many Clusters? | 327 |
| | Beale's <i>F</i> -Type Statistic | 328 |
| | A Pseudo Hotelling's T^2 Test | 329 |
| | The Cubic Clustering Criterion | 329 |
| | Clustering Order | 334 |
| | Estimating the Number of Clusters | 339 |
| | Principal Components Plots | 348 |
| | Clustering with SPSS | 355 |
| | SAS's FASTCLUS Procedure | 369 |
| 9.4 | Multidimensional Scaling | 385 |
| | Exercises | 395 |
| 10. | MEAN VECTORS AND VARIANCE-COVARIANCE MATRICES | 397 |
| 10.1 | Inference Procedures for Variance-Covariance Matrices | 397 |
| | A Test for a Specific Variance-Covariance Matrix | 398 |
| | A Test for Sphericity | 400 |

| | | |
|-------------|---|------------|
| | A Test for Compound Symmetry | 403 |
| | A Test for the Huynh–Feldt Conditions | 405 |
| | A Test for Independence | 406 |
| | A Test for Independence of Subsets of Variables | 407 |
| | A Test for the Equality of Several Variance–Covariance Matrices | 408 |
| 10.2 | Inference Procedures for a Mean Vector | 408 |
| | Hotelling's T^2 Statistic | 409 |
| | Hypothesis Test for μ | 409 |
| | Confidence Region for μ | 409 |
| | A More General Result | 411 |
| | Special Case—A Test of Symmetry | 412 |
| | A Test for Linear Trend | 418 |
| | Fitting a Line to Repeated Measures | 418 |
| | Multivariate Quality Control | 419 |
| 10.3 | Two Sample Procedures | 420 |
| | Repeated Measures Experiments | 420 |
| 10.4 | Profile Analyses | 431 |
| 10.5 | Additional Two-Group Analyses | 432 |
| | Paired Samples | 432 |
| | Unequal Variance–Covariance Matrices | 433 |
| | Large Sample Sizes | 433 |
| | Small Sample Sizes | 433 |
| | Summary | 434 |
| | Exercises | 434 |

11. MULTIVARIATE ANALYSIS OF VARIANCE

| | | |
|-------------|---|------------|
| 11.1 | MANOVA | 439 |
| | MANOVA Assumptions | 440 |
| | Test Statistics | 440 |
| | Test Comparisons | 441 |
| | Why Do We Use MANOVAs? | 441 |
| | A Conservative Approach to Multiple Comparisons | 442 |

| | | |
|-----------|--|------------|
| 11.2 | Dimensionality of the Alternative Hypothesis | 455 |
| 11.3 | Canonical Variates Analysis | 456 |
| | The First Canonical Variate | 456 |
| | The Second Canonical Variate | 457 |
| | Other Canonical Variates | 457 |
| 11.4 | Confidence Regions for Canonical Variates | 458 |
| | Summary | 485 |
| | Exercises | 485 |
| 12 | PREDICTION MODELS AND MULTIVARIATE REGRESSION | 489 |
| 12.1 | Multiple Regression | 489 |
| 12.2 | Canonical Correlation Analysis | 494 |
| | Two Sets of Variables | 494 |
| | The First Canonical Correlation | 495 |
| | The Second Canonical Correlation | 495 |
| | Number of Canonical Correlations | 496 |
| | Estimates | 496 |
| | Hypothesis Tests on the Canonical Correlations | 497 |
| | Interpreting Canonical Functions | 508 |
| | Canonical Correlation Analysis with SPSS | 511 |
| 12.3 | Factor Analysis and Regression | 515 |
| | Summary | 522 |
| | Exercises | 522 |
| | APPENDIX A: MATRIX RESULTS | 525 |
| A.1 | Basic Definitions and Rules of Matrix Algebra | 525 |
| A.2 | Quadratic Forms | 527 |
| A.3 | Eigenvalues and Eigenvectors | 528 |
| A.4 | Distances and Angles | 529 |
| A.5 | Miscellaneous Results | 529 |