DISK INCLUDED 

Applied Multivariate Statistical Analysis

Fourth Edition

Richard A. Johnson
Dean W. Wichern

FOURTH EDITION

Applied Multivariate Statistical Analysis

RICHARD A. JOHNSON

University of Wisconsin—Madison

DEAN W. WICHERN

Texas A&M University

PRENTICE HALL, Upper Saddle River, New Jersey 07458

Library of Congress Cataloging-in-Publication Data

Johnson, Richard Arnold.
Applied multivariate statistical analysis / Richard A. Johnson, Dean W. Wichern. -- 4th ed.
p. cm.
Includes bibliographical references and indexes.
ISBN 0-13-834194-X
1. Multivariate analysis. I. Wichern, Dean W. II. Title.
QA278.J63 1998
519.5'35--dc21 97-42907
CIP

Acquisitions Editor: **ANN HEATH**
Marketing Manager: **MELODY MARCUS**
Editorial Assistant: **MINDY McCLARD**
Editorial Director: **TIM BOZIK**
Editor-in-Chief: **JEROME GRANT**
Assistant Vice-President of Production and Manufacturing: **DAVID W. RICCARDI**
Editorial/Production Supervision: **RICHARD DeLORENZO**
Managing Editor: **LINDA MIHATOV BEHRENS**
Executive Managing Editor: **KATHLEEN SCHIAPARELLI**
Manufacturing Buyer: **ALAN FISCHER**
Manufacturing Manager: **TRUDY PISCIOTTI**
Marketing Assistant: **PATRICK MURPHY**
Director of Creative Services: **PAULA MAYLAHN**
Art Director: **JAYNE CONTE**
Cover Designer: **BRUCE KENSELAAR**

Simon & Schuster / A Viacom Company
Upper Saddle River, NJ 07458

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

ISBN 0-13-834194-X

Prentice-Hall International (UK) Limited, *London*
Prentice-Hall of Australia Pty. Limited, *Sydney*
Prentice-Hall Canada Inc., *Toronto*
Prentice-Hall Hispanoamericana, S.A., *Mexico*
Prentice-Hall of India Private Limited, *New Delhi*
Prentice-Hall of Japan, Inc., *Tokyo*
Simon & Schuster Asia Pte. Ltd., *Singapore*
Editora Prentice-Hall do Brasil, Ltda., *Rio de Janeiro*

Preface

INTENDED AUDIENCE

This book originally grew out of our lecture notes for an "Applied Multivariate Analysis" course offered jointly by the Statistics Department and the School of Business at the University of Wisconsin—Madison. Applied Multivariate Statistical Analysis, Fourth Edition, is concerned with statistical methods for describing and analyzing multivariate data. Data analysis, while interesting with one variable, becomes truly fascinating and challenging when several variables are involved. Researchers in the biological, physical, and social sciences frequently collect measurements on several variables. Modern computer packages readily provide the numerical results to rather complex statistical analyses. We have tried to provide readers with the supporting knowledge necessary for making proper interpretations, selecting appropriate techniques, and understanding their strengths and weaknesses. We hope our discussions will meet the needs of experimental scientists, in a wide variety of subject matter areas, as a readable introduction to the statistical analysis of multivariate observations.

LEVEL

Our aim is to present the concepts and methods of multivariate analysis at a level that is readily understandable by readers who have taken two or more statistics

courses. We emphasize the applications of multivariate methods and consequently, have attempted to make the mathematics as palatable as possible. We avoid the use of calculus. On the other hand, the concepts of a matrix and of matrix manipulations are important. We do not assume the reader is familiar with matrix algebra. Rather, we introduce matrices as they appear naturally in our discussions, and we then show how they simplify the presentation of multivariate models and techniques.

The introductory account of matrix algebra, in Chapter 2, highlights the more important matrix algebra results as they apply to multivariate analysis. The Chapter 2 supplement provides a summary of matrix algebra results for those with little or no previous exposure to the subject. This supplementary material helps make the book self-contained and is used to complete proofs. The proofs may be ignored on the first reading. In this way we hope to make the book accessible to a wide audience.

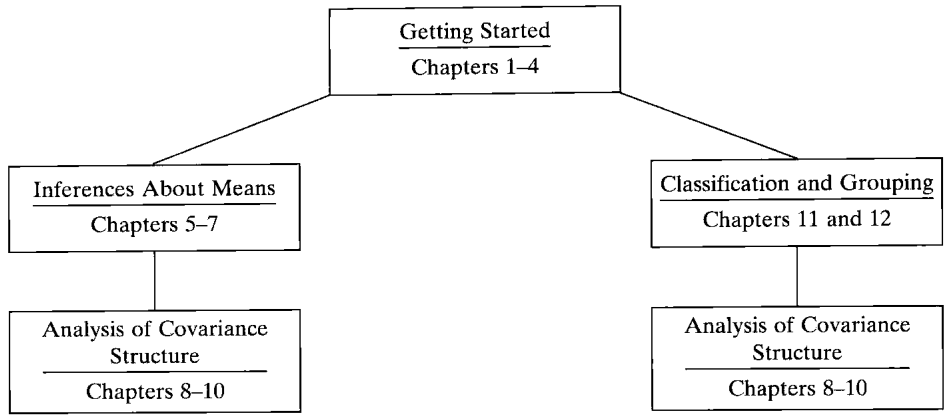
In our attempt to make the study of multivariate analysis appealing to a large audience of both practitioners and theoreticians, we have had to sacrifice a consistency of level. Some sections are harder than others. In particular, we have summarized a voluminous amount of material on regression in Chapter 7. The resulting presentation is rather succinct and difficult the first time through. We hope instructors will be able to compensate for the unevenness in level by judiciously choosing those sections, and subsections, appropriate for their students and by toning them down if necessary.

ORGANIZATION AND APPROACH

The methodological "tools" of multivariate analysis are contained in Chapters 5 through 12. These chapters represent the heart of the book but they cannot be assimilated without much of the material in the introductory Chapters 1 through 4. Even those readers with a good knowledge of matrix algebra or those willing to accept the mathematical results on faith should, at the very least, peruse Chapter 3, Sample Geometry, and Chapter 4, Multivariate Normal Distribution.

Our approach in the methodological chapters is to keep the discussion direct and uncluttered. Typically, we start with a formulation of the population models, delineate the corresponding sample results, and liberally illustrate everything with examples. The examples are of two types: those that are simple and whose calculations can be easily done by hand, and those that rely on real-world data and computer software. These will provide an opportunity to: (1) duplicate our analyses, (2) carry out the analyses dictated by exercises, or (3) analyze the data using methods other than the ones we have used or suggested.

The division of the methodological chapters (5 through 12) into three units allows instructors some flexibility in tailoring a course to their needs. Possible sequences for a one-semester (two quarter) course are indicated schematically.



Each instructor will undoubtedly omit certain sections from some chapters to cover a broader collection of topics than is indicated by these two choices.

For most students, we would suggest a quick pass through the first four chapters (concentrating primarily on the material in Chapter 1, Sections 2.1, 2.2, 2.3, 2.5, 2.6, and 3.6, and the "assessing normality" material in Chapter 4) followed by a selection of methodological topics. For example, one might discuss the comparison of mean vectors, principle components, factor analysis, discriminant analysis, and clustering. The discussions could feature the many "worked out" examples included in these sections of the text. Instructors may rely on diagrams and verbal descriptions to teach the corresponding theoretical developments. If the students have uniformly strong mathematical backgrounds, much of the book can successfully be covered in one term.

We have found individual data-analysis projects useful for integrating material from several of the methods chapters. Here, our rather complete treatments of MANOVA, regression analysis, factor analysis, canonical correlation, discriminant analysis, and so forth are helpful, even though they may not be specifically covered in lectures.

CHANGES TO THE FOURTH EDITION

New Material. Users of the previous editions will notice that we have added and updated some examples and exercises, and have expanded the discussions of viewing multivariate data, generalized variance, assessing normality and transformations to normality, simultaneous confidence intervals, repeated measure designs, and cluster analysis. We have also added a number of new sections including: Detecting Outliers and Data Cleaning (Ch. 4); Multivariate Quality Control Charts, Difficulties Due to Time Dependence in Multivariate Observations (Ch. 5); Repeated Measures Designs and Growth Curves (Ch. 6); Multiple Regression

Models with Time Dependent Errors (Ch. 7); Monitoring Quality with Principal Components (Ch. 8); Correspondence Analysis (Ch. 12); Biplots (Ch. 12); and Procrustes Analysis (Ch. 12). We have worked to improve the exposition throughout the text, and have expanded the t-table in the appendix.

Data Disk. Recognizing the importance of modern statistical packages in the analysis of multivariate data, we have added numerous real-data sets. The full data sets used in the book are saved as ASCII files on the Data Disk which is packaged with each copy of the book. This format will allow easy interface with existing statistical software packages and provide more convenient hands-on data analysis opportunities.

Instructors Solutions Manual. An Instructors Solutions Manual (ISBN 0-13-834202-4) containing complete solutions to most of the exercises in the book is available free upon adoption from Prentice Hall.

For information on additional for sale supplements that may be used with the book or additional titles of interest, please visit the Prentice Hall web site at www.prenhall.com.

ACKNOWLEDGMENTS

We thank our many colleagues who helped improve the applied aspect of the book by contributing their own data sets for examples and exercises. A number of individuals helped guide this revision and we are grateful for their suggestions: Steve Coad, University of Michigan; Richard Kiltie, University of Florida; Sam Kotz, George Mason University; Shyamal Peddada, University of Virginia; K. Sivakumar, University of Illinois at Chicago; Eric Smith, Virginia Tech; and Stanley Wasserman, University of Illinois at Urbana-Champaign. We also acknowledge the feedback of the students we have taught these past 25 years in our applied multivariate analysis courses. Their comments and suggestions are largely responsible for the present iteration of this work. We would also like to give special thanks to Wai Kwong Cheang for his help with the calculations for many of the new examples.

We must thank Deborah Smith for her valuable work on the Data Disk and Solutions Manual, Steve Verrill for computing assistance throughout, and Alison Pollack for implementing a Chernoff faces program. We are indebted to Cliff Gilman for his assistance with the multi-dimensional scaling examples discussed in Chapter 12. Jacquelyn Forer did most of the typing of the original draft manuscript and we appreciate her expertise and willingness to endure the cajoling of authors faced with publication deadlines. Finally we would like to thank Ann Heath, Mindy McClard, Richard DeLorenzo, Brian Baker, Linda Behrens, Alan Fischer, and the rest of the Prentice Hall staff for their help with this project.

R. A. Johnson
D. W. Wichern

Contents

PREFACE *xiii*

PART I Getting Started

1 ASPECTS OF MULTIVARIATE ANALYSIS *1*

1.1 Introduction 1
1.2 Applications of Multivariate Techniques 3
1.3 The Organization of Data 5
1.4 Data Displays and Pictorial Representations 19
1.5 Distance 28
1.6 Final Comments 36
Exercises 36
References 47

2 MATRIX ALGEBRA AND RANDOM VECTORS *49*

2.1 Introduction 49
2.2 Some Basics of Matrix and Vector Algebra 49

2.3 Positive Definite Matrices 61

2.4 A Square-Root Matrix 67

2.5 Random Vectors and Matrices 68

2.6 Mean Vectors and Covariance Matrices 69

2.7 Matrix Inequalities and Maximization 81

Supplement 2A Vectors and Matrices: Basic Concepts 86

Exercises 107

References 115

3 SAMPLE GEOMETRY AND RANDOM SAMPLING 116

3.1 Introduction 116

3.2 The Geometry of the Sample 117

3.3 Random Samples and the Expected Values of the Sample Mean and Covariance Matrix 124

3.4 Generalized Variance 129

3.5 Sample Mean, Covariance, and Correlation as Matrix Operations 145

3.6 Sample Values of Linear Combinations of Variables 148

Exercises 153

References 156

4 THE MULTIVARIATE NORMAL DISTRIBUTION 157

4.1 Introduction 157

4.2 The Multivariate Normal Density and Its Properties 158

4.3 Sampling from a Multivariate Normal Distribution and Maximum Likelihood Estimation 177

4.4 The Sampling Distribution of $\bar{\mathbf{X}}$ and $\mathbf{S}$ 184

4.5 Large-Sample Behavior of $\bar{\mathbf{X}}$ and $\mathbf{S}$ 185

4.6 Assessing the Assumption of Normality 188

4.7 Detecting Outliers and Data Cleaning 200

4.8 Transformations to Near Normality 204

Exercises 214

References 222

PART II Inferences About Multivariate Means and Linear Models

5 INFERENCES ABOUT A MEAN VECTOR ***224***

5.1 Introduction 224

5.2 The Plausibility of μ_0 as a Value for a Normal Population Mean 224

5.3 Hotelling's T^2 and Likelihood Ratio Tests 231

5.4 Confidence Regions and Simultaneous Comparisons of Component Means 235

5.5 Large Sample Inferences about a Population Mean Vector 252

5.6 Multivariate Quality Control Charts 257

5.7 Inferences about Mean Vectors When Some Observations Are Missing 268

5.8 Difficulties Due To Time Dependence in Multivariate Observations 273

Supplement 5A Simultaneous Confidence Intervals and Ellipses as Shadows of the p-Dimensional Ellipsoids 276

Exercises 279

References 288

6 COMPARISONS OF SEVERAL MULTIVARIATE MEANS ***290***

6.1 Introduction 290

6.2 Paired Comparisons and a Repeated Measures Design 291

6.3 Comparing Mean Vectors from Two Populations 302

6.4 Comparing Several Multivariate Population Means (One-Way MANOVA) 314

6.5 Simultaneous Confidence Intervals for Treatment Effects 329

6.6 Two-Way Multivariate Analysis of Variance 331

6.7 Profile Analysis 343

6.8 Repeated Measures Designs and Growth Curves 350

6.9 Perspectives and a Strategy for Analyzing Multivariate Models 355

Exercises 358

References 375

7 MULTIVARIATE LINEAR REGRESSION MODELS ***377***

7.1 Introduction 377

7.2 The Classical Linear Regression Model 377

7.3 Least Squares Estimation 381

7.4 Inferences About the Regression Model 390

7.5 Inferences from the Estimated Regression Function 400

7.6 Model Checking and Other Aspects of Regression 404

7.7 Multivariate Multiple Regression 410

7.8 The Concept of Linear Regression 427

7.9 Comparing the Two Formulations of the Regression Model 438

7.10 Multiple Regression Models with Time Dependent Errors 441

Supplement 7A The Distribution of the Likelihood Ratio for the Multivariate Multiple Regression Model 446

Exercises 448

References 456

PART III Analysis of Covariance Structure

8 PRINCIPAL COMPONENTS ***458***

8.1 Introduction 458

8.2 Population Principal Components 458

8.3 Summarizing Sample Variation by Principal Components 471

8.4 Graphing the Principal Components 484

8.5 Large Sample Inferences 487

8.6 Monitoring Quality with Principal Components 490

Supplement 8A The Geometry of the Sample Principal Component Approximation 498

Exercises 503

References 512

9 FACTOR ANALYSIS AND INFERENCE FOR STRUCTURED COVARIANCE MATRICES 514

9.1 Introduction 514

9.2 The Orthogonal Factor Model 515

9.3 Methods of Estimation 521

9.4 Factor Rotation 540

9.5 Factor Scores 550

9.6 Perspectives and a Strategy for Factor Analysis 557

9.7 Structural Equation Models 565

Supplement 9A Some Computational Details for Maximum Likelihood Estimation 572

Exercises 575

References 585

10 CANONICAL CORRELATION ANALYSIS 587

10.1 Introduction 587

10.2 Canonical Variates and Canonical Correlations 587

10.3 Interpreting the Population Canonical Variables 595

10.4 The Sample Canonical Variates and Sample Canonical Correlations 601

10.5 Additional Sample Descriptive Measures 610

10.6 Large Sample Inferences 615

Exercises 619

References 627

PART IV Classification and Grouping Techniques

11 DISCRIMINATION AND CLASSIFICATION ***629***

11.1 Introduction 629

11.2 Separation and Classification for Two Populations 630

11.3 Classification with Two Multivariate Normal Populations 639

11.4 Evaluating Classification Functions 649

11.5 Fisher's Discriminant Function—Separation of Populations 661

11.6 Classification with Several Populations 665

11.7 Fisher's Method for Discriminating among Several Populations 683

11.8 Final Comments 697

Exercises 703

References 723

12 CLUSTERING, DISTANCE METHODS AND ORDINATION ***726***

12.1 Introduction 726

12.2 Similarity Measures 728

12.3 Hierarchical Clustering Methods 738

12.4 Nonhierarchical Clustering Methods 754

12.5 Multidimensional Scaling 760

12.6 Correspondence Analysis 770

12.7 Biplots for Viewing Sampling Units and Variables 779

12.8 Procrustes Analysis: A Method for Comparing Configurations 782

Exercises 790

References 798

APPENDIX ***800***

Table 1 Standard Normal Probabilities 801

Table 2 Student's *t*-Distribution Percentage Points 802

Table 3 χ^2 Distribution Percentage Points 803

Table 4 *F*-Distribution Percentage Points ($\alpha = .10$) 804

Table 5 *F*-Distribution Percentage Points ($\alpha = .05$) 806

Table 6 *F*-Distribution Percentage Points ($\alpha = .01$) 808

DATA INDEX ***811***

SUBJECT INDEX ***812***

CHAPTER

1

Aspects of Multivariate Analysis

1.1 INTRODUCTION

Scientific inquiry is an iterative learning process. Objectives pertaining to the explanation of a social or physical phenomenon must be specified and then tested by gathering and analyzing data. In turn, an analysis of the data gathered by experimentation or observation will usually suggest a modified explanation of the phenomenon. Throughout this iterative learning process, variables are often added or deleted from the study. Thus, the complexities of most phenomena require an investigator to collect observations on many different variables. This book is concerned with statistical methods designed to elicit information from these kinds of data sets. Because the data include simultaneous measurements on many variables, this body of methodology is called *multivariate analysis*.

The need to understand the relationships between many variables makes multivariate analysis an inherently difficult subject. Often, the human mind is overwhelmed by the sheer bulk of the data. Additionally, more mathematics is required to derive multivariate statistical techniques for making inferences than in a univariate setting. We have chosen to provide explanations based upon algebraic concepts and to avoid the derivations of statistical results that *require* the calculus of many variables. Our objective is to introduce several useful multivariate techniques in a clear manner, making heavy use of illustrative examples and a minimum of mathematics. Nonetheless, some mathematical sophistication and a desire to think quantitatively will be required.

Most of our emphasis will be on the *analysis* of measurements obtained without actively controlling or manipulating any of the variables on which the measurements are made. Only in Chapters 6 and 7 shall we treat a few experimental plans (designs) for generating data that prescribe the active manipulation of important variables. Although the experimental design is ordinarily the most important part of a scientific investigation, it is frequently impossible to control the generation of appropriate data in certain disciplines. (This is true, for exam-

ple, in business, economics, ecology, geology, and sociology.) You should consult [7] and [8] for detailed accounts of design principles that, fortunately, also apply to multivariate situations.

It will become increasingly clear that many multivariate methods are based upon an underlying probability model known as the multivariate normal distribution. Other methods are ad hoc in nature and are justified by logical or commonsense arguments. Regardless of their origin, multivariate techniques must, invariably, be implemented on a computer. Recent advances in computer technology have been accompanied by the development of rather sophisticated statistical software packages, making the implementation step easier.

Multivariate analysis is a "mixed bag." It is difficult to establish a classification scheme for multivariate techniques that both is widely accepted and indicates the appropriateness of the techniques. One classification distinguishes techniques designed to study interdependent relationships from those designed to study dependent relationships. Another classifies techniques according to the number of populations and the number of sets of variables being studied. Chapters in this text are divided into sections according to inference about treatment means, inference about covariance structure, and techniques for sorting or grouping. This should not, however, be considered an attempt to place each method into a slot. Rather, the choice of methods and the types of analyses employed are largely determined by the objectives of the investigation. Below, we list a smaller number of practical problems designed to illustrate the connection between the choice of a statistical method and the objectives of the study. These problems, plus the examples in the text, should provide you with an appreciation for the applicability of multivariate techniques across different fields.

The objectives of scientific investigations to which multivariate methods most naturally lend themselves include the following:

1. *Data reduction or structural simplification.* The phenomenon being studied is represented as simply as possible without sacrificing valuable information. It is hoped that this will make interpretation easier.
2. *Sorting and grouping.* Groups of "similar" objects or variables are created, based upon measured characteristics. Alternatively, rules for classifying objects into well-defined groups may be required.
3. *Investigation of the dependence among variables.* The nature of the relationships among variables is of interest. Are all the variables mutually independent or are one or more variables dependent on the others? If so, how?
4. *Prediction.* Relationships between variables must be determined for the purpose of predicting the values of one or more variables on the basis of observations on the other variables.
5. *Hypothesis construction and testing.* Specific statistical hypotheses, formulated in terms of the parameters of multivariate populations, are tested. This may be done to validate assumptions or to reinforce prior convictions.

We conclude this brief overview of multivariate analysis with a quotation from F. H. C. Marriott [19], page 89. The statement was made in a discussion of cluster analysis, but we feel it is appropriate for a broader range of methods. You should keep it in mind whenever you attempt or read about a data analysis. It allows one to maintain a proper perspective and not be overwhelmed by the elegance of some of the theory:

> If the results disagree with informed opinion, do not admit a simple logical interpretation, and do not show up clearly in a graphical presentation, they are probably wrong. There is no magic about numerical methods, and many ways in which they can break down. They are a valuable aid to the interpretation of data, not sausage machines automatically transforming bodies of numbers into packets of scientific fact.

1.2 APPLICATIONS OF MULTIVARIATE TECHNIQUES

The published applications of multivariate methods have increased tremendously in recent years. It is now difficult to cover the variety of real-world applications of these methods with brief discussions, as we did in earlier editions of this book. However, in order to give some indication of the usefulness of multivariate techniques, we offer the following short descriptions of the results of studies from several disciplines. These descriptions are organized according to the categories of objectives given in the previous section. Of course, many of our examples are multifaceted and could be placed in more than one category.

Data reduction or simplification

- Using data on several variables related to cancer patient responses to radiotherapy, a simple measure of patient response to radiotherapy was constructed. (See Exercise 1.15.)
- Track records from many nations were used to develop an index of performance for both male and female athletes. (See [10] and [21].)
- Multispectral image data collected by a high-altitude scanner were reduced to a form that could be viewed as images (pictures) of a shoreline in two dimensions. (See [22].)
- Data on several variables relating to yield and protein content were used to create an index to select parents of subsequent generations of improved bean plants. (See [14].)

Sorting and grouping

- Data on several variables related to computer use were employed to create clusters of categories of computer jobs that allow a better determination of existing (or planned) computer utilization. (See [2].)