Applied Multivariate Techniques

Subhash Sharma
University of South Carolina

John Wiley & Sons, Inc.

New York Chichester Brisbane Toronto Singapore

QA
278
S485

Acquisitions Editors	Tim Kent and Petra Sellers
Assistant Editor	Ellen Ford
Marketing Manager	Leslie Hines
Production Editor	Jennifer Knapp
Cover and Text Design	Harry Nolan
Cover Photograph	Telegraph Colour Library/FPG International Corp.
Manufacturing Manager	Mark Cirillo
Illustration Coordinator	Edward Starr
Outside Production Manager	J. Carey Publishing Service

This book was set in 10/12 Times Roman by Publication Services and printed and bound by Courier Stoughton. The cover was printed by Lehigh Press.

Recognizing the importance of preserving what has been written, it is a policy of John Wiley & Sons, Inc. to have books of enduring value published in the United States printed on acid-free paper, and we exert our best efforts to that end.

Copyright ©1996 by John Wiley & Sons, Inc.

All rights reserved. Published simultaneously in Canada.

Reproduction or translation of any part of this work beyond that permitted by Sections 107 and 108 of the 1976 United States Copyright Act without the permission of the copyright owner is unlawful. Requests for permission or further information should be addressed to the Permissions Department, John Wiley & Sons, Inc.

Library of Congress Cataloging in Publication Data:
Sharma, Subhash.
Applied multivariate techniques / Subhash Sharma.
p. cm.
Includes bibliographical references.
ISBN 0-471-31064-6 (cloth : alk. paper)
1. Multivariate analysis. I. Title.
QA278.S485 1996
519.5'35–dc20 95-12400
CIP

Printed in the United States of America

10 9 8 7 6 5 4 3

Dedication

Dedicated to my students, my parents,
my wife, Swaran,
and my children, Navin and Nikhil

Preface

This book is the result of many years of teaching graduate courses on multivariate statistics. The students in these courses were primarily from business, sciences, and behavioral sciences, and were interested in getting a good working knowledge of the multivariate data analytic techniques without getting bogged down with derivations and/or rigorous proofs. That is, consistent with the needs of today's managers, the students were more interested in knowing when to correctly use a particular technique and its interpretation rather than the mechanics of the technique. The available textbooks were either too technical or were too applied and cookbook in nature. The technical books concentrated more on derivation of the techniques and less on interpretation of the results. On the other hand, books with a cookbook approach did not provide much discussion of the techniques and essentially provided a laundry list of the dos and don'ts. This motivated me to develop notes for the various topics that would emphasize the concepts of a given technique and its application without using matrix algebra and proofs. Extensive in-class testing and refining of these notes resulted in this book.

My approach here is to make statistics a "kinder and gentler" subject by introducing students to the various multivariate techniques used in businesses without intimidating them with mathematical derivations. The main emphasis is on *when to use* the various data analytic techniques and *how to interpret* the resulting output obtained from the most widely used statistical packages (e.g., **SPSS** and **SAS**). This book achieves these objectives using the following strategy.

ORGANIZATION

Most of the chapters are divided into two parts, the text and an appendix. The text provides a conceptual understanding of technique, with basic concepts illustrated by a small hypothetical data set and geometry. Geometry is very effective in providing a clear, concise, and nonmathematical treatment of the technique. However, because some students are unfamiliar with geometrical concepts and data manipulations, Chapter 2 covers the basic high-school level geometrical concepts used throughout the book, and Chapter 3 discusses fundamental data manipulation techniques.

Next, wherever appropriate, the same data set is used to provide an analytical discussion of the technique. This analytical approach essentially reinforces the concepts discussed using geometry. Again, high-school level math is used in the chapter, and no matrix algebra or higher-level math are employed. This is followed by using the same hypothetical data to obtain the output from either **SPSS** or **SAS**. A detailed discussion of the interpretation of the output is provided and, whenever necessary, computations of

the various interpretive statistics reported in the output are illustrated in order to give a better understanding of these statistics and their use in interpreting the results. Finally, wherever necessary, an actual data set is used to illustrate application of the technique.

Most of the chapters also contain an appendix. The appendices are technical in nature, and are meant for students who already have taken a basic course in linear algebra. However, the chapters are completely independent of the appendices and on their own provide a solid understanding of the basic concepts of a given technique, and how to meaningfully interpret statistical output.

We have not provided an appendix or a chapter to review matrix algebra because it simply is not needed. The student can become a sophisticated user of the technique without a working knowledge of matrix algebra. Furthermore, the discussion provided in typical review chapters is almost always insufficient for those who have never had a formal course on matrix algebra. And those who have had a formal course in matrix algebra are better served by reviewing the appropriate matrix algebra textbook.

TOPICS COVERED

The multivariate techniques covered in this book are divided into two categories: interdependence and dependence techniques (Chapter 1 provides a detailed distinction between the two types of techniques). In the interdependence methods no distinction is made between dependent and independent variables and as such the focus is on analyzing information contained in one large set of variables. For the dependence techniques, a distinction is made between one set of variables (normally referred to as independent variables) and another set of variables (normally referred to as dependent variables) and the focus is on how the two sets of variables are related. The chapters in the book are organized such that all the interdependence methods are covered first, followed by the dependence methods. Principal components analysis, factor analysis, confirmatory factor analysis, and cluster analysis are the interdependence topics covered in this text. The dependence techniques covered are two-group and multiple-group discriminant analysis, logistic regression analysis, multivariate analysis of variance, canonical correlation, and structural equations. Many of these techniques make a number of assumptions, such as data coming from a multivariate normal distribution and equality of groups with respect to variances and covariances. Chapter 12 discusses the procedures used to test these assumptions.

SUPPLEMENTAL MATERIALS

Many of the end-of-chapter exercises require hand calculations or spreadsheets and are designed to reinforce the concepts discussed in the chapter; others require the use of statistical packages to analyze the accompanying data sets. The enclosed *data diskette* contains data sets used in the book and the end-of-chapter exercises. To further enhance learning, the reader can analyze the data sets using other statistical software and compare the results to those reported in the book. However, it should be noted that the best learning takes place through the use of data sets with which students are familiar and that are from their own fields of study. Consequently, it is recommended that the reader also obtain data sets from their disciplines and analyze them using the appropriate techniques.

An *Instructor's Manual* to accompany the text offers detailed answers to all end-of-chapter exercises, including computer output for questions that require students to perform data analysis. The *Instructor's Manual* also contains transparency masters for all the exhibits.

ACKNOWLEDGMENTS

This book would not have been possible without the help and encouragement provided by many people. First and foremost, I would like to thank the numerous graduate and doctoral students who provided comments on the initial drafts of the various chapters that were used as class notes during the last ten years. Their insightful comments led to numerous rewrites which substantially improved the clarity and readability of the book. To them I am greatly indebted. I would like to thank Soumen Mukherjee and Anthony Miyazaki, both doctoral students at the University of South Carolina, for numerous readings of the manuscript and helping me prepare the end-of-chapter questions and the *Instructor's Manual.* I would also like to thank numerous colleagues who spent countless hours reading the initial drafts, providing valuable comments and insights, and for using various chapters as supplemental material in their classes. I am particularly indebted to Professors Terence A. Shimp and William O. Bearden both of the University of South Carolina, Donald Liechtenstein, University of Colorado at Boulder, and George Franke, University of Alabama. Special thanks go to Professor Srinivas Durvasula, Marquette University, for using the entire draft in his MBA class and providing detailed student comments. I would also like to thank Professors Barry Babin, University of Southern Mississippi, John Lastovicka, Arizona State University, Jagdip Singh, Case Western Reserve University, and Phil Wirtz, Georgetown University, for reviewing the book and providing valuable comments and insights, which substantially improved the book.

Thanks are also due to Jennie Smyrl and Edie Beaver, administrative assistants in the College of Business, University of South Carolina, for patiently and skillfully handling the numerous demands placed on them. I am also thankful to the College of Business and the University of South Carolina's administration for granting me the sabbatical in 1994 that allowed me the time to concentrate on finishing the book.

Much thanks are also due to the excellent support provided by John Wiley & Sons. Special thanks to Tim Kent who over the years worked with me in putting together the proposal, Whitney Blake, executive editor, and Ellen Ford, assistant editor, for providing invaluable advice and comments. I am also thankful to Jenni Knapp and Jennifer Carey for their help during the various production stages. Thanks are also due to Ed Starr and Harry Nolan for putting together the excellent illustrations and designs used. Finally, I would like to thank my wife and children who were a constant source of inspiration and provided the impetus for starting and finishing the book.

Contents

CHAPTER 1 INTRODUCTION 1

1.1 Types of Measurement Scales 1
1.1.1 Nominal Scale 2
1.1.2 Ordinal Scale 2
1.1.3 Interval Scale 2
1.1.4 Ratio Scale 3
1.1.5 Number of Variables 3
1.2 Classification of Data Analytic Methods 4
1.3 Dependence Methods 5
1.3.1 One Dependent and One Independent Variable 5
1.3.2 One Dependent Variable and More Than One Independent Variable 5
1.3.3 More Than One Dependent and One or More Independent Variables 9
1.4 Interdependence Methods 10
1.4.1 Metric Variables 11
1.4.2 Nonmetric Data 12
1.5 Structural Models 13
1.6 Overview of the Book 14
Questions 15

CHAPTER 2 GEOMETRIC CONCEPTS OF DATA MANIPULATION 17

2.1 Cartesian Coordinate System 17
2.1.1 Change in Origin and Axes 18
2.1.2 Euclidean Distance 19
2.2 Vectors 19
2.2.1 Geometric View of the Arithmetic Operations on Vectors 20
2.2.2 Projection of One Vector onto Another Vector 23
2.3 Vectors in a Cartesian Coordinate System 23
2.3.1 Length and Direction Cosines 24
2.3.2 Standard Basis Vectors 25
2.4 Algebraic Formulae for Vector Operations 25
2.4.1 Arithmetic Operations 25
2.4.2 Linear Combination 26
2.4.3 Distance and Angle between Any Two Vectors 27
2.4.4 Scalar Product and Vector Projections 27
2.4.5 Projection of a Vector onto Subspace 28
2.4.6 Illustrative Example 29

2.5 Vector Independence and Dimensionality 30
2.5.1 Dimensionality 30
2.6 Change in Basis 31
2.7 Representing Points with Respect to New Axes 32
2.8 Summary 33
Questions 34

CHAPTER 3 FUNDAMENTALS OF DATA MANIPULATION 36

3.1 Data Manipulations 36
3.1.1 Mean and Mean-Corrected Data 36
3.1.2 Degrees of Freedom 36
3.1.3 Variance, Sum of Squares, and Cross Products 38
3.1.4 Standardization 39
3.1.5 Generalized Variance 39
3.1.6 Group Analysis 40
3.2 Distances 42
3.2.1 Statistical Distance 42
3.2.2 Mahalanobis Distance 44
3.3 Graphical Representation of Data in Variable Space 45
3.4 Graphical Representation of Data in Observation Space 47
3.5 Generalized Variance 50
3.6 Summary 51
Questions 52
Appendix 54
A3.1 Generalized Variance 54
A3.2 Using PROC IML in SAS for Data Manipulations 55

CHAPTER 4 PRINCIPAL COMPONENTS ANALYSIS 58

4.1 Geometry of Principal Components Analysis 59
4.1.1 Identification of Alternative Axes and Forming New Variables 59
4.1.2 Principal Components Analysis as a Dimensional Reducing Technique 64
4.1.3 Objectives of Principal Components Analysis 66
4.2 Analytical Approach 66
4.3 How To Perform Principal Components Analysis 67
4.3.1 SAS Commands and Options 67
4.3.2 Interpreting Principal Components Analysis Output 68
4.4 Issues Relating to the Use of Principal Components Analysis 71
4.4.1 Effect of Type of Data on Principal Components Analysis 72
4.4.2 Is Principal Components Analysis the Appropriate Technique? 75
4.4.3 Number of Principal Components to Extract 76
4.4.4 Interpreting Principal Components 79
4.4.5 Use of Principal Components Scores 80
4.5 Summary 81
Questions 81
Appendix 84
A4.1 Eigenstructure of the Covariance Matrix 84

A4.2 Singular Value Decomposition 85
A4.2.1 Singular Value Decomposition of the Data Matrix 85
A4.3 Spectral Decomposition of a Matrix 86
A4.3.1 Spectral Decomposition of the Covariance Matrix 86
A4.4 Illustrative Example 87

CHAPTER 5 FACTOR ANALYSIS 90

5.1 Basic Concepts and Terminology of Factor Analysis 90
5.1.1 Two-Factor Model 93
5.1.2 Interpretation of the Common Factors 96
5.1.3 More Than Two Factors 96
5.1.4 Factor Indeterminacy 97
5.2 Objectives of Factor Analysis 99
5.3 Geometric View of Factor Analysis 99
5.3.1 Estimation of Communalities Problem 100
5.3.2 Factor Rotation Problem 100
5.3.3 More Than Two Factors 102
5.4 Factor Analysis Techniques 102
5.4.1 Principal Components Factoring (PCF) 103
5.4.2 Principal Axis Factoring 107
5.4.3 Which Technique Is the Best? 108
5.4.4 Other Estimation Techniques 108
5.5 How to Perform Factor Analysis 109
5.6 Interpretation of SAS Output 110
5.6.1 Are the Data Appropriate for Factor Analysis? 116
5.6.2 How Many Factors? 116
5.6.3 The Factor Solution 117
5.6.4 How Good Is the Factor Solution? 118
5.6.5 What Do the Factors Represent? 118
5.6.6 Rotation 119
5.7 An Empirical Illustration 121
5.7.1 Identifying and Evaluating the Factor Solution 123
5.7.2 Interpreting the Factor Structure 125
5.8 Factor Analysis versus Principal Components Analysis 125
5.9 Exploratory versus Confirmatory Factor Analysis 128
5.10 Summary 129
Questions 129
Appendix 132
A5.1 One-Factor Model 132
A5.2 Two-Factor Model 133
A5.3 More Than Two Factors 135
A5.4 Factor Indeterminacy 136
A5.4.1 Communality Estimation Problem 136
A5.4.2 Factor Rotation Problem 136
A5.5 Factor Rotations 137
A5.5.1 Orthogonal Rotation 137
A5.6 Factor Extraction Methods 141
A5.6.1 Principal Components Factoring (PCF) 141
A5.6.2 Principal Axis Factoring (PAF) 142
A5.7 Factor Scores 142

CHAPTER 6 CONFIRMATORY FACTOR ANALYSIS 144

6.1 Basic Concepts of Confirmatory Factor Analysis 144
6.1.1 Covariance or Correlation Matrix? 144
6.1.2 One-Factor Model 145
6.1.3 Two-Factor Model with Correlated Constructs 147
6.2 Objectives of Confirmatory Factor Analysis 148
6.3 LISREL 148
6.3.1 LISREL Terminology 148
6.3.2 LISREL Commands 150
6.4 Interpretation of the LISREL Output 152
6.4.1 Model Information and Parameter Specifications 152
6.4.2 Initial Estimates 152
6.4.3 Evaluating Model Fit 157
6.4.4 Evaluating the Parameter Estimates and the Estimated Factor Model 162
6.4.5 Model Respecification 164
6.5 Multigroup Analysis 170
6.6 Assumptions 173
6.7 An Illustrative Example 174
6.8 Summary 176
Questions 177
Appendix 180
A6.1 Squared Multiple Correlations 181
A6.2 Maximum Likelihood Estimation 181

CHAPTER 7 CLUSTER ANALYSIS 185

7.1 What Is Cluster Analysis? 185
7.2 Geometrical View of Cluster Analysis 186
7.3 Objective of Cluster Analysis 187
7.4 Similarity Measures 187
7.5 Hierarchical Clustering 188
7.5.1 Centroid Method 188
7.5.2 Single-Linkage or the Nearest-Neighbor Method 191
7.5.3 Complete-Linkage or Farthest-Neighbor Method 192
7.5.4 Average-Linkage Method 192
7.5.5 Ward's Method 193
7.6 Hierarchical Clustering Using SAS 194
7.6.1 Interpreting the SAS Output 195
7.7 Nonhierarchical Clustering 202
7.7.1 Algorithm I 203
7.7.2 Algorithm II 205
7.7.3 Algorithm III 205
7.8 Nonhierarchical Clustering Using SAS 207
7.8.1 Interpreting the SAS Output 208
7.9 Which Clustering Method Is Best? 211
7.9.1 Hierarchical Methods 211
7.9.2 Nonhierarchical Methods 217

7.10 Similarity Measures 218
7.10.1 Distance Measures 218
7.11 Reliability and External Validity of a Cluster Solution 221
7.11.1 Reliability 221
7.11.2 External Validity 221
7.12 An Illustrative Example 221
7.12.1 Hierarchical Clustering Results 221
7.12.2 Nonhierarchical Clustering Results 228
7.13 Summary 232
Questions 233
Appendix 235

CHAPTER 8 TWO-GROUP DISCRIMINANT ANALYSIS 237

8.1 Geometric View of Discriminant Analysis 237
8.1.1 Identifying the "Best" Set of Variables 238
8.1.2 Identifying a New Axis 239
8.1.3 Classification 242
8.2 Analytical Approach to Discriminant Analysis 244
8.2.1 Selecting the Discriminator Variables 244
8.2.2 Discriminant Function and Classification 245
8.3 Discriminant Analysis Using SPSS 245
8.3.1 Evaluating the Significance of Discriminating Variables 246
8.3.2 The Discriminant Function 250
8.3.3 Classification Methods 254
8.3.4 Histograms for the Discriminant Scores 262
8.4 Regression Approach to Discriminant Analysis 262
8.5 Assumptions 263
8.5.1 Multivariate Normality 263
8.5.2 Equality of Covariance Matrices 264
8.6 Stepwise Discriminant Analysis 264
8.6.1 Stepwise Procedures 265
8.6.2 Selection Criteria 265
8.6.3 Cutoff Values for Selection Criteria 266
8.6.4 Stepwise Discriminant Analysis Using SPSS 267
8.7 External Validation of the Discriminant Function 273
8.7.1 Holdout Method 273
8.7.2 U-Method 273
8.7.3 Bootstrap Method 274
8.8 Summary 274
Questions 275
Appendix 277
A8.1 Fisher's Linear Discriminant Function 277
A8.2 Classification 278
A8.2.1 Statistical Decision Theory Method for Developing Classification Rules 279
A8.2.2 Classification Rules for Multivariate Normal Distributions 281
A8.2.3 Mahalanobis Distance Method 283
A8.3 Illustrative Example 284
A8.3.1 Any Known Distribution 284
A8.3.2 Normal Distribution 285

CHAPTER 9 MULTIPLE-GROUP DISCRIMINANT ANALYSIS 287

9.1 Geometrical View of MDA 287
9.1.1 How Many Discriminant Functions Are Needed? 288
9.1.2 Identifying New Axes 289
9.1.3 Classification 293
9.2 Analytical Approach 293
9.3 MDA Using SPSS 294
9.3.1 Evaluating the Significance of the Variables 294
9.3.2 The Discriminant Function 294
9.3.3 Classification 303
9.4 An Illustrative Example 304
9.4.1 Labeling the Discriminant Functions 307
9.4.2 Examining Differences in Brands 307
9.5 Summary 308
Questions 309
Appendix 310
A9.1 Classification for More than Two Groups 311
A9.1.1 Equal Misclassification Costs 311
A9.1.2 Illustrative Example 312
A9.2 Multivariate Normal Distribution 312
A9.2.1 Classification Regions 313
A9.2.2 Mahalanobis Distance 315

CHAPTER 10 LOGISTIC REGRESSION 317

10.1 Basic Concepts of Logistic Regression 317
10.1.1 Probability and Odds 317
10.1.2 The Logistic Regression Model 319
10.2 Logistic Regression with Only One Categorical Variable 321
10.2.1 Model Information 321
10.2.2 Assessing Model Fit 323
10.2.3 Parameter Estimates and Their Interpretation 324
10.2.4 Association of Predicted Probabilities and Observed Responses 325
10.2.5 Classification 326
10.3 Logistic Regression and Contingency Table Analysis 327
10.4 Logistic Regression for Combination of Categorical and Continuous Independent Variables 328
10.4.1 Stepwise Selection Procedure 329
10.5 Comparison of Logistic Regression and Discriminant Analysis 332
10.6 An Illustrative Example 333
10.7 Summary 335
Questions 336
Appendix 339
A10.1 Maximum Likelihood Estimation 339
A10.2 Illustrative Example 340

CHAPTER 11 MULTIVARIATE ANALYSIS OF VARIANCE 342

11.1 Geometry of MANOVA 342
11.1.1 One Independent Variable at Two Levels and One Dependent Variable 343

11.1.2 One Independent Variable at Two Levels and Two or More Dependent Variables 343
11.1.3 More Than One Independent Variable and p Dependent Variables 344
11.2 Analytic Computations for Two-Group MANOVA 346
11.2.1 Significance Tests 346
11.2.2 Effect Size 348
11.2.3 Power 349
11.2.4 Similarities between MANOVA and Discriminant Analysis 350
11.3 Two-Group MANOVA 350
11.3.1 Cell Means and Homogeneity of Variances 351
11.3.2 Multivariate Significance Tests and Power 351
11.3.3 Univariate Significance Tests and Power 353
11.3.4 Multivariate and Univariate Significance Tests 353
11.4 Multiple-Group MANOVA 355
11.4.1 Multivariate and Univariate Effects 356
11.4.2 Orthogonal Contrasts 356
11.5 MANOVA for Two Independent Variables or Factors 366
11.5.1 Significance Tests for the $GENDER \times AD$ Interaction 367
11.6 Summary 370
Questions 371

CHAPTER 12 ASSUMPTIONS 374

12.1 Significance and Power of Test Statistics 374
12.2 Normality Assumptions 375
12.3 Testing Univariate Normality 375
12.3.1 Graphical Tests 376
12.3.2 Analytical Procedures for Assessing Univariate Normality 378
12.3.3 Assessing Univariate Normality Using SPSS 378
12.4 Testing for Multivariate Normality 380
12.4.1 Transformations 382
12.5 Effect of Violating the Equality of Covariance Matrices Assumption 383
12.5.1 Tests for Checking Equality of Covariance Matrices 385
12.6 Independence of Observations 387
12.7 Summary 388
Questions 388
Appendix 389

CHAPTER 13 CANONICAL CORRELATION 391

13.1 Geometry of Canonical Correlation 391
13.1.1 Geometrical Illustration in the Observation Space 397
13.2 Analytical Approach to Canonical Correlation 397
13.3 Canonical Correlation Using SAS 398
13.3.1 Initial Statistics 401
13.3.2 Canonical Variates and the Canonical Correlation 401
13.3.3 Statistical Significance Tests for the Canonical Correlations 402
13.3.4 Interpretation of the Canonical Variates 404
13.3.5 Practical Significance of the Canonical Correlation 404

13.4 Illustrative Example 406
13.5 External Validity 409
13.6 Canonical Correlation Analysis as a General Technique 409
13.7 Summary 409
Questions 410
Appendix 412
A13.1 Effect of Change in Scale 415
A13.2 Illustrative Example 415

CHAPTER 14 COVARIANCE STRUCTURE MODELS 419

14.1 Structural Models 419
14.2 Structural Models with Observable Constructs 420
14.2.1 Implied Matrix 420
14.2.2 Representing Structural Equations as LISREL Models 421
14.2.3 An Empirical Illustration 422
14.3 Structural Models with Unobservable Constructs 426
14.3.1 Empirical Illustration 428
14.4 An Illustrative Example 435
14.4.1 Assessing the Overall Model Fit 435
14.4.2 Assessing the Measurement Model 437
14.5 Summary 440
Questions 440
Appendix 444
A14.1 Implied Covariance Matrix 444
A14.1.1 Models with Observable Constructs 444
A14.1.2 Models with Unobservable Constructs 446
A14.2 Model Effects 449
A14.2.1 Effects among the Endogenous Constructs 450
A14.2.2 Effects of Exogenous Constructs on Endogenous Constructs 452
A14.2.3 Effects of the Constructs on Their Indicators 452

STATISTICAL TABLES 455

REFERENCES 469

TABLES, FIGURES, AND EXHIBITS 473

INDEX 483

CHAPTER 1

Introduction

Let the data speak! There are a number of different statistical techniques that can be used to analyze the data. Obviously, the objective of data analysis is to extract the relevant information contained in the data which can then be used to solve a given problem.[1] The given problem is normally formulated into one or more null hypotheses. The collected sample data are used to statistically test for the rejection or nonrejection of the null hypotheses, which leads to the solution of the problem. That is, the null hypotheses represent the problem and the "relevant information" contained in the data are used to statistically test the null hypotheses. The purpose of this chapter is to give the reader a brief overview of the different techniques that are available to extract relevant information contained in the data set (i.e., to test the null hypotheses representing a given problem). A number of classification schemes exist for classifying the statistical techniques. The following section discusses one such classification scheme. For an example of other classification schemes see Andrews et al. (1981). Since most of the classification schemes, including the one discussed in this chapter, are based on types of measurement scales and the number of variables, we first provide a brief discussion of these topics.

1.1 TYPES OF MEASUREMENT SCALES

Measurement is a process by which numbers or symbols are attached to given characteristics or properties of stimuli according to predetermined rules or procedures. For example, individuals can be described with respect to a number of characteristics such as age, education, income, gender, and brand preferences. Appropriate measurement scales can be used to measure these characteristics. Stevens (1946) postulated that all measurement scales can be classified into the following four types: nominal, ordinal, interval, and ratio. This typology for classifying measurement scales has been adopted in social and behavioral sciences. Following is a brief discussion of the four types of scales. However, we would like to caution the reader that considerable debate, without a clear resolution, regarding the use of Stevens's typology for classifying measurement scales has appeared in the statistical literature (see Velleman and Wilkinson (1993) for further details).

[1] The term *information* is used very loosely and may not necessarily have the same meaning as in *information theory*.

1.1.1 Nominal Scale

Consider the gender variable. Typically we use numerals (although this is not necessary) to represent subjects' genders. For example, we can arbitrarily assign number 1 for males and number 2 for females. The assigned numbers themselves do not have any meaning, and therefore it would be inappropriate to compute such statistics as mean and standard deviation of the gender variable. The numbers are simply used for categorizing subjects into different groups or for counting how many are in each category. Such measurement scales are called nominal scales and the resulting data are called nominal data. The statistics that are appropriate for nominal scales are the ones based on counts such as mode and frequency distributions.

1.1.2 Ordinal Scale

Suppose we want to measure subjects' preferences for four brands of colas, Brands A, B, C, and D. We could ask each subject to rank order the four brands by assigning a 1 to the most preferred brand, a 2 to the next most preferred brand, and so on. Consider the following rank ordering given by one particular subject.

Brand	Rank
A	1
C	2
D	3
B	4

From the preceding table we conclude that the subject prefers Brand A to Brand C, Brand C to Brand D, and Brand D to Brand B. However, even though the differences in the successive numerical values of the ranks are equal, we cannot state by how much the subject prefers one brand over another brand. That is, successive categories do not represent equal differences of the measured attribute. Such measurement scales are referred to as ordinal scales and the resulting data are called ordinal data. Valid statistics that can be computed for ordinal-scaled data are mode, median, frequency distributions, and nonparametric statistics such as rank order correlation. For further details on nonparametric statistics the reader is referred to Segal (1967). Variables measured using nominal and ordinal scales are commonly referred to as nonmetric variables.

1.1.3 Interval Scale

Suppose that instead of asking subjects to rank order the brands, we ask them to rate their brand preference according to the following five-point scale:

Scale Point	Preference
1	Very high preference
2	High preference
3	Moderate preference
4	Low preference
5	Very low preference

If we assume that successive categories represent equal degrees of preference, then we could say that the difference in a subject's preference for the two brands that received ratings of 1 and 2 is the same as the difference in the subject's preference for two other brands that received ratings of 4 and 5. However, we still cannot say that the subject's preference for a brand that received a rating of 5 is five times the preference for a brand that received a rating of 1. The following example clarifies this point. Suppose we multiply each rating point by 2 and then add 10. This would result in the following transformed scale:

Scale Point	Preference
12	Very high preference
14	High preference
16	Moderate preference
18	Low preference
20	Very low preference

From the preceding table it is clear that the differences between the successive categories are equal; however, the ratio of the last to the first category is not the same as that for the original scale. The ratio is 5 for the original scale and 1.67 for the transformed scale. This is because by adding a constant we have changed the value of the base category (i.e., very low preference). The scale does not have a natural base value or point. That is, the base value is arbitrary. Measurement scales whose successive categories represent equal levels of the characteristic that is being measured and whose base values are arbitrary are called interval scales, and the resulting data are called interval data. Properties of the interval scale are preserved under the following transformation:

$$Y_t = a + bY_o$$

where Y_o and Y_t, respectively, are the original and transformed-scale values and a and b are constants. All statistics, except the ones based on ratios such as the coefficient of variation, can be computed for interval-scaled data.

1.1.4 Ratio Scale

Ratio scales, in addition to having all the properties of the interval scale, have a natural base value that cannot be changed. For example, a subject's age has a natural base value, which is zero. Ratio scales can be transformed by multiplying by a constant; however, they cannot be transformed by adding a constant as this will change the base value. That is, only the following transformation is valid for the ratio scale:

$$Y_t = bY_o.$$

Since a ratio scale has a natural base value, statements such as "Subject A's age is twice Subject B's age" are valid. Data resulting from ratio scales are referred to as ratio data. There is no restriction on the kind of statistics that can be computed for ratio-scaled data. Variables measured using interval and ratio scales are called metric variables.

1.1.5 Number of Variables

For ordinal-, interval-, and ratio-scaled data, determining the number of variables is straightforward. The number of variables is simply equal to the number of variables

used to measure the respective characteristics. However, the procedure for determining the number of variables for nominal scales is quite different from that for the other types of scales. Consider, for example, the case where a researcher is interested in determining the effect of gender, a nominal variable, on coffee consumption. The two levels of gender, male and female, can be numerically represented by one dummy or binary variable, D_1. Arbitrarily, a value of 0 may be assigned to D_1 for all male subjects and a value of 1 for all female subjects. That is, the nominal variable, gender, is measured by one dummy variable.

Now suppose that the researcher is interested in determining the effect of a subject's occupation (i.e., professional, technical, or blue collar) on his/her coffee consumption. The nominal variable, occupation, cannot be represented by one dummy variable. As shown, two dummy variables are required:

	Dummy Variables	
Occupation	D_1	D_2
Professional	0	0
Technical	0	1
Blue collar	1	0

That is, occupation, a single nominal variable, is measured by the two dummy variables D_1 and D_2. In yet another example, suppose the researcher is interested in determining the effect of gender *and* occupation on coffee consumption. Three dummy or binary variables (one for gender and two for occupation) are needed to represent the two nominal variables. Therefore, the *number of variables* for nominal variables is equal to the number of dummy variables needed to represent them.

1.2 CLASSIFICATION OF DATA ANALYTIC METHODS

Consider a data set consisting of n observations on p variables. Further assume that the p variables can be divided into two groups or subsets. Statistical methods for analyzing these types of data sets are referred to as *dependence methods*. The dependence methods test for the presence or absence of relationships between the two sets of variables. However, if the researcher, based on controlled experiments and/or some relevant theory, designates variables in one subset as independent variables and variables in the other subset as dependent variables, then the objective of the dependence methods is to determine whether the set of independent variables affects the set of dependent variables individually and/or jointly. That is, statistical techniques only test for the presence or absence of relationships between two sets of variables. Whether the presence of the relationship is due to one set of variables affecting another set of variables or to some other phenomenon can only be established by following the scientific procedures for establishing cause-and-effect relationships (e.g., controlled experimentation). In this and all subsequent chapters, the use of cause-and-effect terms implies that proper scientific principles for establishing cause-and-effect relationships have been followed.

On the other hand, data sets do exist for which it is impossible to conceptually designate one set of variables as dependent and another set of variables as independent. For these types of data sets the objectives are to identify *how* and *why* the variables are related among themselves. Statistical methods for analyzing these types of data sets are called *interdependence* methods.

1.3 DEPENDENCE METHODS

Dependence methods can be further classified according to:

1. The number of independent variables—one or more than one.
2. The number of dependent variables—one or more than one.
3. The type of measurement scale used for the dependent variables (i.e., metric or nonmetric).
4. The type of measurement scale used for the independent variables (i.e., metric or nonmetric).

Table 1.1 gives a list of the statistical methods classified according to the above criteria. A brief discussion of the statistical methods listed in Table 1.1 is provided in the following sections.

1.3.1 One Dependent and One Independent Variable

Statistical methods for a single independent and a single dependent variable are often referred to as *univariate methods*, whereas statistical methods for data sets with more than one independent and/or more than one dependent variable are classified as *multivariate methods*. Univariate methods are special cases of multivariate methods. Therefore, the univariate methods are discussed along with their multivariate counterparts.

1.3.2 One Dependent Variable and More Than One Independent Variable

Consider the example where the marketing manager of a firm is interested in determining the relationship between the dependent variable, Purchase Intention (PI), and the following independent variables: income (I), education (E), age (A), and lifestyle (L). This purchase-behavior example is used to discuss the similarities and differences among the data analytic techniques. Wherever necessary, additional examples are provided to further illustrate the different techniques.

For the purchase-behavior example, the relationship between the dependent variable, PI, and the independent variables can be represented by the following linear model:

$$PI = \beta_0 + \beta_1 I + \beta_2 E + \beta_3 A + \beta_4 L + \epsilon. \tag{1.1}$$

One of the objectives of a given technique is to estimate the parameters β_0, β_1, β_2, β_3, and β_4 of the above model. Most of the dependence method techniques discussed below are special cases of the linear model given by Eq. 1.1.

Regression

Multiple regression is used when the dependent variable and the multiple independent variables of the model given in Eq. 1.1 are measured using a metric scale resulting in metric data such as income measured in dollars. Simple regression is a special name for multiple regression when there is only one independent variable. For example, simple regression would be used if the manager were interested in determining the relationship between PI and I.

$$PI = \beta_0 + \beta_1 I + \epsilon$$

Table 1.1 **Dependence Statistical Methods**

	Dependent Variable(s)			
	One		More than One	
	Metric	Nonmetric	Metric	Nonmetric
Independent Variable(s)				
One				
Metric	• Regression	• Discriminant analysis • Logistic regression	• Canonical correlation	• Multiple-group discriminant analysis (MDA)
Nonmetric	• *t*-test	• Discrete discriminant analysis	• MANOVA (multivariate analysis of variance)	• Discrete MDA
More than One				
Metric	• Multiple regression	• Discriminant analysis • Logistic regression	• Canonical correlation	• MDA
Nonmetric	• ANOVA	• Discrete discriminant analysis • Conjoint analysis (MONANOVA)	• MANOVA	• Discrete MDA

Analysis of Variance

In many situations, a nominal scale is used to measure the independent variables. For instance, rather than obtaining the subjects' exact incomes the researcher can categorize the subjects as having high, medium, or low incomes. Table 1.2 gives an example of how nominal or categorical variables can be used to measure the independent variables. *An*alysis *o*f *va*riance (ANOVA) is the appropriate technique for estimating the parameters of the linear model given in Eq. 1.1 when the independent variables are nominal or categorical.

As another example, consider the case where a medical researcher is interested in the following research issues: (1) Does gender affect cholesterol levels? (2) Does occupation affect cholesterol levels? and, (3) Do gender and occupation jointly affect cholesterol levels? In this example, the independent variables, gender and occupation, are nominal (i.e., categorical), and the dependent variable, cholesterol level, is metric. Again, ANOVA is the appropriate statistical method for this type of data set. Therefore, ANOVA is a special case of multiple regression in which there are multiple independent variables and one dependent variable. The difference is in the level of measurement used for the independent variables. If the number of nonmetric independent variables, as measured by dummy variables, is one, then ANOVA reduces to a simple t-test. For example, a t-test would be used to determine the effect of gender on subjects' cholesterol levels. That is, is the difference in the average cholesterol levels of males and females statistically significant?

Discriminant Analysis

Suppose that the dependent variable, PI, in the purchase-behavior example is measured using a nominal scale. That is, respondents are asked to indicate whether they will or will not purchase a given product. The independent variables A, I, E, and L, on the other hand, are measured using an interval or a ratio scale. We now have a data set in which the dependent variable is categorical or nominal and the independent variables are metric or continuous. The problem reduces to determining whether the two groups, potential

Table 1.2 **Independent Variables Measured Using Nominal Scale**

Independent Variable	Category
Income	High income
	Medium income
	Low income
Education	Less than high school
	High school graduate
	College graduate
	Graduate school or more
Age	Young
	Middle aged
	Senior citizen
Life style	Outgoing
	Homebody

purchasers and nonpurchasers of the product, are significantly different with respect to the independent variables. And if they are, then can the independent variables be used to develop a prediction equation or classification rule for classifying consumers into one of the two groups? Two-group discriminant analysis is a special technique developed for such a situation. The model for two-group discriminant analysis is the same as that given in Eq. 1.1. Therefore, as is discussed in later chapters, one can use multiple regression to achieve the objectives of two-group discriminant analysis. That is, two-group discriminant analysis is a special case of multiple regression.

As another example, consider a data set consisting of two groups of firms: high- and low-performance firms. An industry analyst is interested in identifying the financial ratios that provide the best discrimination between the two types of firms. Furthermore, the analyst is also interested in developing a procedure or rule to classify future firms into one of the two groups. Here again, two-group discriminant analysis would be the appropriate technique.

Logistic Regression

One of the assumptions in discriminant analysis is that the data come from a multivariate normal distribution. Furthermore, situations do arise where the independent variables are a combination of metric and nominal variables. The multivariate normality assumption would definitely not hold when the independent variables are combinations of metric and nominal variables. Violation of the multivariate normality assumption affects the statistical significance tests and the classification rates. Logistic regression analysis, which does not make any distributional assumption for the independent variables, is more robust to the violation of the multivariate normality assumption than discriminant analysis, and therefore is an alternative procedure to discriminant analysis. The model for logistic regression is not the same as that given by Eq. 1.1, and hence logistic regression analysis is not a special case of multiple regression analysis.

Discrete Discriminant Analysis

In the purchase-behavior example, if the dependent variable is measured as above (i.e., it is categorical) and the independent variables are measured as given in Table 1.2, then one would use discrete discriminant analysis. It should be noted that the estimation techniques and the classification procedures used in discrete discriminant analysis are not the same as those for discriminant analysis. Specifically, discrete discriminant analysis uses rules based on multinomial classification that are quite different from the classification rules used by discriminant analysis. For further discussion on discrete discriminant analysis see Goldstein and Dillon (1978).

Conjoint Analysis

For the purchase-behavior example, assume that the independent variables are measured as given in Table 1.2 and the dependent variable, PI, is measured using an ordinal scale. The problem is very similar to that of ANOVA except that the dependent variable is now ordinal. In such a case one resorts to an estimation technique known as *mon*otonic *an*alysis *o*f *va*riance (MONANOVA). MONANOVA belongs to a class of multivariate techniques called *conjoint analysis*. As illustrated in the following example, conjoint analysis is a very popular technique for designing new products or services.

Suppose a financial institution is interested in introducing a new type of checking account. Based on previous research, management has identified the attributes consumers

Table 1.3 **Attributes and Their Levels for Checking Account Example**

1. Service fee
   - No service fee
   - A flat fee of $5.00 per month
   - $2.00 per month plus $0.05 for each check written
2. Cancelled check return policy
   - Cancelled checks are returned
   - Cancelled checks are not returned
3. Account overdraft privilege
   - No overdraft allowed
   - A $5.00 charge for each overdraft
4. Phone transaction
   - A $0.50 charge per transaction
   - Free, unlimited phone transactions
5. Minimum balance
   - No minimum balance
   - Minimum balance of $500

use in selecting a checking account. Table 1.3 gives the attributes and their levels. The attributes can be variously combined to obtain a total of 48 different types of checking accounts. Management is interested in estimating the utilities that consumers attach to each level of the attributes. These utilities (also referred to as *part worths*) can then be used to design and offer the most desirable checking account.

In the above example, the independent variables are clearly nonmetric. ANOVA is the appropriate technique if the dependent variable used to measure consumers' preference for a given checking account, formed by a combination of the attributes, is metric. On the other hand, if the dependent variable is ordinal (i.e., nonmetric) then MONANOVA is one of the suggested techniques.

1.3.3 More Than One Dependent and One or More Independent Variables

Canonical Correlation

In the purchase-behavior example, assume that in addition to purchase intention we also have measured consumers' taste reactions (TR) to the product. The manager is interested in knowing how the two sets of variables—the I, E, A, and L and the PI and TR—are related. Canonical correlation analysis is the appropriate technique to analyze the relationship between the two sets of variables. Canonical correlation procedure does not differentiate between the two sets of variables. However, if based on some theory the manager determines that one set of variables (i.e., I, E, A, and L) is independent and the other set of variables (i.e., PI and TR) is dependent, then the manager can use canonical correlation analysis to determine how the set of independent variables jointly affects the set of dependent variables. Notice that canonical correlation reduces to multiple regression in the case of one dependent variable. That is, multiple regression itself is a special case of canonical correlation.

Multivariate Analysis of Variance

Suppose that in the purchase-behavior example the independent variables are nominal (as given in Table 1.2) and the two dependent variables are metric. *M*ultivariate *an*alysis *o*f *va*riance (MANOVA) is the appropriate multivariate method for this type of data set.

In another example, assume that one is interested in determining how a firm's financial health, measured by a number of financial ratios, is affected by such factors as size of the firm, industry characteristics, type of strategy employed by the firm, and characteristics of the CEO and the board of directors. One could use MANOVA to determine the effect of the independent variables on the dependent variables.

Multiple-Group Discriminant Analysis

Multiple-group discriminant analysis (MDA) is the appropriate method if the independent variables are metric and the dependent variables are nonmetric. In the purchase-behavior example, suppose that three groups of consumers are identified: the first group is willing to purchase the product and likes the taste of the product; the second group is not willing to purchase the product, but likes the product's taste; and the third group of consumers is unwilling to purchase the product and does not like the taste of the product. The problem reduces to determining how the three groups differ with respect to the independent variables and to identify a prediction equation or rule for classifying future customers into one of the three groups.

In another example, suppose that firms can be classified as: (1) high-performance firms; (2) medium-performance firms; and (3) low-performance firms. An industry analyst is interested in identifying the relevant financial ratios that provide the best discrimination among the three types of firms. The financial analyst is also interested in developing a procedure to classify future firms into one of the three types of firms. The analyst can use MDA for achieving these objectives. Notice that two-group discriminant analysis is a special case of MDA.

Discrete Multiple-Group Discriminant Analysis

In the above purchase-behavior example, suppose that the independent variables are categorical. In such a case one would use discrete multiple-group discriminant analysis. In a second example, suppose that the management of a telephone company is interested in determining the differences among households that own one, two, or more than two phones with respect to such categorical variables as gender, occupation, socioeconomic status, location, and type of home. In this example, both the independent and dependent variables are nonmetric. Discrete multiple-group discriminant analysis would be the appropriate multivariate method; however, once again it should be noted that the estimation techniques and classification procedures for discrete discriminant analysis are quite different from those of discriminant analysis.

1.4 INTERDEPENDENCE METHODS

As mentioned previously, situations do exist in which it is impossible or incorrect to delineate one set of variables as independent and another set as dependent. In these situations the major objective of data analysis is to understand or identify *why and how* the variables are correlated among themselves. Table 1.4 gives a list of interdependence multivariate methods. The multivariate methods for the case of two variables are the

Table 1.4 **Interdependence Statistical Methods**

	Type of Data	
Number of Variables	Metric	Nonmetric
Two	• Simple correlation	• Two-way contingency table • Loglinear models
More than two	• Principal components • Factor analysis	• Multiway contingency tables • Loglinear models • Correspondence analysis

same as the methods for more than two variables and, consequently, are not discussed separately.

1.4.1 Metric Variables

Principal Components Analysis

Suppose a financial analyst has a number of financial ratios (say 100) which he/she can use to determine the financial health of any given firm. For this purpose, the financial analyst can use all 100 ratios or use a few (say two) composite indices. Each composite index is formed by summing or taking a weighted average of the 100 ratios. Clearly, it is easier to compare the firms by using the two composite indices than by using 100 financial ratios. The analyst's problem reduces to identifying a procedure or rule to form the two composite indices. Principal components analysis is a suitable technique for such a purpose. It is sometimes classified as a data reduction technique because it attempts to reduce a *large* number of variables to a *few* composite indices.

Factor Analysis

Suppose an educational psychologist has available students' grades in a number of courses (e.g., math, chemistry, history, English, and French) and observes that the grades are correlated among themselves. The psychologist is interested in determining why the grades are correlated. That is, what are the few underlying reasons or factors that are responsible for the correlation among the course grades. Factor analysis can be used to identify the underlying factors. Once again, since factor analysis attempts to identify a *few* factors that are responsible for the correlation among a *large* number of variables, it is also classified as a data reduction technique. In this sense, factor analysis can be viewed as a technique that attempts to identify groups or clusters of variables such that correlations of the variables within each cluster are higher than correlations of variables across clusters.

Cluster Analysis

Cluster analysis is a technique for grouping observations into clusters or groups such that the observations in each cluster or group are similar with respect to the variables used to form clusters, and observations across groups are as different as possible with respect to the clustering variables. For example, nutritionists might be interested in grouping or clustering food items (i.e., fish, beef, chicken, vegetables, and milk) into groups such that the food items within each group are as homogeneous as possible but food items across the groups are different with respect to the food items' nutrient values. Note that in cluster analysis, observations are clustered with respect to certain characteristics of the observations, whereas in factor analysis variables are clustered or grouped with respect to the correlation between the variables.

1.4.2 Nonmetric Data

Loglinear Models

Consider the contingency or cross classification table presented in Table 1.5. The data in the table can be analyzed by a number of different methods, one of the most popular being to use crosstabulation or contingency table analysis to determine if there is a relationship between the two variables. Alternatively, one could use *Loglinear* models to estimate the probability of any given observation falling into one of the cells as a function of the independent variables—marital status and occupation. Loglinear models can also be used to examine the relationship among more than two categorical variables.

Correspondence Analysis

Suppose that we have a large contingency or crosstabulation table (say a 20×20 table). Interpretation of such a large table could be simplified if a few components representing most of the relationships between the row and column variables could be identified. Correspondence analysis attains this objective. In this respect, the purpose of *correspondence analysis* is similar to that of principal components analysis. In fact, correspondence analysis can be viewed as equivalent to principal components analysis for nonmetric data. Loglinear models and correspondence analysis can be generalized to multiway contingency tables. Multiway contingency tables are crosstabulations for more than two variables.

Table 1.5 **Contingency Table**

	Marital Status				
Occupation	Married	Widowed	Divorced	Separated	Never Married
Professional	30	20	20	25	5
Clerical	30	40	10	10	10
Blue collar	25	30	20	5	20

کتابخانه مرکزی دانشگاه تربیت مدرس

1.5 STRUCTURAL MODELS

In recent years a number of statistical methods have appeared for analyzing relationships among a number of variables represented by a system of linear equations. Some researchers have labeled these methods as second-generation multivariate methods. In the following section we provide a brief discussion of the second-generation multivariate methods.

Consider the causal model shown in Figure 1.1. The model depicts the relationship among dependent and independent variables and is usually referred to as a path or a structural model. The model can be represented by the following system of equations:

$$\begin{aligned} Y_1 &= a_1X_1 + e_1 \\ Y_2 &= a_2X_2 + e_2 \\ Y_3 &= b_1Y_1 + b_2Y_2 + e_3, \end{aligned} \tag{1.2}$$

where a_1 and b_1 are the path, or structural, or regression coefficients and the e_1 are the errors in equations. A number of statistical packages (e.g., SAS) have routines or procedures (e.g., SYSLIN) to estimate the parameters of the system of equations given by Eq. 1.2.

Now suppose that the dependent variables (i.e., Y) and the independent variables (i.e., X) cannot be directly observed. For example, constructs such as attitudes, personality, and intelligence cannot be directly observed. Such constructs are referred to as *latent* or *unobservable* constructs. However, one can obtain multiple observable measures of these latent constructs. Figure 1.2 represents the modified path model given in Figure 1.1. In the figure, x and y are, respectively, the observable measures of the independent and the dependent variables. The part of the model in Figure 1.2 which depicts the relationship among the unobservable constructs and its indicators is referred to as

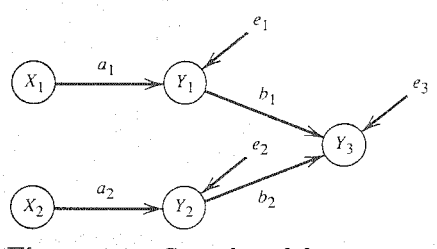


Figure 1.1 Causal model.

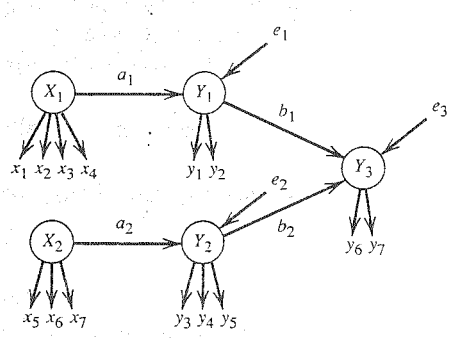


Figure 1.2 Causal model for unobservable constructs.

the *measurement model*, and the part of the model that represents relationships among the latent constructs is called the *structural model*. Estimation of model parameters can be broken down into the following two parts:

1. Estimate the unobservable constructs using the observable measures. That is, first estimate the parameters of the measurement model. Techniques like factor or confirmatory factor analysis can be used for this purpose.
2. Use the estimates of the unobservable constructs, commonly referred to as *factor scores*, to estimate the coefficients of the structural model.

Recently, estimation procedures have been developed to *simultaneously* estimate the parameters of the structural and measurement models given in Figure 1.2. These estimation procedures are available in the following computer packages: (1) the LISREL procedure in **SPSS**; (2) the CALIS procedure in **SAS**; and (3) the EQS procedure in **BIOMED**.

1.6 OVERVIEW OF THE BOOK

Obviously, it is not possible to cover all the techniques presented in Tables 1.1 and 1.4. This book covers the following multivariate techniques:

1. Principal components analysis.
2. Factor analysis.
3. Confirmatory factor analysis.
4. Cluster analysis.
5. Two-group discriminant analysis.
6. Multiple-group discriminant analysis.
7. Logistic regression.
8. MANOVA.
9. Canonical correlation.
10. Structural models.

Apart from regression and ANOVA, these are the most widely used multivariate techniques. Multiple regression and ANOVA are not covered because these two techniques are normally covered in a single course and require a separate textbook to provide good coverage. The four interdependence techniques are covered first, followed by the remaining six dependence techniques. Also included is a chapter discussing the assumptions made in MANOVA and discriminant analysis.

The following discussion format is used for presenting the material in the book:

1. Wherever appropriate, the concepts of the techniques are discussed using hypothetical data and geometry. Geometry is used very liberally because it lends itself to a very lucid presentation of most of the statistical techniques.
2. Geometrical discussion is followed by an analytical discussion of the technique. The analytical discussion is nonmathematical and does not use any matrix or linear algebra.

3. Next, a detailed discussion of how to interpret the resulting output from such statistical packages as SPSS and SAS is provided.[2] A discussion of the various issues faced by the applied researcher is also included. Only the relevant portion of the computer output is included. The interested reader can easily obtain the full output as almost all the data sets are provided either in tables or on the floppy diskette.
4. Most chapters have appendices which contain the technical details of the multivariate techniques. The appendices require a fairly good knowledge of matrix algebra; however, the applied researcher can safely omit the material in the appendices.

The next two chapters provide an overview of the basic geometrical and analytical concepts employed in the discussion of the statistical techniques. The remaining chapters discuss the foregoing statistical techniques.

QUESTIONS

1.1 For each of the measurement situations described below, indicate what *type* of scale is being used.

(a) An owner of a Ford Escort is asked to indicate her satisfaction with her car's handling ease using the following scale:

−2	−1	0	1	2
very dissatisfied				very satisfied

(b) In a consumer survey, a housewife is asked to indicate her annual household income using the following classification:

(i)	\$0–\$25,000	Level A
(ii)	\$25,001–\$45,000	Level B
(iii)	\$45,001–\$65,000	Level C
(iv)	\$65,001–\$80,000	Level D
(v)	More than \$80,000	Level E

(c) The housewife in (b) is asked to indicate her annual household income in dollars.

(d) A prospective car buyer is asked to rank the following criteria, used in deciding which car to buy, in order of their importance: (i) Manufacturer of the car; (ii) Terms of payment; (iii) Price of the car; (iv) Safety measures such as air bags, antilock brakes; (v) Size of the car; (vi) Automatic v. stick shift; and (vii) Number of miles to a gallon of gas.

(e) In a weight-reduction program the weight (in pounds) of each participant is measured every day.

1.2 For each variable listed, certain measurement scales are indicated. In each case suggest suitable operational measures of the indicated scale type(s). (a) Classroom temperature: ratio scaled, interval scaled; (b) Age: nominal, ratio scaled; (c) Importance of various criteria used to select a store for grocery shopping: ordinal, interval scaled; (d) Opinion on the importance of sex education in high school: interval scaled; and (e) Marital status: nominal.

1.3 Construct dummy variables to represent the nominal variable "race." The possible races are: (a) Caucasian; (b) Asian; (c) African–American; and (d) Latin–American.

1.4 A marketing research company believes that the sales (S) of a product are a function of the number of retail outlets (N_R) in which it is available, the advertising dollars (A) spent on the product, and the number of years (N_Y) the product has already been available on the

[2]These packages are chosen as they are the most widely used commercial packages.

market. The company has information on S, N_R, A, and N_Y for 35 competing brands at a given point in time. Suggest a suitable statistical method that will help the company test the relationship between sales and N_R, A, and N_Y.

1.5 In a nationwide survey of its customers, a leading marketer of consumer packaged goods collected information about various buying habits. The company wants to identify distinct segments among the consumers and design marketing strategies tailored to individual segments. Suggest a suitable statistical method to the marketing research department of the company to help it accomplish this task.

1.6 An experiment is conducted to determine the impact of background music on sales in a department store. During the first week no background music is played and the total store sales are measured. During the second week fast-tempo background music is played and total store sales are measured. During the third and final week of the experiment slow-tempo background music is played and total store sales are measured.

Suggest a suitable statistical method to determine if there are significant differences between the store sales under no-music, fast-tempo music, and slow-tempo music conditions.

1.7 ABC Tour & Travel Company advertises its tour packages by mailing brochures about tourist resorts. The company feels it could increase its marketing efficiency if it were able to segregate consumers likely to go on its tours from those not likely to go, based on consumer demographics and lifestyle considerations. You decide to help the company by undertaking some consumer research.

From the company's files you extract the names and addresses of consumers who had received the brochures in the past two years. You select two random samples of consumers who went on the tours and those who didn't. Having done this, you interview the selected consumers and collect demographic and lifestyle information (using nonmetric scales) about them.

Describe a statistical method that you would use to help predict the tour-going potential of consumers based on their demographics and lifestyles.

1.8 How do structural models (e.g., covariance structure analysis) differ from ordinary multivariate methods (e.g., multivariate regression analysis)?

CHAPTER 2

Geometric Concepts of Data Manipulation

A picture is worth a thousand words. A clear and intuitive understanding of most of the multivariate statistical techniques can be obtained by using geometry. In this chapter the necessary background material needed for understanding the geometry of the multivariate statistical techniques discussed in this book is provided. For presentation clarity, the discussion is limited to two dimensions; however, the geometrical concepts discussed can be generalized to more than two dimensions.

2.1 CARTESIAN COORDINATE SYSTEM

Figure 2.1 presents four points, A, B, C, and D, in a two-dimensional space. It is obvious that the location of each of these points in the two-dimensional space can only be specified relative to each other, or relative to some reference point and reference axes. Let O be the reference point. Furthermore, let us draw two perpendicular lines, X_1 and X_2, through point O.

The points in the space can now be represented based on how far they are from O. For example, point A can be represented as $(2, 3)$, indicating that this point is reached by moving 2 units to the right of O along X_1 and then 3 units above O and parallel to X_2. Alternatively, point A can be reached by moving 3 units above O along X_2 and then 2 units to the right of O and parallel to X_1. Similarly, point B can be represented as $(-4, 2)$, meaning that this point is reached by moving 4 units to the left of O along X_1 and 2 units above O and parallel to X_2. Note that movement to the right of or above O is assigned a positive sign, and movement to the left of or below O is assigned a negative sign. This system of representing points in a space is known as the *Cartesian coordinate system.* Point O is called the *origin,* and the X_1 and X_2 lines are known as *rectangular Cartesian axes* and will simply be referred to as *axes.* The values 2 and -4 are known as X_1 *coordinates* of the points A and B, respectively, and the values 3 and 2 as the X_2 *coordinates* of points A and B, respectively.

In general, a p-dimensional space is represented by p axes passing through the origin with the axes perpendicular to each other. Any point, say A, in p dimensions is represented as $(a_1, a_2, \ldots, a_p)$, where a_p is the coordinate of the point for the pth axis. This representation implies that the point A can be reached by moving a_1 units along the first axis (i.e., X_1), then moving a_2 units parallel to the second axis (i.e., X_2), and so on. Henceforth, this convention will be used to represent points in a given dimensional space.

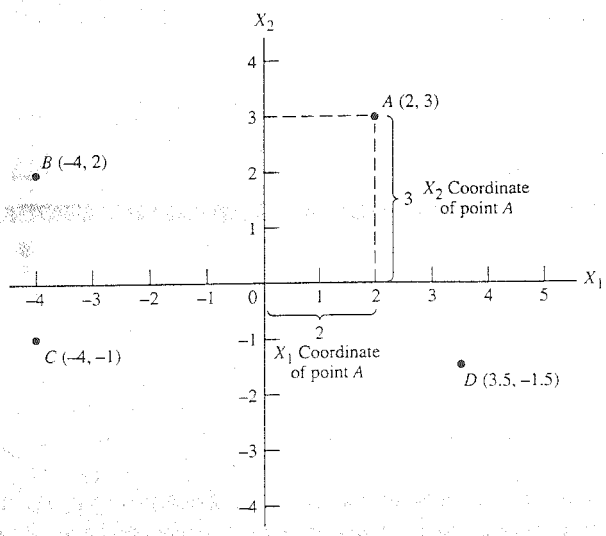


Figure 2.1 Points represented relative to a reference point.

2.1.1 Change in Origin and Axes

Suppose that the origin O and, therefore, the axes, X_1 and X_2, are moved to another location in the space. The representation of the *same* points with respect to the new origin and axes will be *different*. However, the position of the points in the space with respect to each other (i.e., the orientation of the points) does not change. Figure 2.2 gives the representation of the same points (i.e., A and B) with respect to the new origin (O^*) and the associated set of new axes (X_1^* and X_2^*).[1] Notice that points A and B can be represented as $(2, 3)$ and $(-4, 2)$, respectively, with respect to the origin O, and as

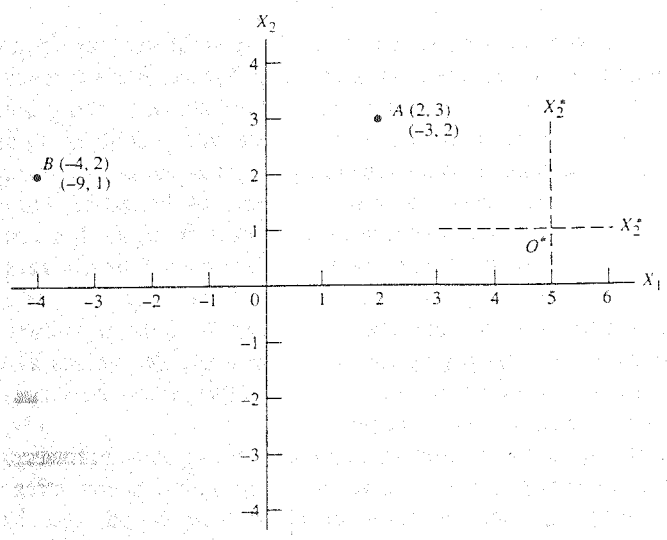


Figure 2.2 Change in origin and axes.

[1]Henceforth, the term *origin* will be used to refer to both the origin and the associated set of reference axes defining the Cartesian coordinate system.

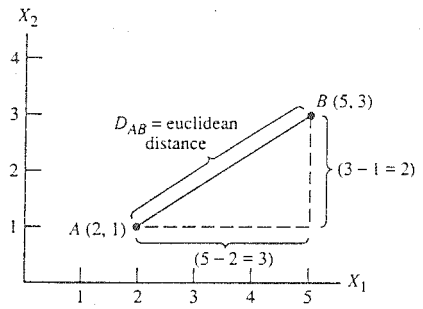


Figure 2.3 Euclidean distance between two points.

$(-3, 2)$ and $(-9, 1)$, respectively, with respect to the new origin, O^*. The new origin O^* can itself be represented as $(5, 1)$ with respect to the old origin, O. Algebraically, any point represented with respect to O can be represented with respect to the new origin O^* by subtracting the coordinates of the origin O^* with respect to O from the respective coordinates of the point. For example, point A can be represented with respect to the new origin O^* as $(2 - 5, 3 - 1)$ or $(-3, 2)$.

2.1.2 Euclidean Distance

One of the measures of how far apart two points are is the straight-line distance between the two points. The straight-line distance between any two points is referred to as the *euclidean distance* between the two points. The Pythagorean theorem can be used to compute the euclidean distance between the two points. In Figure 2.3, according to the Pythagorean theorem, the euclidean distance, D_{AB}, between points A and B is equal to

$$\begin{aligned} D_{AB} &= \sqrt{(5-2)^2 + (3-1)^2} \\ &= \sqrt{13}, \end{aligned}$$

or the squared euclidean distance, D^2_{AB}, is equal to

$$D^2_{AB} = 13.$$

In general the euclidean distance between any two points in a p-dimensional space is given by

$$D_{AB} = \sqrt{\sum_{j=1}^{p} (a_j - b_j)^2}, \tag{2.1}$$

where a_j and b_j are coordinates of points A and B for the jth axis representing the jth dimension.

2.2 VECTORS

Vectors in a space are normally represented as directed line segments or arrows. The vector or the arrow begins at an *initial* point and ends at a *terminal* point. Or in other words, a vector is a line joining two points (i.e., an initial and a terminal point). Notationally, vectors are represented as lowercase bold letters, and points as uppercase italic

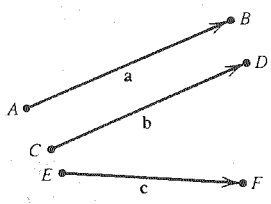


Figure 2.4 Vectors.

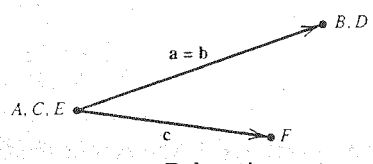


Figure 2.5 Relocation or translation of vectors.

letters. For example, in Figure 2.4 **a** is a vector joining points A and B. The length of the vector is simply the euclidean distance between the two points, and is referred to as the *norm* or *magnitude* of the vector. Sometimes, the points A and B are, respectively, referred to as the tail and head of the vector. Clearly, a vector has a length and a direction. Vectors having the same length and direction are referred to as *equivalent vectors*. In Figure 2.4, vectors **a** and **b** are equivalent as they have the same length and direction; vector **c** is not equivalent to vectors **a** and **b** as vector **c** has a different direction than **a** and **b**. That is, vectors having the *same length and direction* are considered to be equivalent even if they are *located at different positions in the space*. In other words, vectors are completely defined with respect to their magnitude and direction. Consequently, vectors can be moved or translated in the space such that they have the same tail or initial point. The vector does not change if its magnitude and direction are not affected by the move or translation. Figure 2.5 gives the new location of vectors **a**, **b**, and **c** such that they have the same initial point. Note that vectors **a** and **b** overlap, indicating that they are equivalent.

2.2.1 Geometric View of the Arithmetic Operations on Vectors

Vectors can be subjected to a number of operations such as (1) multiplying or dividing a vector by a real number; (2) addition and/or subtraction of two or more vectors; (3) multiplication of two vectors; and (4) projecting one vector onto another vector. A geometrical view of these operations is provided in the following sections. In order to differentiate between points, vectors, and real numbers the following representation is used: (1) points are represented by uppercase italic letters; (2) vectors are represented by lowercase bold letters; and (3) real numbers are represented by lowercase italic letters.

Multiplication of a Vector by a Real Number

A vector **a** multiplied by a real number k results in a new vector **b** whose length is $|k|$ times the length or magnitude of vector **a**, where $|k|$ is the absolute value of k. The real number, k, is commonly referred to as a *scalar*. For positive-valued scalars the new vector **b** has the same direction as that of vector **a**, and for negative-valued scalars the new vector **b** has an opposite direction as that of vector **a**. In Figure 2.6, for example,

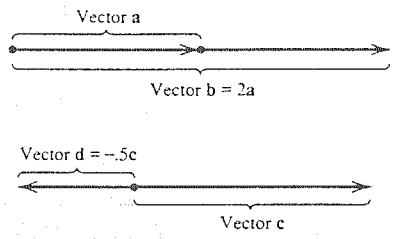


Figure 2.6 Scalar multiplication of a vector.

vector **a** multiplied by 2 results in a new vector **b** whose length is twice that of vector **a**, and whose direction is the same as that of vector **a**. On the other hand, vector **c**, when multiplied by $-.5$, results in a new vector **d** whose length is half that of vector **c** and whose direction is opposite to that of vector **c**. Notice that multiplication of a vector by $-.5$ is the same as dividing the vector by -2. To summarize, multiplying a vector by any scalar k

- *Stretches* the vector if $|k| > 1$ and *compresses* the vector if $|k| < 1$. The amount of stretching or compression depends on the absolute value of the scalar. If the value of the scalar is zero, the new vector has zero length. A vector of zero length is referred to as a *null* or *zero* vector. The null vector has no direction and, therefore, any direction that is convenient for the given problem may be assigned to it.
- The direction of the vector is preserved for positive scalars and for negative scalars the direction is reversed. The reversal of vector direction is called *reflection*.

That is, vectors can be reflected and/or stretched or compressed by multiplying them with a scalar.

Addition and Subtraction of Vectors

ADDITION OF VECTORS. The sum or addition of two vectors results in a third vector. That is, $\mathbf{c} = \mathbf{a} + \mathbf{b}$, and is obtained as follows:

- Reposition **b** such that its initial point coincides with the terminal point of **a**.
- The sum, $\mathbf{a} + \mathbf{b}$, is given by **c** whose initial point is the same as the initial point of **a** and the terminal point is the same as the terminal point of the repositioned vector **b**.

Figure 2.7 shows the concept of vector addition. The new position of **b**, such that its initial point is the same as the terminal point of **a**, is given by the dotted vector. Figure 2.7 also shows the addition $\mathbf{b} + \mathbf{a}$. Once again, the dotted vector shows the new position of **a** such that its initial point is the same as the terminal point of **b**. Notice that $\mathbf{a}+\mathbf{b} = \mathbf{b}+\mathbf{a}$.

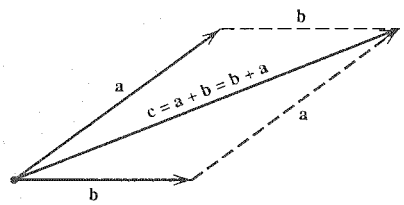


Figure 2.7 Vector addition.

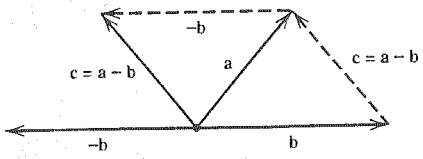


Figure 2.8 Vector subtraction.

That is, vector addition is commutative. Also, notice that $\mathbf{a} + \mathbf{b}$ is given by the diagonal of the parallelogram formed by $\mathbf{a}$ and $\mathbf{b}$, and this is sometimes referred to as the *parallelogram law of vector addition.*

SUBTRACTION OF VECTORS. Subtraction of two vectors is a special case of vector addition. For example, $\mathbf{c} = \mathbf{a} - \mathbf{b}$ can be obtained by first multiplying $\mathbf{b}$ by -1 to yield $-\mathbf{b}$, and then adding $\mathbf{a}$ and $-\mathbf{b}$. Figure 2.8 shows the process of vector subtraction. Notice that $\mathbf{c} = \mathbf{a} - \mathbf{b}$ can be moved so that its initial point coincides with the terminal point of $\mathbf{b}$, and its terminal point coincides with the terminal point of $\mathbf{a}$. That is, $\mathbf{c} = \mathbf{a} - \mathbf{b}$ is also given by the vector whose initial point is at the terminal of $\mathbf{b}$, and the terminal point is at the terminal point of $\mathbf{a}$.

Addition and subtraction of more than two vectors is a straightforward extension of the above procedure. For example, the sum of three vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ is obtained by first adding any two vectors, say $\mathbf{a}$ and $\mathbf{b}$, to give another vector which can then be added to the third vector, $\mathbf{c}$. It will become clear in the later chapters that addition and subtraction of vectors is analytically equivalent to forming linear combinations or weighted sums of variables to obtain new variables, which is the basis of most of the multivariate statistical techniques.

Multiplication of Two Vectors

The product of two vectors is defined such that it results in a single number or a scalar, and therefore multiplication of two vectors is referred to as the *scalar product* or *inner*

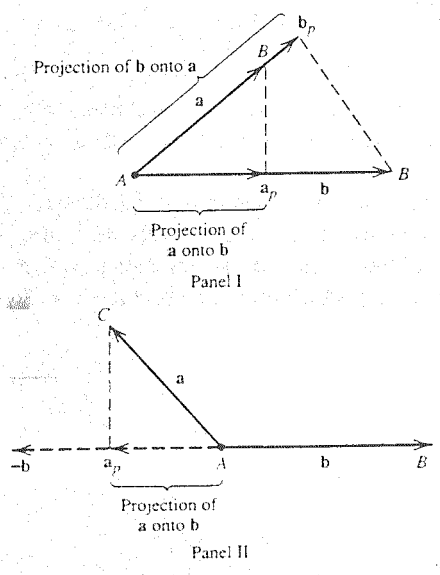


Figure 2.9 Vector projections.

dot product of two vectors. The scalar product of two perpendicular vectors is zero. Multiplication of two vectors is discussed further in Section 2.4.4.

2.2.2 Projection of One Vector onto Another Vector

Any given vector can be projected onto other vectors.[2] Panel I of Figure 2.9 shows the projection of **a** onto **b** and the resulting projection vector $\mathbf{a}_p$. The projection of **a** onto **b** is obtained by dropping a perpendicular from the terminal point of **a** onto **b**. The projection of **a** onto **b** results in another vector called the *projection vector*, and is normally denoted as $\mathbf{a}_p$. The initial point of $\mathbf{a}_p$ is the same as the initial point of **b** and the terminal point lies somewhere on **b** or $-\mathbf{b}$. The length or magnitude of $\mathbf{a}_p$ is called the *component* of **a** along **b**. As shown in Panel I of Figure 2.9, $\mathbf{b}_p$ is the projection vector obtained by projecting **b** onto **a**. Panel II of Figure 2.9 shows the projection vector $\mathbf{a}_p$ whose direction is in the direction of $-\mathbf{b}$.

2.3 VECTORS IN A CARTESIAN COORDINATE SYSTEM

Consider the coordinate system given in Figure 2.10, which has the origin at point $O = (0, 0)$, and X_1 and X_2 are the two reference axes. Let $A = (a_1, a_2)$ be a point whose X_1 and X_2 coordinates, respectively, are a_1 and a_2. Point A can be represented by vector **a** whose terminus is point A, and the initial point is the origin O. Typically, vector **a** is represented as $\mathbf{a} = (a_1\ a_2)$, where a_1 and a_2 are called the components of the vector. Vector **a** is also referred to as a 2-tuple vector where the number of tuples is equal to the number of components or elements of the vector. Note that the components of the vector are the same as the coordinates of the point A. That is, point A in a Cartesian coordinate system can be represented by a vector whose terminus is at the respective point and the tail is at the origin. Indeed, all points in a coordinate system can be represented as vectors such that respective points are the terminuses and the origin is the initial point *for all the vectors*. In general, any point in a p-dimensional space can be represented as a *p-component* vector in the p-dimensional space. That is, point A in a p-dimensional space can be represented as a p-tuple vector $\mathbf{a} = (a_1\ a_2 \cdots a_p)$. The origin O in a p-dimensional space is represented by the null vector $\mathbf{0} = (0\,0 \cdots 0)$. Thus, any vector in a p-dimensional Cartesian coordinate system can be located by its p components (i.e., coordinates).

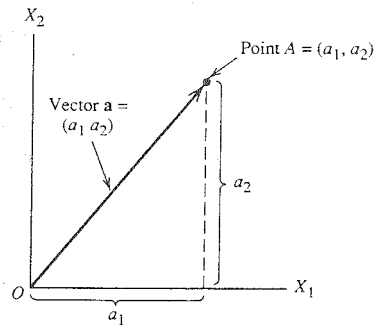


Figure 2.10 Vectors in a Cartesian coordinate system.

[2]A vector can also be projected onto spaces. Projection of vectors onto a space is discussed later.

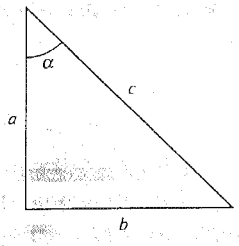


Figure 2.11 Trigonometric functions.

2.3.1 Length and Direction Cosines

We first provide a very brief discussion of the relevant trigonometric functions used in this chapter. Figure 2.11 gives a right-angle triangle. The cosine of angle α is given by the adjacent side a divided by the hypotenuse c. The sine of the angle is given by the opposite side b divided by the hypotenuse c. That is,

$$\cos\alpha = \frac{a}{c} \qquad \sin\alpha = \frac{b}{c}.$$

Also, $(\cos\alpha)^2 + (\sin\alpha)^2 = 1$ or $\cos^2\alpha + \sin^2\alpha = 1$.

The location of each vector in the Cartesian system or space can also be determined by the length of the vector and the angle it makes with the axes. The length of any vector is given by the euclidean distance between the terminal point of the vector and the initial point (i.e., the origin). For example, in Figure 2.12 the length of vector **a** is given by

$$D_{AO} = \|\mathbf{a}\| = \sqrt{(a_1 - 0)^2 + (a_2 - 0)^2} = \sqrt{a_1^2 + a_2^2}, \tag{2.2}$$

where $\|\mathbf{a}\|$ represents the length of vector **a**. In general, the length of a vector in p dimensions will be given by

$$\|\mathbf{a}\| = \sqrt{\sum_{j=1}^{p} a_j^2} \tag{2.3}$$

where a_j is the jth component (i.e., the jth coordinate).

As depicted in Figure 2.12, vector **a** makes angles of α and β, respectively, with X_1 and X_2 axes. From basic trigonometry, the cosine of the angle is given by

$$\cos\alpha = \frac{a_1}{\|\mathbf{a}\|} = \frac{a_1}{\sqrt{a_1^2 + a_2^2}} \tag{2.4}$$

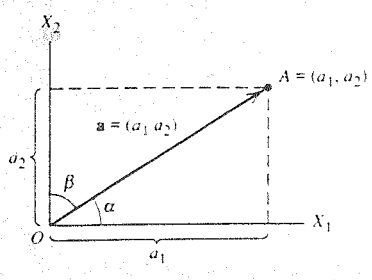


Figure 2.12 Length and direction cosines.

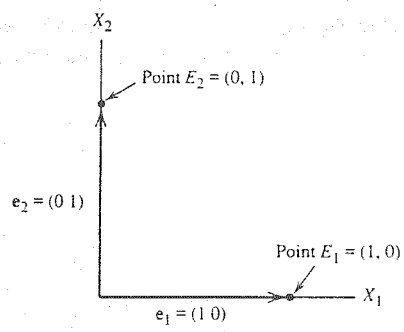


Figure 2.13 Standard basis vectors.

and

$$\cos\beta = \frac{a_2}{\|\mathbf{a}\|} = \frac{a_2}{\sqrt{a_1^2 + a_2^2}}. \tag{2.5}$$

The cosines of the angles between a vector and the axes are called *direction cosines*. The following two observations can be made.

1. If vector **a** is of unit length, then the direction cosine gives the component of **a** along the respective axes.
2. The sum of the square of direction cosines is equal to one. That is,

$$\cos^2\alpha + \cos^2\beta = \frac{a_1^2}{a_1^2 + a_2^2} + \frac{a_2^2}{a_1^2 + a_2^2} = 1,$$

and this relationship holds for any dimensional space.

2.3.2 Standard Basis Vectors

In Figure 2.13, let $E_1 = (1, 0)$ and $E_2 = (0, 1)$, respectively, be points on the X_1 and X_2 axes. These points can be represented as vectors $\mathbf{e}_1 = (1\ 0)$ and $\mathbf{e}_2 = (0\ 1)$, respectively. That is, the Cartesian axes can themselves be represented as vectors in a given dimensional space. In general, a p-dimensional space is represented by the p vectors, $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p$. These vectors are sometimes referred to as the *standard basis* vectors.

Note that $\|\mathbf{e}_1\| = 1$ and $\|\mathbf{e}_2\| = 1$ and the angle between the two vectors is equal to 90°. Vectors which are of unit length and orthogonal to each other are called *orthonormal* vectors. Thus the Cartesian axes can be represented by a set of orthonormal basis vectors. Henceforth the term *basis vectors* will be used to imply a set of orthonormal standard basis vectors that represent the respective axes of the Cartesian coordinate system.

2.4 ALGEBRAIC FORMULAE FOR VECTOR OPERATIONS

2.4.1 Arithmetic Operations

Section 2.2.1 provided a geometrical view of the various arithmetic operations on vectors. Representation of vectors in a Cartesian coordinate system facilitates the use of

algebraic equations to represent the various arithmetic operations on the vectors discussed in Section 2.2.1. This section gives these equations. Consider the two vectors $\mathbf{a} = (a_1\, a_2 \cdots a_p)$ and $\mathbf{b} = (b_1\, b_2 \cdots b_p)$. The various arithmetic operations are given by the following equations:

- Scalar Multiplication of a Vector

$$k\mathbf{a} = (ka_1\, ka_2 \cdots ka_p). \tag{2.6}$$

- Vector Addition and Subtraction

$$\mathbf{a} + \mathbf{b} = (a_1 + b_1\; a_2 + b_2 \cdots a_p + b_p). \tag{2.7}$$

$$\mathbf{a} - \mathbf{b} = (a_1 - b_1\; a_2 - b_2 \cdots a_p - b_p). \tag{2.8}$$

- Scalar Product of Two Vectors

$$\mathbf{ab} = a_1b_1 + a_2b_2 + \cdots + a_pb_p. \tag{2.9}$$

2.4.2 Linear Combination

Each point or vector in a space can be represented as a linear combination of the basis vectors. As depicted in Figure 2.14, $\mathbf{a}_1$ and $\mathbf{a}_2$ are two vectors that result by multiplying $\mathbf{e}_1$ and $\mathbf{e}_2$, respectively, by scalars a_1 and a_2. That is,

$$\mathbf{a}_1 = a_1\mathbf{e}_1 = (a_1\, 0)$$
$$\mathbf{a}_2 = a_2\mathbf{e}_2 = (0\, a_2).$$

The sum of the above two vectors results in a new vector

$$\mathbf{a} = \mathbf{a}_1 + \mathbf{a}_2 = a_1\mathbf{e}_1 + a_2\mathbf{e}_2$$
$$= (a_1\, 0) + (0\, a_2) = (a_1\, a_2)$$

whose terminus is A. Note that the vector $\mathbf{a}$ is given by the weighted sum of the two basis vectors. The weights a_1 and a_2 are called the coordinates of point A with respect to basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$, respectively. The weights are also the components of the vector $\mathbf{a}$ representing point A. This weighted sum is referred to as a *linear combination*. That is, vector $\mathbf{a}$ is a linear combination of the basis vectors.

It is interesting to note that $\mathbf{a}_1$ and $\mathbf{a}_2$ are the respective projection vectors resulting from the projection of vector $\mathbf{a}$ onto the basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$. The lengths of the projection vectors, which are also the components of $\mathbf{a}$ along $\mathbf{e}_1$ and $\mathbf{e}_2$, are a_1 and a_2

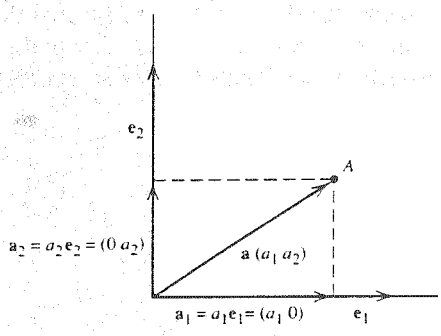


Figure 2.14 Linear combinations.

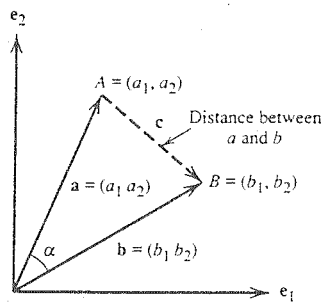


Figure 2.15 Distance and angle between any two vectors.

respectively. In general, any vector in a p-dimensional space can be represented as a linear combination of the basis vectors. That is, $\mathbf{a} = (a_1\ a_2 \cdots a_p)$ can be represented as

$$\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 + \cdots + a_p\mathbf{e}_p. \tag{2.10}$$

2.4.3 Distance and Angle between Any Two Vectors

The distance between any two vectors is given by the euclidean distance between the two vectors. From basic trigonometry, we know that the length of any side, c, of a triangle is given by

$$c = \sqrt{a^2 + b^2 - 2ab\cos\alpha} \tag{2.11}$$

where a and b are the lengths of the other two sides and α is the angle between the two sides. From Eq. 2.11, the distance between vectors $\mathbf{a}$ and $\mathbf{b}$ in Figure 2.15 is given by

$$\|\mathbf{c}\| = \sqrt{\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2 - 2 \cdot \|\mathbf{a}\| \bullet \|\mathbf{b}\| \cos\alpha}, \tag{2.12}$$

where $\cos\alpha$ is the angle between the two vectors.

2.4.4 Scalar Product and Vector Projections

The scalar product of two vectors is defined as

$$\mathbf{ab} = \|\mathbf{a}\| \bullet \|\mathbf{b}\| \cos\alpha, \tag{2.13}$$

where $\mathbf{ab}$ is the scalar product of $\mathbf{a}$ and $\mathbf{b}$. Representation of the scalar product by the above equation facilitates the discussion of the linkage between scalar products and vector projections.

Geometrically, the scalar product of two vectors is related to the concept of vector projections and the length of projection vectors. Panel I of Figure 2.16 shows the projection of $\mathbf{a}$ onto $\mathbf{b}$ and the resulting projection vector $\mathbf{a}_p$. The length of the projection vector, $\mathbf{a}_p$, is given by

$$\|\mathbf{a}_p\| = \|\mathbf{a}\| \cos\alpha, \tag{2.14}$$

where α is the angle between the vectors $\mathbf{a}$ and $\mathbf{b}$. Substituting the value of $\cos\alpha$ from Eq. 2.13, Eq. 2.14 can be rewritten as

$$\|\mathbf{a}_p\| = \|\mathbf{a}\| \bullet \frac{\mathbf{ab}}{\|\mathbf{a}\| \bullet \|\mathbf{b}\|} = \frac{\mathbf{ab}}{\|\mathbf{b}\|}. \tag{2.15}$$

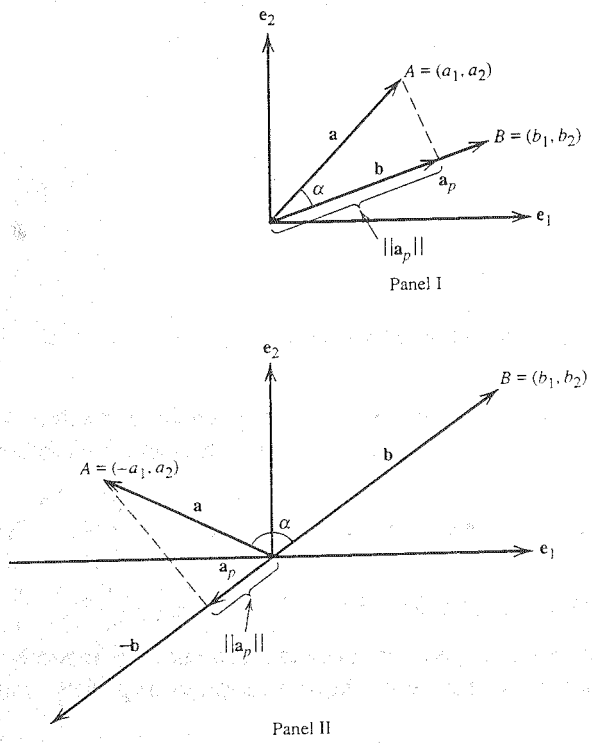


Figure 2.16 Geometry of vector projections and scalar products.

Or

$$\|\mathbf{a}_p\| = \mathbf{ab} \tag{2.16}$$

if $\|\mathbf{b}\| = 1$. The length, $\|\mathbf{a}_p\|$, of the projection vector, $\mathbf{a}_p$, is known as the component of $\mathbf{a}$ along $\mathbf{b}$. From Eq. 2.16, it is clear that the scalar product is the *signed length* of the projection vector. Since lengths are always positive, the sign attached to the length does not imply a positive or negative length; rather, it denotes the direction of the projection vector. If the angle between the two vectors is acute then the scalar product or the signed length of the projection vector will be positive, implying that the projection vector is in the direction of $\mathbf{b}$. On the other hand, as depicted in Panel II of Figure 2.16, if the angle between the two vectors is obtuse then the scalar product or the signed length will be negative, implying that the direction of the projection vector is in the direction of $-\mathbf{b}$.

Also note that for orthogonal vectors, projection of one vector onto another vector will result in a projection vector of zero length. That is, as is obvious from Eq. 2.13, the scalar product of orthogonal vectors is zero (as $\cos 90° = 0$). It can be shown that the projection vector is given by

$$\mathbf{a}_p = \frac{\|\mathbf{a}_p\| \cdot \mathbf{b}}{\|\mathbf{b}\|}, \tag{2.17}$$

which is equal to $\|\mathbf{a}_p\| \cdot \mathbf{b}$ if $\|\mathbf{b}\| = 1$.

2.4.5 Projection of a Vector onto Subspace

In Section 2.2.2 we discussed projection of one vector onto another vector. This concept can be extended to projection of a vector onto a p-dimensional subspace. For presenta-

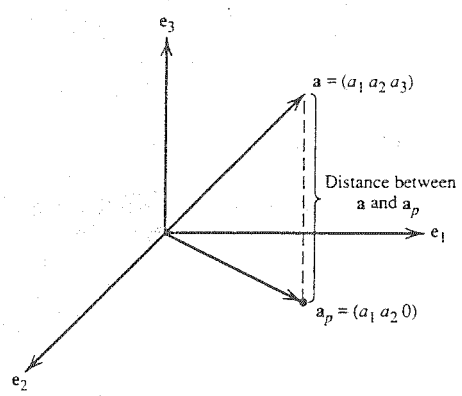


Figure 2.17 Projection of a vector onto a subspace.

tion clarity, let us consider the case where a vector in three dimensions is projected onto a two-dimensional subspace (i.e., a plane). Figure 2.17 gives vector $\mathbf{a} = (a_1\ a_2\ a_3)$ in a three-dimensional space defined by the basis vectors $\mathbf{e}_1$, $\mathbf{e}_2$, and $\mathbf{e}_3$. The projection vector, $\mathbf{a}_p = (a_1\ a_2\ 0)$, is obtained by dropping a perpendicular from the terminus of $\mathbf{a}$ onto the plane defined by $\mathbf{e}_1$ and $\mathbf{e}_2$. The distance between $\mathbf{a}$ and $\mathbf{a}_p$ is the shortest distance between $\mathbf{a}$ and any other vector in the two-dimensional space defined by $\mathbf{e}_1$ and $\mathbf{e}_2$. This concept can be extended to projecting any vector onto a p-dimensional subspace.

2.4.6 Illustrative Example

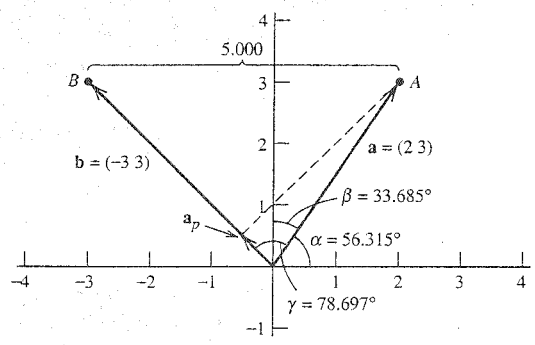
Consider the vectors $\mathbf{a} = (2\ 3)$ and $\mathbf{b} = (-3\ 3)$ shown in Figure 2.18.

1. The lengths of vectors $\mathbf{a}$ and $\mathbf{b}$ are (see Eq. 2.3):

$$\|\mathbf{a}\| = \sqrt{2^2 + 3^2} = 3.606$$
$$\|\mathbf{b}\| = \sqrt{-3^2 + 3^2} = 4.243.$$

2. The angles between vector $\mathbf{a}$ and the basis vectors is given by (see Eqs. 2.4 and 2.5):

$$\cos\alpha = \frac{2}{3.606} \qquad \alpha = 56.315°$$



$$\cos\beta = \frac{3}{3.606} \qquad \beta = 33.685°.$$

3. The scalar product of the two vectors is (see Eq. 2.9)

$$\mathbf{ab} = 2 \times (-3) + 3 \times 3 = 3.$$

4. The cosine of the angle and the angle between the two vectors is given by (see Eq. 2.13):

$$\cos\gamma = \frac{\mathbf{ab}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|} = \frac{3}{3.606 \times 4.243} = .196$$

or $\gamma = 78.697°$.

5. The distance between two vectors is given by (see Eq. 2.12):

$$\sqrt{3.606^2 + 4.243^2 - 2 \times 3.606 \times 4.243 \times .196} = 5.000,$$

or from Eq. 2.1

$$\sqrt{(2-(-3))^2 + (3-3)^2} = 5.000.$$

6. The length of the projection of vector **a** on **b** is given by (see Eq. 2.15):

$$\|\mathbf{a}_p\| = \frac{\mathbf{ab}}{\|\mathbf{b}\|} = \frac{3}{4.243} = .707.$$

7. The projection vector $\mathbf{a}_p$ is given by (see Eq. 2.17):

$$\mathbf{a}_p = \frac{.707}{4.243}(-3\ 3) = (-.500\ .500).$$

2.5 VECTOR INDEPENDENCE AND DIMENSIONALITY

It was seen that the Cartesian axes, X_1 and X_2, can be represented by the two orthonormal vectors $\mathbf{e}_1 = (1\ 0)$ and $\mathbf{e}_2 = (0\ 1)$, respectively. Consider any other vector $\mathbf{a} = (a_1\ a_2)$ represented as a linear combination of vectors $\mathbf{e}_1$ and $\mathbf{e}_2$. That is,

$$\begin{aligned}\mathbf{a} &= a_1\mathbf{e}_1 + a_2\mathbf{e}_2 \\ &= a_1(1\ 0) + a_2(0\ 1) \\ &= (a_1\ a_2)\end{aligned}$$

where a_1 and a_2 are, respectively, the X_1 and X_2 coordinates. The three vectors, **a**, $\mathbf{e}_1$, and $\mathbf{e}_2$, are said to be linearly dependent as any one of them can be represented by a linear combination of the other two vectors. On the other hand, vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ are linearly independent as neither vector can be represented as a linear combination of the other vector. Similarly, vectors $\mathbf{e}_1$ and **a** are linearly independent and so are vectors $\mathbf{e}_2$ and **a**. In general, a set of p vectors, $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_p$, is linearly independent if no one vector is a linear combination of the other vector(s).

2.5.1 Dimensionality

Any point or vector in a two-dimensional space can be represented as a linear combination of the two basis vectors, $\mathbf{e}_1$ and $\mathbf{e}_2$. Alternatively, it can be said that the two basis vectors span the entire two-dimensional space. The number of linearly indepen-

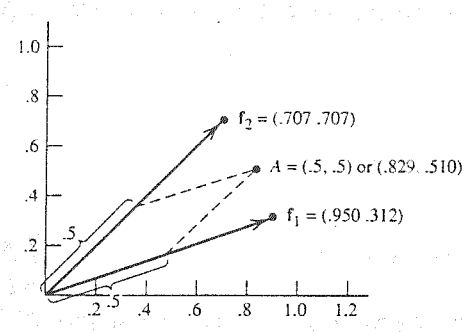


Figure 2.19 Change in basis.

dent vectors that span a given space determines the dimensionality of the space. Since the two basis vectors, $\mathbf{e}_1$ and $\mathbf{e}_2$, are orthonormal, the basis represented by these two vectors is called *orthonormal basis*.

The basis vectors representing a given dimension do not have to be orthonormal. For example, consider the vectors

$$\mathbf{f}_1 = .950\mathbf{e}_1 + .312\mathbf{e}_2 = (.950 .312)$$

and

$$\mathbf{f}_2 = .707\mathbf{e}_1 + .707\mathbf{e}_2 = (.707 .707)$$

shown in Figure 2.19. Each of the vectors, $\mathbf{f}_1$ and $\mathbf{f}_2$, is a linear combination of vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ and has a unit length. However, the two vectors are not orthogonal to each other since the scalar product of $\mathbf{f}_1$ and $\mathbf{f}_2$ is not zero. Vectors which are not orthogonal to each other are referred to as oblique vectors. Furthermore, $\mathbf{f}_1$ and $\mathbf{f}_2$ are linearly independent and therefore can be used to form the basis for the two-dimensional space. That is, they can be used as basis vectors and the basis is referred to as an *oblique basis*.

2.6 CHANGE IN BASIS

Any vector or point in a p-dimensional space can be represented with respect to an orthonormal or an oblique basis. For example, point A in Figure 2.19 can be represented with respect to the orthonormal basis given by $\mathbf{e}_1$ and $\mathbf{e}_2$, or by the oblique basis given by $\mathbf{f}_1$ and $\mathbf{f}_2$. First, let us represent $A = (.5, .5)$ with respect to $\mathbf{f}_1$ and $\mathbf{f}_2$. The representation implies that point A can be reached by first traveling 0.5 units from O along $\mathbf{f}_1$ and then 0.5 units parallel to $\mathbf{f}_2$. However, representation using an orthonormal basis is easier to work with and, therefore, this basis is used in most of the multivariate techniques. The process of changing the representation from one basis to another basis is called *change in basis*. A brief discussion of the process used for changing the basis follows.

Vector $\mathbf{a}$, representing point A, is given by

$$\mathbf{a} = .5\mathbf{f}_1 + .5\mathbf{f}_2 \tag{2.18}$$

with respect to the oblique basis vectors $\mathbf{f}_1$ and $\mathbf{f}_2$. Vectors $\mathbf{f}_1$ and $\mathbf{f}_2$ can themselves be represented as

$$\mathbf{f}_1 = .950\mathbf{e}_1 + .312\mathbf{e}_2 \tag{2.19}$$

and

$$\mathbf{f}_2 = .707\mathbf{e}_1 + .707\mathbf{e}_2 \tag{2.20}$$

with respect to the orthonormal basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$. Substituting Eqs. 2.19 and 2.20 in Eq. 2.18 results in

$$\begin{aligned}\mathbf{a} &= .5(.950\mathbf{e}_1 + .312\mathbf{e}_2) + .5(.707\mathbf{e}_1 + .707\mathbf{e}_2)\\ &= (.950 \times .5 + .707 \times .5)\mathbf{e}_1 + (.312 \times .5 + .707 \times .5)\mathbf{e}_2\\ &= .829(1\ 0) + .510(0\ 1)\\ &= (.829\ .510).\end{aligned}$$

That is, $\mathbf{a} = (.829\ .510)$ with respect to the orthonormal basis vectors, or $\mathbf{a} = (.5\ .5)$ with respect to the oblique basis vectors. Alternatively, one can say that the coordinates of point A with respect to the orthonormal basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ are, respectively, .829 and .510, and with respect to the oblique basis vectors are, respectively, .5 and .5. In general, any arbitrary oblique basis can be transformed into an orthonormal basis using *Gram-Schmidt orthonormalization* procedure. For further details about this procedure see Green (1976).

2.7 REPRESENTING POINTS WITH RESPECT TO NEW AXES

Many statistical techniques essentially reduce to representing points with respect to a new basis, which is typically orthonormal. This section illustrates how points can be represented with respect to new bases and hence new axes. In Figure 2.20, let $\mathbf{e}_1$ and $\mathbf{e}_2$ be the two orthonormal basis vectors representing the axes X_1 and X_2, respectively, and let $\mathbf{a} = (a_1\ a_2)$ represent point A. That is,

$$\mathbf{a} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2. \tag{2.21}$$

Let $\mathbf{e}_1^*$ and $\mathbf{e}_2^*$ be another orthonormal basis such that $\mathbf{e}_1^*$ and $\mathbf{e}_2^*$, respectively, make an angle of $\theta°$ with $\mathbf{e}_1$ and $\mathbf{e}_2$. Using Eq. 2.14, the length of the projection vector of $\mathbf{e}_1$ on $\mathbf{e}_1^*$ is given by

$$\|\mathbf{e}_1\| \cos\theta$$

which is equal to $\cos\theta$ as $\|\mathbf{e}_1\|$ is of unit length. That is, the component or coordinate of $\mathbf{e}_1$ with respect to $\mathbf{e}_1^*$ is equal to $\cos\theta$. Similarly, the component or coordinate of $\mathbf{e}_1$ with respect to $\mathbf{e}_2^*$ is given by $\cos(90 + \theta) = -\sin\theta$. The vector $\mathbf{e}_1$ can now be represented

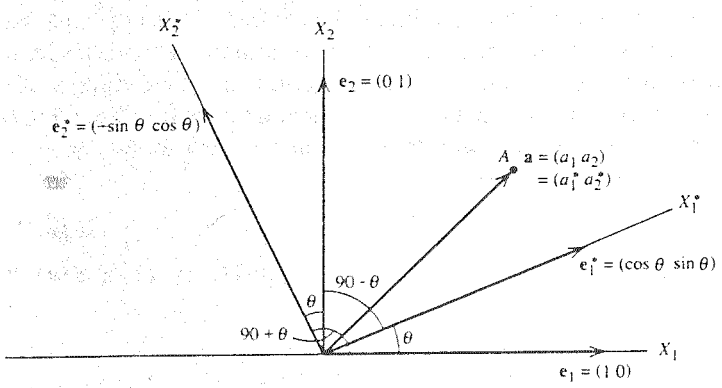


Figure 2.20 Representing points with respect to new axes.

with respect to $\mathbf{e}_1^*$ and $\mathbf{e}_2^*$ as

$$\mathbf{e}_1 = (\cos\theta \quad -\sin\theta)$$

or

$$\mathbf{e}_1 = \cos\theta \times \mathbf{e}_1^* - \sin\theta \times \mathbf{e}_2^*. \tag{2.22}$$

Similarly, $\mathbf{e}_2$ can be represented as

$$\mathbf{e}_2 = \sin\theta \times \mathbf{e}_1^* + \cos\theta \times \mathbf{e}_2^*. \tag{2.23}$$

Substituting Eqs. 2.22 and 2.23 in Eq. 2.21 we get

$$\begin{aligned}\mathbf{a} &= a_1(\cos\theta \times \mathbf{e}_1^* - \sin\theta \times \mathbf{e}_2^*) + a_2(\sin\theta \times \mathbf{e}_1^* + \cos\theta \times \mathbf{e}_2^*)\\ &= (\cos\theta \times a_1 + \sin\theta \times a_2)\mathbf{e}_1^* + (-\sin\theta \times a_1 + \cos\theta \times a_2)\mathbf{e}_2^*.\end{aligned}$$

That is, the coordinates of point A with respect to $\mathbf{e}_1^*$ and $\mathbf{e}_2^*$ are

$$a_1^* = \cos\theta \times a_1 + \sin\theta \times a_2 \tag{2.24}$$

$$a_2^* = -\sin\theta \times a_1 + \cos\theta \times a_2. \tag{2.25}$$

It is clear that the coordinates of A with respect to the new axes are linear combinations of the coordinates with respect to the old axes. The following points can be summarized from the preceding discussion.

1. The new axis, X_1^*, can be viewed as the resulting axis obtained by rotating the X_1 axis counterclockwise by $\theta°$, and the new axis, X_2^*, can be viewed as the resulting axis obtained by rotating the X_2 axis counterclockwise by $\theta°$. That is, the original axes are rotated to obtain a new set of axes. Such a rotation is called an orthogonal rotation and the new set of axes can be used as the new basis vectors.
2. Points can be represented with respect to any axes in the given dimensional space. The coordinates of the points with respect to the new axes are linear combinations of the coordinates with respect to the original axes.

2.8 SUMMARY

In this chapter we provided a geometrical view of some of the basic concepts that will be used to discuss many of the multivariate statistical techniques. The material presented in this chapter can be summarized as follows:

1. Points in a given space can be represented as vectors.
2. Points or vectors in a space can only be located relative to some reference point and a set of linearly independent vectors. This reference point is called the origin and is represented by the null vector $\mathbf{0}$. The set of linearly independent vectors are called the basis vectors or axes.
3. Vectors can be multiplied by a scalar, k. Multiplying a vector by k stretches the vector if $|k| > k$ and compresses the vector if $|k| < 1$. The direction of the vector is preserved for positive scalars, whereas for negative scalars the direction is reversed.
4. Vectors can be added or subtracted to form new vectors. The new vector can be viewed as a linear combination of the vectors added or subtracted.
5. The basis vectors can be orthogonal or oblique. If the basis vectors are orthogonal, the space is commonly referred to as an orthogonal space; if the basis vectors are not orthogonal, the space is referred to as an oblique space.

6. One can easily change the basis vectors. That is, one can easily change the representation of a given vector or point from one basis to another basis. In other words, an arbitrary basis (oblique or orthogonal) can easily be transformed into an orthonormal basis.
7. Coordinates of points with respect to a new set of axes that are obtained by rotation of the original set of axes are linear combinations of the coordinates of the points with respect to the original axes.

QUESTIONS

2.1 The coordinates of points in three-dimensional space are given as A = (4, −1, 0); B = (2, 2, 2); C = (−5, 3, −2).

Compute the euclidean distances between points:

(a) A and B
(b) B and C
(c) C and O (where O is origin).

2.2 If **a**, **b**, and **c** are the vectors representing the points A, B, and C (of Question 2.1) in three-dimensional space, compute:

(a) **a** + **c**
(b) 3**a** − 2**b** + 5**c**
(c) Scalar product **ac**.

2.3 Vectors **a**, **b**, and **c** are given as: **a** = (3 2); **b** = (−5 0); **c** = (3 −2).

Compute: (a) Lengths of **a**, **b**, and **c**; (b) Angle between **a** & **b**, and **a** & **c**; (c) Distance between **a** & **b**, and **a** & **c**; (d) Projection of **a** on **b**, and **a** on **c**.

If vector **d** = (2 −3), determine: (e) Whether **a** and **d** are orthogonal; and (f) Projection of **a** on **d**.

2.4 The coordinates of points A, B, and C with respect to orthogonal axes X_1 and X_2 are: A = (3, −2); B = (−5, 3); C = (0, 1).

Compute the euclidean distances between points A and B, B and C, C and A. If the origin O is shifted to a point O^* such that the coordinates of O^* with respect to the previous origin O are (2, −5), compute the coordinates of points A, B, and C with respect to O^*. Confirm that the shift of origin has not changed the orientation of the points by recomputing the distances between A and B, B and C, C and A, using the new coordinates.

2.5 (a) Points A and B have the following coordinates with respect to orthogonal axes X_1 and X_2: A = (3, −2); B = (5, 1).

If the axes X_1 and X_2 are rotated 20° counterclockwise to produce a new set of orthogonal axes X_1^* and X_2^*, find the coordinates of A and B with respect to X_1^* and X_2^*.

(b) Coordinates of a point A with respect to an orthogonal set of axes X_1 and X_2 are (5, 2). The axes X_1 and X_2 are rotated clockwise by an angle θ. If the new coordinates of the point A with respect to the rotated axes are (3.69, 3.93), find θ.

2.6 $\mathbf{e}_1$ and $\mathbf{e}_2$ are the basis vectors representing the orthogonal axes E_1 and E_2, and $\mathbf{f}_1$ and $\mathbf{f}_2$ are oblique vectors representing the oblique axes F_1 and F_2. Vectors **a** and **b** are given as follows:

$$\mathbf{a} = 0.500\mathbf{e}_1 + 0.866\mathbf{e}_2$$
$$\mathbf{b} = 0.700\mathbf{f}_1 + 0.500\mathbf{f}_2.$$

If the relationship between the orthogonal and oblique axes is given by

$$\mathbf{f}_1 = 0.800\mathbf{e}_1 + 0.600\mathbf{e}_2$$
$$\mathbf{f}_2 = 0.707\mathbf{e}_1 + 0.707\mathbf{e}_2$$

represent **a** with respect to $\mathbf{f}_1$ and $\mathbf{f}_2$ and **b** with respect to $\mathbf{e}_1$ and $\mathbf{e}_2$. What is the angle between **a** and **b**?

2.7 Two cars start from a stadium after a football game. Car A travels east at an average speed of 50 miles per hour while car B travels northeast at an average speed of 55 miles per hour. What is the distance (euclidean) between the two cars after 1 hour and 45 minutes?

2.8 Cities A and B are separated by a 2.5-mile-wide river. Tom wants to swim across from a point X in city A to a point Y in city B that is directly across from point X. If the speed of the current in the river is 15 miles per hour (flowing from Tom's right to his left), in what direction should Tom swim from X to reach Y in 1 hour (indicate direction as an angle from the straight line connecting X and Y).

2.9 A spaceship *Enterprise* from planet Earth meets a spaceship *Bakh-ra* from planet Kling-on, in outer space. The instruments on *Bakh-ra* have ceased working because of a malfunction. *Bakh-ra*'s captain requests the captain of *Enterprise* to help her determine her position.

Enterprise's instruments indicate that its position is (0.5,2). The instruments use the Sun as the origin of an orthogonal system of axes, and measure distance in light years. The Kling-on inhabitants, however, use an oblique system of axes (with the Sun as the origin). *Enterprise*'s computers indicate that the relation between the two systems of axes is given by:

$$\mathbf{k}_1 = 0.810\mathbf{e}_1 + 0.586\mathbf{e}_2$$
$$\mathbf{k}_2 = 0.732\mathbf{e}_1 + 0.681\mathbf{e}_2$$

where the $\mathbf{k}_i$'s and $\mathbf{e}_i$'s are the basis vectors used by the inhabitants of Kling-on and Earth respectively.

As captain of the *Enterprise* how would you communicate *Bakh-ra*'s position to its captain using their system of axes?

According to Earth scientists (who use an orthogonal system of axes), Kling-on's position with respect to the Sun is (2.5, 3.2) (units in light years) and Earth's position with respect to the Sun is (5.2, −1.5). What is the distance between Earth and Kling-on?

Note: In solving this problem assume that the Sun, Earth, Kling-on, and the two spaceships are on the same plane.

Hint: It might be helpful to sketch a picture of the relative positions of the ships, planets, etc. before solving the problem.

CHAPTER 3

Fundamentals of Data Manipulation

Almost all the statistical techniques use summary measures such as means, sum of squares and cross products, variances and covariances, and correlations as inputs for performing the necessary data analysis. These summary measures are computed from the raw data. The purpose of this chapter is to provide a brief review of summary measures and the data manipulations used to obtain them.

3.1 DATA MANIPULATIONS

For discussion purposes, we will use a hypothetical data set given in Table 3.1. The table gives two financial ratios, X_1 and X_2, for 12 hypothetical companies.[1]

3.1.1 Mean and Mean-Corrected Data

A common measure that is computed for summarizing the data is the central tendency. One of the measures of central tendency is the mean or the average. The mean, $\bar{X}_j$, for the jth variable is given by:

$$\bar{X}_j = \frac{\sum_{i=1}^{n} X_{ij}}{n} \tag{3.1}$$

where X_{ij} is the ith observation for the jth variable and n is the number of observations. Data can also be represented as deviations from the mean or the average. Such data are usually referred to as *mean-corrected data*, which are typically used to compute the summary measures. Table 3.1 also gives the mean for each variable and the mean-corrected data.

3.1.2 Degrees of Freedom

Almost all of the summary measures and various statistics use degrees of freedom in their computation. Although the formulae used for computing degrees of freedom vary across statistical techniques, the conceptual meaning or the definition of degrees of freedom remains the same. In the following section we provide an intuitive explanation of this important concept.

[1]The financial ratios could be any of the standard accounting ratios (e.g., current ratio, liquidity ratio) that are used for assessing the financial health of a given firm.

Table 3.1 **Hypothetical Financial Data**

	Original Data		Mean-Corrected Data		Standardized Data	
Firm	X_1	X_2	x_1	x_2	x_1	x_2
1	13.000	4.000	7.917	3.833	1.619	1.108
2	10.000	6.000	4.917	5.833	1.006	1.686
3	10.000	2.000	4.917	1.833	1.006	0.530
4	8.000	−2.000	2.917	−2.167	0.597	−0.627
5	7.000	4.000	1.917	3.833	0.392	1.108
6	6.000	−3.000	0.917	−3.167	0.187	−0.915
7	5.000	0.000	−0.083	−0.167	−0.017	−0.048
8	4.000	2.000	−1.083	1.833	−0.222	0.530
9	2.000	−1.000	−3.083	−1.167	−0.631	−0.337
10	0.000	−5.000	−5.083	−5.167	−1.040	−1.493
11	−1.000	−1.000	−6.083	−1.167	−1.244	−0.337
12	−3.000	−4.000	−8.083	−4.167	−1.653	−1.204
Mean	5.083	.167	0.000	0.000	0.000	0.000
SS			262.917	131.667	11.000	11.000
Var	23.902	11.970	23.902	11.970	1.000	1.000

The degrees of freedom represent the independent pieces of information contained in the data set that are used for computing a given summary measure or statistic. We know that the sum, and hence the mean, of the mean-corrected data is zero. Therefore, the value of any nth mean-corrected observation can be determined from the sum of the remaining $n-1$ mean-corrected observations. That is, there are only $n-1$ *independent* mean-corrected observations, or only $n-1$ pieces of information in the mean-corrected data. The reason there are only $n-1$ independent mean-corrected observations is that the mean-corrected observations were obtained by subtracting the mean from each observation, and one piece or bit of information is used up for computing the mean. The degrees of freedom for the mean-corrected data, therefore, is $n-1$. Any summary measure computed from sample mean-corrected data (e.g., variance) will have $n-1$ degrees of freedom.

(امکان، جدول توافقی ؟) تقاطعی

As another example, consider the two-way contingency table or crosstabulation given in Table 3.2 which represents the joint-frequency distribution for two variables: the number of telephone lines owned by a household and the household income. The numbers in the column and row totals are marginal frequencies for each variable, and

Table 3.2 **Contingency Table**

	Number of Phone Lines Owned		
Income	One	Two or More	Total
Low	150		200
High			200
Total	200	200	400

the number in the cell is the joint frequency. Only one joint frequency is given in the table: the number of households that own one phone line and have a low income, which is equal to 150. The other joint frequencies can be computed from the marginal frequencies and the one joint frequency. For example, the number of low-income households with two or more phone lines is equal to 50 (i.e., 200 − 150); the number of high-income households with just one phone line is equal to 50 (i.e., 200 − 150); and the number of high-income households with two or more phone lines is equal to 150 (i.e., 200 − 50). That is, if the marginal frequencies of the two variables are known, then *only one* joint-frequency value is necessary to compute the remaining joint-frequency values. The other three joint-frequency values are *dependent* on the marginal frequencies and the one known joint-frequency value. Therefore, the crosstabulation has only one degree of freedom or one independent piece of information.[2]

3.1.3 Variance, Sum of Squares, and Cross Products

پراکندگی

Another summary measure that is computed is a measure for the amount of dispersion in the data set. Variance is the most commonly used measure of dispersion in the data, and it is directly proportional to the amount of variation or information in the data.[3] For example, if all the companies in Table 3.1 had the same value for X_1, then this financial ratio would not contain any information and the variance of X_1 would be zero. There simply would be nothing to explain in the data; all the firms would be homogeneous with respect to X_1. On the other hand, if all the firms had different values for X_1 (i.e., the firms were heterogeneous with respect to this ratio) then one of our objectives could be to determine why the ratio was different across the firms. That is, our objective is to account for or explain the variation in the data. The variance for the jth variable is given by

$$s_j^2 = \frac{\sum_{i=1}^{n} x_{ij}^2}{n-1} = \frac{SS}{df} \tag{3.2}$$

where x_{ij} is the mean-corrected data for the ith observation and the jth variable and n is the number of observations. The numerator in Eq. 3.2 is the sum of squared deviations from the mean and is typically referred to as the *sum of squares* (SS), and the denominator is the *degrees of freedom* (df). Variance, then, is the average square of mean-corrected data for each degree of freedom. The sums of squares for x_1 and x_2, respectively, are 262.917 and 131.667. The variances for the two ratios are, respectively, 23.902 and 11.970.

تغییر توأم

The linear relationship or association between the two ratios can be measured by the covariation between two variables. Covariance, a measure of the covariation between two variables, is given by:

$$s_{jk} = \frac{\sum_{i=1}^{n} x_{ij}x_{ik}}{n-1} = \frac{SCP}{df} \tag{3.3}$$

where s_{jk} is the covariance between variables j and k, x_{ij} is the mean-corrected value of the ith observation for the jth variable, x_{ik} is the mean-corrected value of the ith observation for the kth variable, and n is the number of observations. The numerator

[2]The general computational formula for obtaining the degrees of freedom for a contingency table is given by $(c-1)(r-1)$ where c is the number of columns and r is the number of rows.

[3]Once again, it should be noted that the term *information* is used very loosely and may not necessarily have the same meaning as in *information theory*.

is the sum of the cross products of the mean-corrected data for the two variables and is referred to as the *sum of the cross products* (*SCP*), and the denominator is the *df*. Covariation, then, is simply the average cross product between two variables for each degree of freedom. The *SCP* between the two ratios is 136.375 and hence the covariance between the two ratios is 12.398.

The *SS* and the *SCP* are usually summarized in a sum of squares and cross products (**SSCP**) matrix, and the variances and covariances are usually summarized in a covariance (**S**) matrix. The $\mathbf{SSCP}_t$ and $\mathbf{S}_t$ matrices for the data set of Table 3.1 are:[4]

$$\mathbf{SSCP}_t = \begin{bmatrix} 262.917 & 136.375 \\ 136.375 & 131.667 \end{bmatrix},$$

$$\mathbf{S}_t = \frac{\mathbf{SSCP}_t}{df} = \begin{bmatrix} 23.902 & 12.398 \\ 12.398 & 11.970 \end{bmatrix}.$$

Note that the above matrices are symmetric as the *SCP* (or covariance) between variables j and k is the same as the *SCP* (or covariance) between variables k and j. As mentioned previously, variance of a given variable is a measure of its variation in the data and covariance between two variables is a measure of the amount of covariation between them. However, variances of variables can only be compared if the variables are measured using the same units. Also, although the lower bound for the absolute value of the covariance is zero, implying that the two variables are not linearly associated, it has no upper bound. This makes it difficult to compare the association between two variables across data sets. For this reason data are sometimes standardized.[5]

3.1.4 Standardization

Standardized data are obtained by dividing the mean-corrected data by the respective standard deviation (square root of the variance). Table 3.1 also gives the standardized data. The variances of standardized variables are always 1 and the covariation of standardized variables will always lie between -1 and $+1$. The value will be 0 if there is no linear association between the two variables, -1 if there is perfect inverse linear relationship between the two variables, and $+1$ for a perfect direct linear relationship between the two variables. A special name has been given to the covariance of standardized data. Covariance of two standardized variables is called the *correlation coefficient* or *Pearson product moment correlation*. Therefore, the correlation matrix (**R**) is the covariance matrix for standardized data. For the data in Table 3.1, the correlation matrix is:

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.733 \\ 0.733 & 1.000 \end{bmatrix}.$$

3.1.5 Generalized Variance

In the case of p variables, the covariance matrix consists of p variances and $p(p-1)/2$ covariances. Hence, it is useful to have a single or composite index to measure the amount of variation for all the p variables in the data set. Generalized variance is one such measure. Further discussion of generalized variance is provided in Section 3.5.

[4]The subscript t is used to indicate that the respective matrices are for the total sample.

[5]Sometimes the data are standardized even though the units of measurement are the same. We will discuss this in the next chapter on principal components analysis.

3.1.6 Group Analysis

In a number of situations, one is interested in analyzing data from two or more groups. For example, suppose that the first seven observations (i.e., $n_1 = 7$) in Table 3.1 are data for successful firms and the next five observations (i.e., $n_2 = 5$) are data for failed firms. That is, the total data set consists of two groups of firms: Group 1 consisting of successful firms, and Group 2 consisting of failed firms. One might be interested in determining the extent to which firms in each group are similar to each other with respect to the two variables, and also the extent to which firms of the two groups are different with respect to the two variables. For this purpose:

1. Data for each group can be summarized separately to determine the similarities within each group. This is called within-group analysis.
2. Data can also be summarized to determine the differences between the groups. This is called between-group analysis.

Within-Group Analysis

Table 3.3 gives the original, mean corrected, and standardized data for the two groups, respectively. The **SSCP**, **S**, and **R** matrices for Group 1 are

$$\mathbf{SSCP}_1 = \begin{bmatrix} 45.714 & 33.286 \\ 33.286 & 67.714 \end{bmatrix}, \quad \mathbf{S}_1 = \begin{bmatrix} 7.619 & 5.548 \\ 5.548 & 11.286 \end{bmatrix},$$

Table 3.3 **Hypothetical Financial Data for Groups**

	Original Data		Mean-Corrected Data		Standardized Data	
Firm	X_1	X_2	x_1	x_2	z_{x_1}	z_{x_2}
Group 1						
1	13.000	4.000	4.571	2.429	1.656	0.723
2	10.000	6.000	1.571	4.429	0.569	1.318
3	10.000	2.000	1.571	0.429	0.569	0.128
4	8.000	−2.000	−0.429	−3.571	−0.155	−1.063
5	7.000	4.000	−1.429	2.429	−0.518	0.723
6	6.000	−3.000	−2.429	−4.571	−0.880	−1.361
7	5.000	0.000	−3.429	−1.571	−1.242	−0.468
Mean	8.429	1.571	0.000	0.000	0.000	0.000
SS			45.714	67.714	6.000	6.000
Var	7.619	11.286	7.619	11.286	1.000	1.000
Group 2						
8	4.000	2.000	3.600	3.800	1.332	1.369
9	2.000	−1.000	1.600	0.800	0.592	0.288
10	0.000	−5.000	−0.400	−3.200	−0.148	−1.153
11	−1.000	−1.000	−1.400	0.800	−0.518	0.288
12	−3.000	−4.000	−3.400	−2.200	−1.258	−0.793
Mean	0.400	−1.800	0.000	0.000	0.000	0.000
SS			29.200	30.800	4.000	4.000
Var	7.300	7.700	7.300	7.700	1.000	1.000

and

$$\mathbf{R}_1 = \begin{bmatrix} 1.000 & 0.598 \\ 0.598 & 1.000 \end{bmatrix}.$$

And the SSCP, S, and R matrices for Group 2 are

$$\mathbf{SSCP}_2 = \begin{bmatrix} 29.200 & 22.600 \\ 22.600 & 30.800 \end{bmatrix}, \quad \mathbf{S}_2 = \begin{bmatrix} 7.300 & 5.650 \\ 5.650 & 7.700 \end{bmatrix},$$

and

$$\mathbf{R}_2 = \begin{bmatrix} 1.000 & 0.754 \\ 0.754 & 1.000 \end{bmatrix}.$$

The SSCP matrices of the two groups can be combined or pooled to give a pooled SSCP matrix. The pooled within-group $\mathbf{SSCP}_w$ is obtained by adding the respective *SS*s and *SCP*s of the two groups and is given by:

$$\begin{aligned} \mathbf{SSCP}_w &= \mathbf{SSCP}_1 + \mathbf{SSCP}_2 \\ &= \begin{bmatrix} 74.914 & 55.886 \\ 55.886 & 98.514 \end{bmatrix}. \end{aligned}$$

The pooled covariance matrix, $\mathbf{S}_w$, can be obtained by dividing $\mathbf{SSCP}_w$ by the pooled degrees of freedom (i.e., $n_1 - 1$ plus $n_2 - 1$, or $n_1 + n_2 - 2$, or in general $n_1 + n_2 + \cdots + n_g - G$ where G is the number of groups) and is given by:

$$\mathbf{S}_w = \begin{bmatrix} 7.491 & 5.589 \\ 5.589 & 9.851 \end{bmatrix}.$$

Similarly, the reader can check that the pooled correlation matrix is given by:

$$\mathbf{R}_w = \begin{bmatrix} 1.000 & 0.651 \\ 0.651 & 1.000 \end{bmatrix}$$

The pooled $\mathbf{SSCP}_w$, $\mathbf{S}_w$, and the $\mathbf{R}_w$ matrices give the pooled or combined amount of variation that is present in each group. In other words, the matrices provide information about the similarity or homogeneity of observations in each group. If the observations in each group are similar with respect to a given variable then the *SS* of that variable will be zero; if the observations are not similar (i.e., they are heterogeneous) then the *SS* will be greater than zero. The greater the heterogeneity the greater the *SS* and vice versa.

Between-Group Analysis

The between-group sum of squares measures the degree to which the means of groups differ from the overall or total sample means. Computationally, between-group sum of squares can be obtained by the following formula:

توجه کنید که چون برای داده‌های بعدی است نه میانگین تقسیم شده پس باید $\bar{X}$ بزرگ باشد نه کوچک

$$SS_j = \sum_{g=1}^{G} n_g(\underbrace{\bar{x}_{jg}}_{\bar{X}_{jg}} - \underbrace{\bar{x}_{j.}}_{\bar{X}_{j.}})^2 \qquad j = 1, \ldots, p \tag{3.4}$$

where SS_j is the between-group sum of squares for variable j, n_g is the number of observations in group g, $\bar{x}_{jg}$ is the mean for the jth variable in the gth group, $\bar{x}_{j.}$ is the mean of the jth variable for the total data, and G is the number of groups. For example, from Tables 3.1 and 3.3 the between-group *SS* for X_1 is equal to

$$\begin{aligned} SS_1 &= 7(8.429 - 5.083)^2 + 5(0.400 - 5.083)^2 \\ &= 188.022. \end{aligned}$$

The between-group SCP is given by:

$$SCP_{jk} = \sum_{g=1}^{G} n_g(\bar{x}_{jg} - \bar{x}_{j.})(\bar{x}_{kg} - \bar{x}_{k.}), \tag{3.5}$$

which from Tables 3.1 and 3.3 is equal to

$$\begin{aligned} SCP_{12} &= 7(8.429 - 5.083)(1.571 - 0.167) + 5(0.400 - 5.083)(-1.800 - 0.167) \\ &= 78.942. \end{aligned}$$

However, it is not necessary to use the above equations to compute $\mathbf{SSCP}_b$ as

$$\mathbf{SSCP}_t = \mathbf{SSCP}_w + \mathbf{SSCP}_b. \tag{3.6}$$

For example,

$$\begin{aligned} \mathbf{SSCP}_b &= \begin{bmatrix} 262.917 & 136.375 \\ 136.375 & 131.667 \end{bmatrix} - \begin{bmatrix} 74.914 & 55.886 \\ 55.886 & 98.514 \end{bmatrix} \\ &= \begin{bmatrix} 188.003 & 80.489 \\ 80.489 & 33.153 \end{bmatrix}. \end{aligned}$$

The differences between the SSs and the SCPs of the above matrix and the ones computed using Eqs. 3.4 and 3.5 are due to rounding errors.

The identity given in Eq. 3.6 represents the fact that the total information can be divided into two components or parts. The first component, $\mathbf{SSCP}_w$, is information due to within-group differences and the second component, $\mathbf{SSCP}_b$, is information due to between-group differences. That is, the within-group $SSCP$ matrices provide information regarding the similarities of observations within groups and the between-group $SSCP$ matrices give information regarding differences in observations between or across groups.

It was seen above that the $\mathbf{SSCP}_t$ matrix could be decomposed into $\mathbf{SSCP}_w$ and $\mathbf{SSCP}_b$ matrices. Similarly, the degrees of freedom for the total sample can be decomposed into within-group and between-group dfs. That is,

$$df_t = df_w + df_b.$$

It will be seen in later chapters that many multivariate techniques, such as discriminant analysis and MANOVA, involve further analysis of the between-group and within-group $SSCP$ matrices. For example, it is obvious that the greater the difference between the two groups of firms the greater will be the between-group sum of squares relative to the within-group sum of squares and vice versa.

3.2 DISTANCES

In Chapter 2 we discussed the use of euclidean distance as a measure of the distance between two points or observations in a p-dimensional space. This section discusses other measures of the distance between two points and will show that the euclidean distance is a special case of Mahalanobis distance.

3.2.1 Statistical Distance

In Panel I of Figure 3.1, assume that x is a random variable having a normal distribution with a mean of 0 and a variance of 4.0 (i.e., $x \sim N(0, 4)$). Let $x_1 = -2$ and $x_2 = 2$

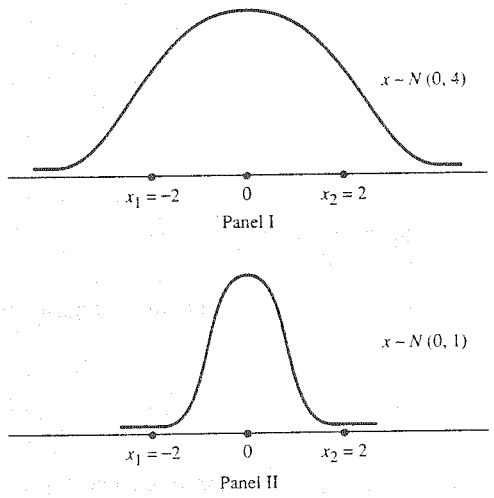


Figure 3.1 Distribution for random variable.

be two observations or values of the random variable x. From Chapter 2, the distance between the two observations can be measured by the squared euclidean distance and is equal to 16 (i.e., $\{2-(-2)\}^2$). An alternative way of representing the distance between the two observations might be to determine the probability of any given observation selected at random falling between the two observations, x_1 and x_2 (i.e., -2 and 2). From the standard normal distribution table, this probability is equal to 0.6826.

If, as shown in Panel II of Figure 3.1, the two observations or values are from a normal distribution with a mean of 0 and a variance of 1, then the probability of a random observation falling between x_1 and x_2 is 0.9544. Therefore, one could argue that the two observations, $x_1 = -2$ and $x_2 = 2$, from the normal distribution with a variance of 4 are *statistically* closer than if the two observations were from a normal distribution whose variance is 1.0, even though the euclidean distances between the observations are the same for both the distributions. It is, therefore, intuitively obvious that the euclidean distance measure must be adjusted to take into account the variance of the variable. This adjusted euclidean distance is referred to as the *statistical* distance or *standard* distance. The squared statistical distance between the two observations is given by

$$SD_{ij}^2 = \left(\frac{x_i - x_j}{s}\right)^2 \tag{3.7}$$

where SD_{ij} and s are, respectively, the statistical distance between observations i and j and the standard deviation. Using Eq. 3.7, the squared statistical distances between the two points are 4 and 16, respectively, for distributions with a variance of 4 and 1. The attractiveness of using the statistical distance in the case of two or more variables is discussed below.

Figure 3.2 gives a scatterplot of observations from a bivariate distribution (i.e., 2 variables). It is clear from the figure that if the euclidean distance is used, then observation A is closer to observation C than to observation B. However, there appears to be a greater probability that observations A and B are from the same distribution than observations A and C are. Consequently, if one were to use the statistical distance then one would conclude that observations A and B are closer to each other than observations

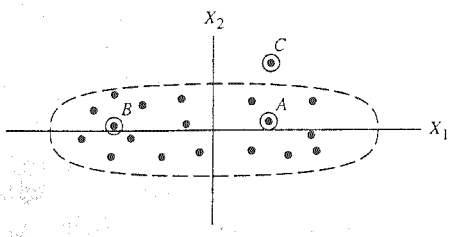


Figure 3.2 Hypothetical scatterplot of a bivariate distribution.

A and C. The formula for squared statistical distance, SD_{ik}^2, between observations i and k for p variables is

$$SD_{ik}^2 = \sum_{j=1}^{p} \left(\frac{x_{ij} - x_{kj}}{s_j} \right)^2. \tag{3.8}$$

Note that in the equation, each term is the square of the standardized value for the respective variable. Therefore, the statistical distance between two observations is the same as the euclidean distance between two observations for standardized data.

3.2.2 Mahalanobis Distance

The scatterplot given in Figure 3.2 is for uncorrelated variables. If the two variables, x_1 and x_2, are correlated then the statistical distance should take into account the covariance or the correlation between the two variables. Mahalanobis distance is defined as the statistical distance between two points that takes into account the covariance or correlation among the variables. The formula for the Mahalanobis distance between observations i and k is given by

$$MD_{ik}^2 = \frac{1}{1 - r^2} \left[\frac{(x_{i1} - x_{k1})^2}{s_1^2} + \frac{(x_{i2} - x_{k2})^2}{s_2^2} - \frac{2r(x_{i1} - x_{k1})(x_{i2} - x_{k2})}{s_1 s_2} \right], \tag{3.9}$$

where s_1^2, s_2^2 are the variances for variables 1 and 2, respectively, and r is the correlation coefficient between the two variables. It can be seen that if the variables are not correlated (i.e., $r = 0$) then the Mahalanobis distance reduces to the statistical distance and if the variances of the variables are equal to one and the variables are uncorrelated then the Mahalanobis distance reduces to the euclidean distance. That is, euclidean and statistical distances are special cases of Mahalanobis distance.

For p-variable case, the Mahalanobis distance between two observations is given by

$$MD_{ik} = (\mathbf{x}_i - \mathbf{x}_k)'\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{x}_k) \tag{3.10}$$

where $\mathbf{x}$ is a $p \times 1$ vector of coordinates and $\mathbf{S}$ is a $p \times p$ covariance matrix. Note that for uncorrelated variables, $\mathbf{S}$ will be a diagonal matrix with variances on the diagonal and for uncorrelated standardized variables $\mathbf{S}$ will be an identity matrix.

Mahalanobis distance is not the only measure of distance between two points that can be used. One could conceivably use other measures of distance depending on the objective of the study. Further discussion about other measures of distance will be provided in Chapter 7. However, irrespective of the distance measure employed, distance measures should be based on the concept of a metric. The metric concept views observations as points in a p-dimensional space. Distances based on this definition of metric possess the following properties.

1. Given two observations, i and k, the distance, D_{ik}, between observations i and k, should be equal to the distance between observations k and i and should be greater than zero. That is, $D_{ik} = D_{ki} > 0$. This property is referred to as *symmetry*.
2. Given three observations, i, k, and l, $D_{il} < D_{ik} + D_{lk}$. This property simply implies that the length of any given side of a triangle is less than the sum of the lengths of the other two sides. This property is referred to as *triangular inequality*.
3. Given two observations i and k, if $D_{ik} = 0$ then i and k are the same observations and if $D_{ik} \neq 0$ then i and k are not the same observations. This property is referred to as *distinguishability* of observations.

3.3 GRAPHICAL REPRESENTATION OF DATA IN VARIABLE SPACE

The data of Table 3.1 can be represented graphically as shown in Figure 3.3. Each observation is a point in the two-dimensional space with each dimension representing a variable. In general, p dimensions are required to graphically represent data having p variables. The dimensional space in which each dimension represents a variable is referred to as variable space.

As discussed in Chapter 2, each point can also be represented by a vector. For presentation clarity only a few points are shown as vectors in Figure 3.3. As shown in the figure, the length of the projection of a vector (or a point) on the X_1 and X_2 axes will give the respective coordinates (i.e., values x_1 and x_2). The means of the ratios can be represented by a vector, called the centroid. Let the centroid, C, be the new origin and let X_1^* and X_2^* be a new set of axes passing through the centroid. As shown in Figure 3.4, the data can also be represented with respect to the new set of axes and the new origin. The length of the projection vectors on the new axes will give the values for the mean-corrected data. The following three observations can be made from Figure 3.4.

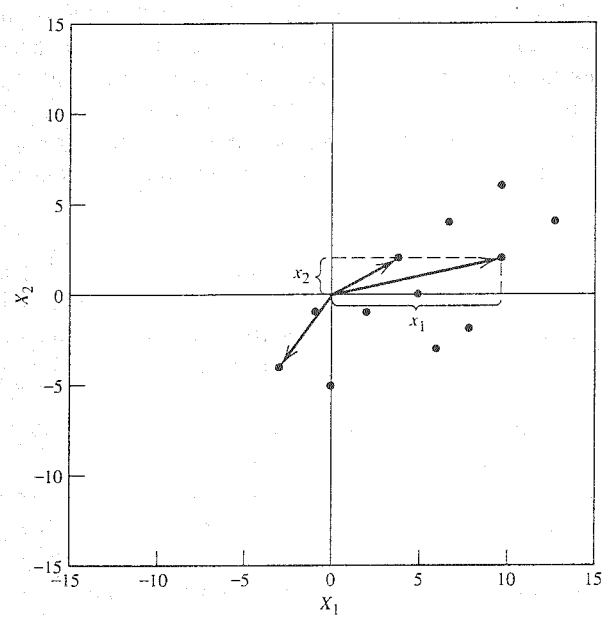


Figure 3.3 Plot of data and points as vectors.

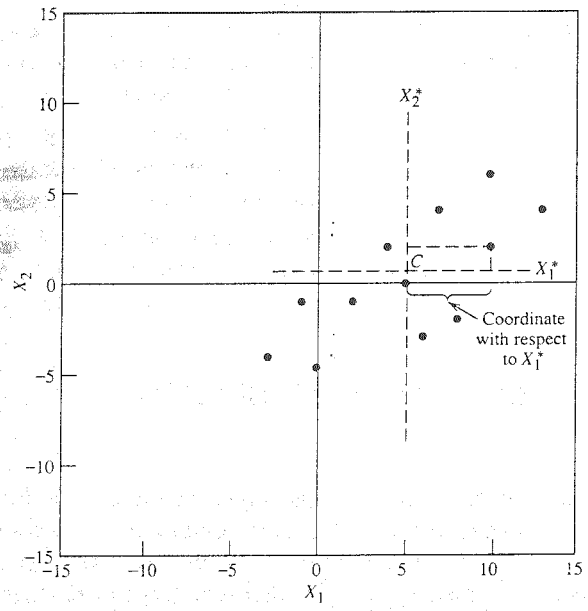


Figure 3.4 Mean-corrected data.

1. The new axes pass through the centroid. That is, the centroid is the origin of the new axes.
2. The new axes are parallel to the respective original axes.
3. The relative positions of the points have not changed. That is, the interpoint distances of the data are not affected. Representing data as deviations from the mean does not affect the orientation of data points and, therefore, without loss of generality, mean-corrected data are used in discussing various statistical techniques. Note that the mean-corrected value for a given variable is obtained by subtracting a constant (i.e., the mean) from each observation. In other words, mean-corrected data represent a change in the measurement scale used. If the subsequent analysis or computations are not affected by the change in scale, then the analysis is said to be *scale invariant*. Almost all of the statistical techniques are scale invariant with respect to mean correcting the data. That is, mean correction of the data does not affect the results.

Standardized data are obtained by dividing the mean-corrected data by the respective standard deviations; that is, the measurement scale of each variable changes and may be different. Division of the data by the standard deviation is tantamount to compressing or stretching the axis. Since the compression or stretching is proportional to the standard deviation, the amount of compression or stretching may not be the same for all the axes. The vectors representing the observations or data points will also move in relation to the amount of stretching and compression of the axes. In Figure 3.5, which gives a representation of the standardized data, it can be observed that the orientation of the data points has changed. And since data standardization changes the configuration of the points or the vectors in the space, the results of some multivariate techniques could be affected. That is, these techniques will not be scale invariant with respect to standardization of the data.

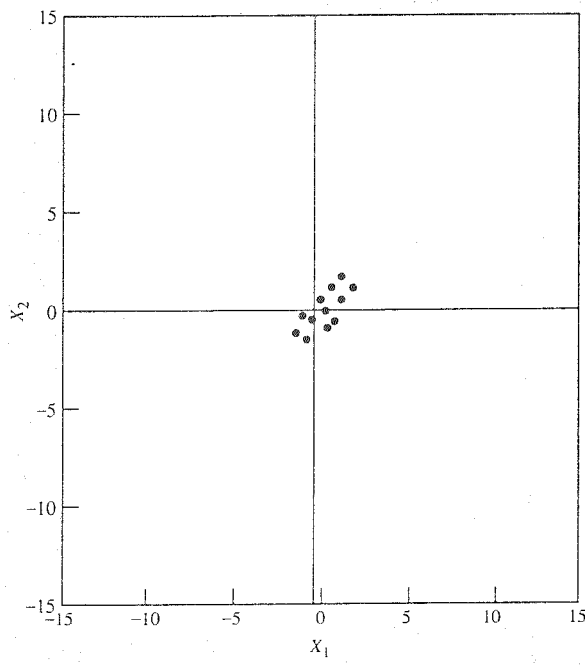


Figure 3.5 Plot of standardized data.

3.4 GRAPHICAL REPRESENTATION OF DATA IN OBSERVATION SPACE

Data can also be represented in a space where each observation is assumed to represent a dimension and the points are assumed to represent the variables. For example, for the data set given in Table 3.1 each observation can be considered as a variable and the x_1 and x_2 variables can be considered as observations. Table 3.4 shows the mean-corrected transposed data. Thus, the transposed data has 12 variables and 2 observations. That is, x_1 and x_2 can be represented as points in the 12-dimensional space. Representing data in a space in which the dimensions are the observations and the points are variables is referred to as *representation of data in the observation space.*

As discussed in Chapter 2, each point can also be represented as a vector whose tail is at the origin and the terminus is at the point. Thus, we have two vectors in a 12-dimensional space, with each vector representing a variable. However, these two vectors will lie in a two-dimensional space embedded in 12 dimensions.[6] Figure 3.6 shows the two vectors, $\mathbf{x}_1$ and $\mathbf{x}_2$, in the two-dimensional space embedded in the 12-dimensional space. The two vectors can be represented as

$$\mathbf{x}_1 = \begin{pmatrix} 7.917 & 4.917 & \cdots & -8.083 \end{pmatrix},$$

and

$$\mathbf{x}_2 = \begin{pmatrix} 3.833 & 5.833 & \cdots & -4.167 \end{pmatrix}.$$

[6]In the case of p variables and n observations, the observation space consists of n dimensions and the vectors lie in a p-dimensional space embedded in an n-dimensional space.

Table 3.4 **Transposed Mean-Corrected Data**

Variables	Observations											
	1	2	3	4	5	6	7	8	9	10	11	12
x_1	7.917	4.917	4.917	2.917	1.917	0.917	−0.083	−1.083	−3.083	−5.083	−6.083	−8.083
x_2	3.833	5.833	1.833	−2.167	3.833	−3.167	−0.167	1.833	−1.167	−5.167	−1.167	−4.167

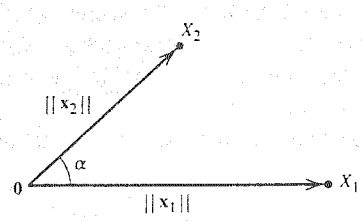


Figure 3.6 Plot of data in observation space.

Note that

1. Since the data are mean corrected, the origin is at the centroid. The average of the mean-corrected ratios is zero, and, therefore, the origin is represented as the null vector $\mathbf{0} = (0\,0)$ implying that the averages of the mean-corrected ratios are zero.
2. Each vector has 12 elements and therefore represents a point in 12 dimensions. However, the two vectors lie in a two-dimensional subspace of the 12-dimensional observation space.
3. Each element or component of $\mathbf{x}_1$ represents the mean-corrected value of x_1 for a given observation. Similarly, each element of $\mathbf{x}_2$ represents the mean-corrected value of x_2 for a given observation.

The squared length of vector, $\mathbf{x}_1$, is given by

$$\|\mathbf{x}_1\|^2 = 7.917^2 + 4.917^2 + \cdots + -8.083^2 = 262.917,$$

which is the same as the *SS* of the mean-corrected data. That is, the squared length of a vector in the observation space gives the *SS* for the respective variable represented by the vector. The variance of x_1 is equal to

$$s_1^2 = \frac{\|\mathbf{x}_1\|^2}{n-1}, \tag{3.11}$$

and the standard deviation is equal to

$$s_1 = \frac{\|\mathbf{x}_1\|}{\sqrt{n-1}}. \tag{3.12}$$

That is, the variance and the standard deviation of a variable are, respectively, equal to the squared length and the length of the vector that has been rescaled by dividing it by $\sqrt{n-1}$ (i.e., the *df*). Using Eqs. 3:11 and 3.12, the variance and standard deviation of $\mathbf{x}_1$ are equal to 23.902 and 4.889, respectively.

Similarly, the squared length of vector $\mathbf{x}_2$ is equal to

$$\|\mathbf{x}_2\|^2 = 3.833^2 + 5.833^2 + \cdots + -4.167^2 = 131.667,$$

and the variance and standard deviation are equal to 11.970 and 3.460, respectively.

For the standardized data, the squared length of the vector is equal to $n-1$ and the length is equal to $\sqrt{n-1}$. That is, standardization is equivalent to rescaling each vector representing the variables in the observation space to have a length of $\sqrt{n-1}$. The scalar product of the two vectors, $\mathbf{x}_1$ and $\mathbf{x}_2$, is given by

$$\begin{aligned}\mathbf{x}_1\mathbf{x}_2 &= (7.917 \times 3.833) + (4.917 \times 5.833) + \cdots + (-8.083 \times -4.167)\\ &= 136.375.\end{aligned}$$

The quantity 136.375 is the *SCP* of mean-corrected data. Therefore, the scalar product of the two vectors gives the *SCP* for the variables represented by the two vectors. Since covariance is equal to $SCP/n-1$, the covariance of two variables is equal to the scalar product of the vectors, which have been rescaled by dividing them by $\sqrt{n-1}$.

From Eq. 2.13 of Chapter 2, the cosine of the angle between the two vectors, $\mathbf{x}_1$ and $\mathbf{x}_2$, is given by

$$\cos\alpha = \frac{\mathbf{x}_1\mathbf{x}_2}{\|\mathbf{x}_1\| \bullet \|\mathbf{x}_2\|} = \frac{136.375}{\sqrt{262.917 \times 131.667}} = .733.$$

This quantity is the same as the correlation between the two variables. Therefore, the cosine of the angle between the two vectors is equal to the correlation between the variables. Notice that if the two vectors are collinear (i.e., they coincide), then the angle between them is zero and the cosine of the angle is one. That is, the correlation between the two variables is one. On the other hand, if the two vectors are orthogonal then the cosine of the angle between them is zero, implying that the two variables are uncorrelated.

3.5 GENERALIZED VARIANCE

As discussed earlier, the covariance matrix for p variables contains p variances and $p(p-1)/2$ covariances. Interpreting these many variances and covariances for assessing the amount of variation in the data could become quite cumbersome for a large number of variables. Consequently, it would be desirable to have a single index that could represent the amount of variation and covariation in the data set. One such index is the generalized variance. Following is a geometric view of the concept of generalized variance.

Figure 3.7 represents variables X_1 and X_2 as vectors in the observation space. The vectors have been scaled by dividing them by $\sqrt{n-1}$, and α is the angle between the two vectors, which can be computed from the correlation coefficient because the correlation between two variables is equal to the cosine of the angle between the respective vectors in the observation space. The figure also shows the parallelogram formed by the two vectors. Recall that if X_1 and X_2 are perfectly correlated then vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ are collinear and the area of the parallelogram is equal to zero. Perfectly correlated variables imply redundancy in the data; i.e., the two variables are not different. On the other hand if the two variables have a zero correlation then the two vectors will be orthogonal, suggesting that there is no redundancy in the data. It is clear from Figure 3.7 that the area of the parallelogram will be minimum (i.e., zero) for collinear vectors and it will be maximum for orthogonal vectors. Therefore, the area of the parallelogram

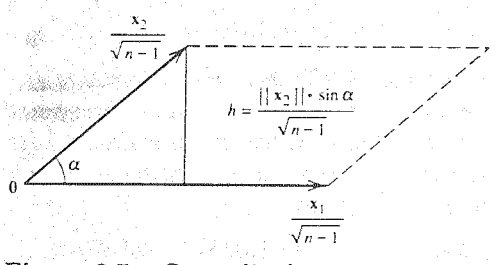


Figure 3.7 Generalized variance.

gives a measure of the amount of redundancy in the data set. The square of the area is used as a measure of the generalized variance.[7] Since the area of a parallelogram is equal to base times height, generalized variance (GV) is equal to

$$GV = \left(\frac{\|\mathbf{x}_1\| \cdot \|\mathbf{x}_2\|}{n-1} \cdot \sin\alpha\right)^2. \tag{3.13}$$

It can be shown (see the Appendix) that the generalized variance is equal to the determinant of the covariance matrix.

For the data set given in Table 3.1, the angle between the two vectors is equal to 42.862° (i.e., $\cos^{-1}.733$), and the generalized variance is

$$GV = \left(\frac{\sqrt{262.917 \times 131.667} \times \sin 42.862}{11}\right)^2 = 132.382.$$

3.6 SUMMARY

Most multivariate techniques use summary measures computed from raw data as inputs for performing the necessary analysis. This chapter discusses these manipulations, a summary of which follows.

1. Procedures for computing the mean, mean-corrected data, sum of squares and cross products, and variance of the variables and standardized data are discussed.
2. Mean correcting the data does not affect the results of the multivariate techniques; however, standardization can affect the results of some of the techniques.
3. Degrees of freedom is an important concept in statistical techniques, and it represents the number of independent pieces of information contained in the data set.
4. When the data can be divided into a number of groups, data manipulation can be done for each group to assess similarities and differences within and across groups. This is called within- and between-group analysis. Within-group analysis pertains to determining similarities of the observations within a group, and between-group analysis pertains to determining differences of the observations across groups.
5. The use of statistical distance, as opposed to euclidean distance, is preferred because it takes into account the variance of the variables. The statistical distance is a special case of Mahalanobis distance, which takes into account the correlation among the variables.
6. Data can be represented in variable or observation space. When data are represented in observation space, each variable is a vector in the n-dimensional space and the length of the vector is proportional to the standard deviation of the variable represented by the vector. The scalar or the inner dot product of two vectors is proportional to the covariance between the two respective variables represented by the vectors. The cosine of the angle between two vectors gives the correlation between the two variables represented by the two vectors.
7. The generalized variance of the data is a single index computed to represent the amount of variation in the data. Geometrically, it is given by the square of the hypervolume of the parallelopiped formed by the vectors representing the variables in the observation space. It is also equal to the determinant of the covariance matrix.

[7]In the case of p variables the generalized variance is given by the square of the hypervolume of the parallelopiped formed by the p vectors in the observation space.

QUESTIONS

3.1 Explain the differences between the three distance measures: euclidean, statistical, and Mahalanobis. Under what circumstances would you use one versus the other?

Given the following data, compute the euclidean, statistical, and Mahalanobis distance between observations 2 and 4 and observations 2 and 3. Which set of observations is more similar? Why? (Assume sample estimates are equal to population values.)

Obs.	X_1	X_2
1	7	8
2	3	1
3	9	8
4	2	4
5	5	5

3.2 Data on number of years of formal education and annual income were collected from 200 respondents. The data are presented in the following table:

Annual Income ($ thous.)	Years of Formal Education 0–10	11–14	>14	Total
20–40	50	X	10	70
41–60	X	30	X	80
>60	X	X	20	50
Total	100	60	40	200

Fill in the missing values (X's) in the above table. How many degrees of freedom does the table have?

3.3 A household appliance manufacturing company conducted a consumer survey on their "IC-Kool" brand of refrigerators. Rating data (on a 10-point scale) were collected on attitude (X_1), opinion (X_2), and purchase intent (PI) for IC-Kool. The data are presented below:

Obs.	Attitude Score (X_1)	Opinion Score (X_2)	Intent Score (PI)
1	4	6	5
2	3	5	6
3	6	8	8
4	8	8	7
5	5	4	5
6	2	7	8
7	4	6	7
8	8	9	8
9	2	3	3
10	2	2	1
11	7	6	4
12	6	5	3
13	2	4	4
14	5	7	4
15	3	5	4

(a) Reconstruct the data by (i) mean correction and (ii) standardization.
(b) Compute the variance, sum of squares, and sum of cross products.
(c) Compute the covariance and correlation matrices.

3.4 For the data in Question 3.3, assume that purchase intent (PI) is determined by opinion alone.

(a) Represent the data on PI and opinion graphically in two-dimensional space.
(b) How does the graphical representation change when the data are (i) mean corrected and (ii) standardized?

3.5 For the data in Question 3.3, assume that observations 1–8 belong to group 1 and the rest belong to group 2.

(a) Compute the within-group and between-group sum of squares.
(b) Deduce from the above if the grouping is justified; i.e., are there similarities within the groups for each of the variables?

3.6 The sum of squares and cross products matrices for two groups are given below:

$$\mathbf{SSCP}_t = \begin{pmatrix} 100 & 56 \\ 56 & 200 \end{pmatrix} \qquad \mathbf{SSCP}_1 = \begin{pmatrix} 10 & 45 \\ 45 & 100 \end{pmatrix} \qquad \mathbf{SSCP}_2 = \begin{pmatrix} 10 & 10 \\ 10 & 15 \end{pmatrix}.$$

Compute the $\mathbf{SSCP}_w$ and $\mathbf{SSCP}_b$ matrices. What conclusions can you draw with respect to the two groups?

3.7

Obs.	X_1	X_2	X_3
1	2	2	1
2	3	1	4
3	1	2	4
4	4	3	2
5	3	5	1
6	5	4	2

For the preceding table, compute (a) SS_1, SS_2, SS_3; and (b) $SCP_{12}, SCP_{23}, SCP_{13}$.

3.8 In a study designed to determine price sensitivity of the sales of Brand X, the following data were collected:

Sales (S) ($ mil.)	Price per Unit (P) ($)
5.1	1.25
5.0	1.30
5.0	1.35
4.8	1.40
4.2	1.50
4.0	1.55

(a) Reconstruct the data by mean correction; (b) Represent the data in subject space, i.e., find the vectors **s** and **p**; (c) Compute the lengths of **s** and **p**; (d) Compute the sum of cross products (SCP) for the variables S and P; (e) Compute the correlation between S and P; and (f) Repeat steps (a) through (e) using matrix operation in PROC IML of SAS.

3.9 The following table gives the mean-corrected data on four variables.

Obs.	X_1	X_2	X_3	X_4
1	2	−3	6	0
2	−1	1	1	0
3	0	1	−3	−1
4	1	2	−2	1
5	−1	1	2	2
6	−1	−2	−4	−2

(a) Compute the covariance matrix Σ; and (b) Compute the generalized variance.

3.10 Show that $\mathbf{SSCP}_t = \mathbf{SSCP}_b + \mathbf{SSCP}_w$.

Appendix

In this appendix we show that generalized variance is equal to the determinant of the covariance matrix. Also, we show how the PROC IML procedure in SAS can be used to perform the necessary matrix operations for obtaining summary measures discussed in Chapter 3.

A3.1 GENERALIZED VARIANCE

The covariance matrix for variables X_1 and X_2 is given by

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} \\ s_{21} & s_2^2 \end{bmatrix}.$$

Since $s_{12} = rs_1s_2$, where r is the correlation between the two variables, the above equation can be rewritten as

$$\mathbf{S} = \begin{bmatrix} s_1^2 & rs_1s_2 \\ rs_2s_1 & s_2^2 \end{bmatrix}.$$

The determinant of the above matrix is given by[1]

$$\begin{aligned} |\mathbf{S}| &= s_1^2s_2^2 - r^2s_1^2s_2^2 \\ &= s_1^2s_2^2(1 - r^2) \\ &= s_1^2s_2^2(1 - \cos^2\alpha) \\ &= (s_1s_2\sin\alpha)^2 \end{aligned} \tag{A3.1}$$

[1]The procedure for computing the determinant of a matrix is quite complex for large matrices. The PROC IML procedure in SAS can be used to compute the determinant of the matrix. The interested reader can consult any textbook on matrix algebra for further details regarding the determinant of matrices.

as $r = \cos\alpha$ and $\sin^2\alpha + \cos^2\alpha = 1$. From Eq. 3.13, the standard deviations of X_1 and X_2 are equal to

$$s_1 = \frac{\|\mathbf{x}_1\|}{\sqrt{n-1}} \tag{A3.2}$$

$$s_2 = \frac{\|\mathbf{x}_2\|}{\sqrt{n-1}}. \tag{A3.3}$$

Substituting Eqs. A3.2 and A3.3 in Eq. A3.1 we get

$$|\mathbf{S}| = \left(\frac{\|\mathbf{x}_1\| \cdot \|\mathbf{x}_2\|}{n-1}\sin\alpha\right)^2. \tag{A3.4}$$

The above equation is the same as Eq. 3.13 for generalized variance.

A3.2 USING PROC IML IN SAS FOR DATA MANIPULATIONS

Suppose we have an $n \times p$ data matrix $\mathbf{X}$ and a $1 \times n$ unit row vector $\mathbf{1}'$.[2] The mean or the average is given by

$$\bar{\mathbf{x}}' = \frac{1}{n}\mathbf{1}'\mathbf{X}, \tag{A3.5}$$

and the mean-corrected data are given by

$$\mathbf{X}_m = \mathbf{X} - \mathbf{1}\mathbf{x}' \tag{A3.6}$$

where $\mathbf{X}_m$ gives the matrix containing the mean-corrected data. The $\mathbf{SSCP}_m$ matrix is given by

$$\mathbf{SSCP}_m = \mathbf{X}_m'\mathbf{X}_m, \tag{A3.7}$$

and the covariance matrix is given by

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}_m'\mathbf{X}_m. \tag{A3.8}$$

Now if we define a diagonal matrix, $\mathbf{D}$, which has variances of the variables in the diagonal, then standardized data are given by

$$\mathbf{X}_s = \mathbf{X}_m\mathbf{D}^{-\frac{1}{2}}. \tag{A3.9}$$

The $\mathbf{SSCP}_s$ of standardized data is given by

$$\mathbf{SSCP}_s = \mathbf{X}_s'\mathbf{X}_s, \tag{A3.10}$$

the correlation matrix is given by

$$\mathbf{R} = \frac{1}{n-1}\mathbf{SSCP}_s, \tag{A3.11}$$

and the generalized variance is given by $|\mathbf{S}|$.

[2]Until now we did not differentiate between row and column vectors. Henceforth, we will use the standard notation to differentiate row from column vectors (i.e., the $'$ symbol will be used to indicate the transpose of a vector or matrix).

For the data set of Table 3.1 the above matrix manipulations can be made by assuming that

$$\mathbf{X}' = \begin{pmatrix} 13 & 10 & 10 & 8 & 7 & 6 & 5 & 4 & 2 & 0 & -1 & -3 \\ 4 & 6 & 2 & -2 & 4 & -3 & 0 & 2 & -1 & -5 & -1 & -4 \end{pmatrix}.$$

and $\mathbf{1}' = (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1)$. Table A3.1 gives the necessary PROC IML commands in SAS for the various matrix manipulations discussed in Chapter 3 and the resulting output is given in Exhibit A3.1. Note that the summary measures given in the exhibit are the same as those reported in Chapter 3. Following is a brief discussion of the PROC IML commands. The reader should consult the SAS/IML User's Guide (1985) for further details.

The DATA TEMP command reads the data from the data file into an SAS data set named TEMP. The PROC IML command invokes the IML procedure and the USE command specifies the SAS data set from which data are to be read. The READ ALL INTO X command reads the data into the X matrix whose rows are equal to the number of observations and whose columns are equal to the number of variables in the TEMP data set. In the N=NROW(X) command, N gives the number of rows for matrix X. ONE=J(N,1,1) command creates an $N \times 1$ vector ONE with all the elements equal to one. The D=DIAG(S) command creates a diagonal matrix D from the symmetric S matrix such that the elements of D are the same as the diagonal elements of S. The INV(D) command computes the inverse of the D matrix and the DET(S) command computes the determinant of the S matrix. The PRINT command requests the printing of the various matrices that have been computed.

Table A3.1 **PROC IML Commands for Data Manipulations**

```
TITLE PROC IML COMMANDS FOR MATRIX MANIPULATIONS ON
DATA IN TABLE 3.1;
OPTIONS NOCENTER;
DATA TEMP;
 INPUT X1 X2;
CARDS;
insert data here
PROC IML;
USE TEMP;
READ ALL INTO X; * READ DATA INTO X MATRIX;
N=NROW(X); * N CONTAINS THE NUMBER OF OBSERVATIONS;
ONE=J(N,1,1); * 12X1 VECTOR CONTAINING ONES;
DF=N-1;
MEAN=(ONE`*X)/N; * MEAN MATRIX CONTAINS THE MEANS;
XM=X-ONE*MEAN;  * XM MATRIX CONTAINS THE MEAN-CORRECTED
DATA;
SSCPM=XM`*XM;
S=SSCPM/DF;
D=DIAG(S);
XS=XM*SQRT(INV(D)); * XS MATRIX CONTAINS THE
STANDARDIZED DATA;
R=XS`*XS/(N-1); * R IS THE CORRELATION MATRIX;
GV=DET(S);  * GV GIVES THE DETERMINANT OF THE S MATRIX;
PRINT MEAN,XM,SSCPM,S,XS,R,GV;
```

Exhibit A3.1

PROC IML output

```
1PROC IML COMMANDS FOR MATRIX MANIPULATIONS ON DATA IN
TABLE 3.1                                          11:12
Friday, July 2, 1993   1

     MEAN
5.0833333 0.1666667


       XM
7.9166667 3.8333333
4.9166667 5.8333333
4.9166667 1.8333333
2.9166667 -2.166667
1.9166667 3.8333333
0.9166667 -3.166667
-0.083333 -0.166667
-1.083333 1.8333333
-3.083333 -1.166667
-5.083333 -5.166667
-6.083333 -1.166667
-8.083333 -4.166667


    SSCPM
262.91667 134.83333
134.83333 131.66667


    S
23.901515 12.257576
12.257576 11.969697


       XS
1.6193087 1.1079879
1.0056759 1.6860685
1.0056759 0.5299072
0.5965874 -0.626254
0.3920432 1.1079879
0.1874989 -0.915294
-0.017045 -0.048173
 -0.22159 0.5299072
-0.630678 -0.337214
-1.039767 -1.493375
-1.244311 -0.337214
-1.653399 -1.204335


        R
        1 0.7246867
0.7246867         1


       GV
135.84573
```

CHAPTER 4

Principal Components Analysis

Consider each of the following scenarios.

- A financial analyst is interested in determining the financial health of firms in a given industry. Research studies have identified a number of financial ratios (say about 120) that can be used for such a purpose. Obviously, it would be extremely taxing to interpret the 120 pieces of information for assessing the financial health of firms. However, the analyst's task would be simplified if these 120 ratios could be reduced to a few indices (say about 3), which are linear combinations of the original 120 ratios.[1]
- The quality control department is interested in developing a few key composite indices from numerous pieces of information resulting from the manufacturing process to determine if the process is or is not in control.
- The marketing manager is interested in developing a regression model to forecast sales. However, the independent variables under consideration are correlated among themselves. That is, there is multicollinearity in the data. It is well known that in the presence of multicollinearity, the standard errors of the parameter estimates could be quite high, resulting in unstable estimates of the regression model. It would be extremely helpful if the marketing manager could form "new" variables, which are linear combinations of the original variables, such that the new variables are uncorrelated among themselves. These new variables could then be used for developing the regression model.

Principal components analysis is the appropriate technique for achieving each of the above objectives. Principal components analysis is a technique for forming new variables which are linear composites of the original variables. The maximum number of new variables that can be formed is equal to the number of original variables, and the new variables are uncorrelated among themselves.

Principal components analysis is often confused with factor analysis, a related but a conceptually distinct technique. There is a considerable amount of confusion concerning the similarities and differences between the two techniques. This may be due to the fact that in many statistical packages (e.g., SPSS) principal components analysis is an option of the factor analysis procedure. This chapter focuses on principal components analysis; the next chapter discusses factor analysis and explains the differences between the two techniques. The following section provides a geometric view of principal components analysis. This is then followed by an algebraic explanation.

[1]The concept is similar to the use of Dow Jones Industrial Average for measuring stock market performance.

4.1 GEOMETRY OF PRINCIPAL COMPONENTS ANALYSIS

Table 4.1 presents a small data set consisting of 12 observations and 2 variables. The table also gives the mean-corrected data, the SSCP, the S (i.e., covariance), and the R (i.e., correlation) matrices. Figure 4.1 presents a plot of the mean-corrected data in the two-dimensional space. From Table 4.1 we can see that the variances of variables x_1 and x_2 are 23.091 and 21.091, respectively, and the total variance of the two variables is 44.182 (i.e., 23.091 + 21.091). Also, x_1 and x_2 are correlated, with the correlation coefficient being 0.746. The percentages of the total variance accounted for by x_1 and x_2 are, respectively, 52.26% and 47.74%.

4.1.1 Identification of Alternative Axes and Forming New Variables

As shown by the dotted line in Figure 4.1, let X_1^* be any axis in the two-dimensional space making an angle of θ degrees with X_1. The projection of the observations onto X_1^* will give the coordinate of the observations with respect to X_1^*. As discussed in Section 2.7 of Chapter 2, the coordinate of a point with respect to a new axis is a linear combination of the coordinates of the point with respect to the original set of axes. That

Table 4.1 **Original, Mean-Corrected, and Standardized Data**

	x_1		x_2	
Observation	**Original**	**Mean Corrected**	**Original**	**Mean Corrected**
1	16	8	8	5
2	12	4	10	7
3	13	5	6	3
4	11	3	2	−1
5	10	2	8	5
6	9	1	−1	−4
7	8	0	4	1
8	7	−1	6	3
9	5	−3	−3	−6
10	3	−5	−1	−4
11	2	−6	−3	−6
12	0	−8	0	−3
Mean	8	0	3	0
Variance	23.091	23.091	21.091	21.091

$$\mathbf{SSCP} = \begin{bmatrix} 254 & 181 \\ 181 & 232 \end{bmatrix}$$

$$\mathbf{S} = \begin{bmatrix} 23.091 & 16.455 \\ 16.455 & 21.091 \end{bmatrix}$$

$$\mathbf{R} = \begin{bmatrix} 1.000 & 0.746 \\ 0.746 & 1.000 \end{bmatrix}$$

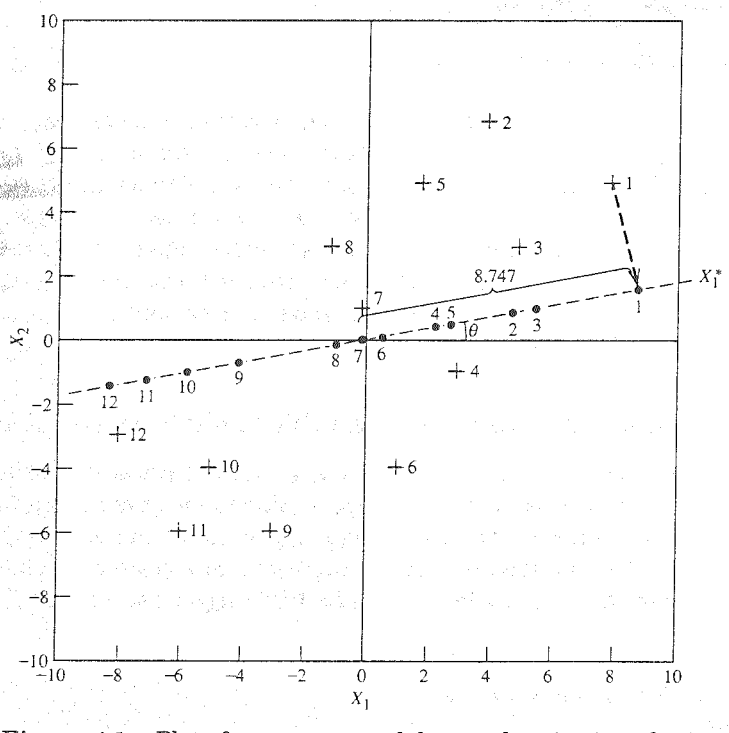


Figure 4.1 Plot of mean-corrected data and projection of points onto X_1^*.

is (see Eq. 2.24, Chapter 2),

$$x_1^* = \cos\theta \times x_1 + \sin\theta \times x_2,$$

where x_1^* is the coordinate of the observation with respect to X_1^*, and x_1 and x_2 are, respectively, coordinates of the observation with respect to X_1 and X_2. It is clear that x_1^*, which is a linear combination of the original variables, can be considered as a new variable.

For a given value of θ, say 10°, the equation for the linear combination is

$$x_1^* = 0.985x_1 + 0.174x_2,$$

which can be used to obtain the coordinates of the observations with respect to X_1^*. These coordinates are given in Figure 4.1 and Table 4.2. For example, in Figure 4.1 the coordinate for the first observation with respect to X_1^* is equal to 8.747. The coordinates or projections of the observations onto X_1^* can be viewed as the corresponding values for the new variable, x_1^*. That is, the value of the new variable for the first observation is 8.747. Table 4.2 also gives the mean and the variance for x_1^*. From the table we can see that (1) the new variable remains mean corrected (i.e., its mean is equal to zero); and (2) the variance of x_1^* is 28.659 accounting for 64.87% (28.659/44.182) of the total variance in the data. Note that the variance accounted for by x_1^* is greater than the variance accounted for by any one of the original variables.

Now suppose the angle between X_1^* and X_1 is, say, 20° instead of 10°. Obviously, one would obtain different values for x_1^*. Table 4.3 gives the percent of total variance accounted for by x_1^* when X_1^* makes different angles with X_1 (i.e., for different new axes). Figure 4.2 gives the plot of the percent variance accounted for by x_1^* and

Table 4.2 **Mean-Corrected Data and New Variable (x_1^*) for a Rotation of 10°**

	Mean-Corrected Data		
Observation	x_1	x_2	x_1^*
1	8	5	8.747
2	4	7	5.155
3	5	3	5.445
4	3	−1	2.781
5	2	5	2.838
6	1	−4	0.290
7	0	1	0.174
8	−1	3	−0.464
9	−3	−6	−3.996
10	−5	−4	−5.619
11	−6	−6	−6.951
12	−8	−3	−8.399
Mean	0.000	0.000	0.000
Variance	23.091	21.091	28.659

Table 4.3 **Variance Accounted for by the New Variable x_1^* for Various New Axes**

Angle with X_1 (θ)	Total Variance	Variance of x_1^*	Percent (%)
0	44.182	23.091	52.263
10	44.182	28.659	64.866
20	44.182	33.434	75.676
30	44.182	36.841	83.387
40	44.182	38.469	87.072
43.261	44.182	38.576	87.312
50	44.182	38.122	86.282
60	44.182	35.841	81.117
70	44.182	31.902	72.195
80	44.182	26.779	60.597
90	44.182	21.091	47.772

the angle between X_1^* and X_1. From the table and the figure, one can see that the percent of the total variance accounted for by x_1^* increases as the angle between X_1^* and X_1 increases and then, after a certain maximum value, the variance accounted for by x_1^* begins to decrease. That is, there is *one and only one* new axis that results in a new variable accounting for the maximum variance in the data. And this axis makes an angle of 43.261° with X_1. The corresponding equation for computing the values of x_1^* is

$$\begin{aligned} x_1^* &= \cos 43.261 \times x_1 + \sin 43.261 \times x_2 \\ &= 0.728x_1 + 0.685x_2. \end{aligned} \tag{4.1}$$

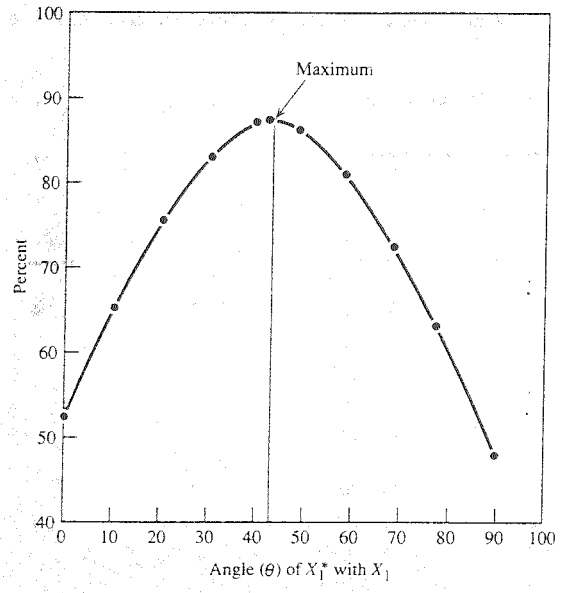


Figure 4.2 Percent of total variance accounted for by X_1^*.

Table 4.4 **Mean-Corrected Data, and x_1^* and x_2^* for the New Axes Making an Angle of 43.261°**

	Mean-Corrected Data		New Variables	
Observation	x_1	x_2	x_1^*	x_2^*
1	8	5	9.253	−1.841
2	4	7	7.710	2.356
3	5	3	5.697	−1.242
4	3	−1	1.499	−2.784
5	2	5	4.883	2.271
6	1	−4	−2.013	−3.598
7	0	1	0.685	0.728
8	−1	3	1.328	2.870
9	−3	−6	−6.297	−2.313
10	−5	−4	−6.382	0.514
11	−6	−6	−8.481	−0.257
12	−8	−3	−7.882	3.298
Mean	0.000	0.000	0.000	0.000
SS			424.334	61.666
Variance			38.576	5.606

SSCP, Covariance, and Correlation Matrices for the New Variables

$$\text{SSCP} = \begin{bmatrix} 424.334 & 0.000 \\ 0.000 & 61.666 \end{bmatrix}$$

$$\text{S} = \begin{bmatrix} 38.576 & 0.000 \\ 0.000 & 5.606 \end{bmatrix}$$

$$\text{R} = \begin{bmatrix} 1.000 & 0.000 \\ 0.000 & 1.000 \end{bmatrix}$$

Table 4.4 gives the values for x_1^* and its mean, SS, and variance. It can be seen that x_1^* accounts for about 87.31% (38.576/44.182) of the total variance in the data.

Note that x_1^* does not account for all the variance in the data. Therefore, it is possible to identify a second axis such that the corresponding second new variable accounts for the maximum of the variance that is not accounted for by x_1^*. Let X_2^* be the second new axis that is orthogonal to X_1^*. Thus, if the angle between X_1^* and X_1 is θ then the angle between X_2^* and X_2 will also be θ. The linear combination for forming x_2^* will be (see Eq. 2.25, Chapter 2)

$$x_2^* = -\sin\theta \times x_1 + \cos\theta \times x_2.$$

For $\theta = 43.261°$ the above equation becomes

$$x_2^* = -0.685x_1 + 0.728x_2. \tag{4.2}$$

Table 4.4 also gives the values of x_2^*, and its mean, SS, and variance, and the **SSCP**, **S**, and **R** matrices. Figure 4.3 gives the plot showing the observations and the new axes. The following observations can be made from the figure and the table:

1. The orientation or the configuration of the points or observations in the two-dimensional space does not change. The observations can, therefore, be represented with respect to the old or the new axes.
2. The projections of the points onto the original axes give the values for the original variables, and the projections of the points onto the new axes give the values for the new variables. The new axes or the variables are called *principal components* and the values of the new variables are called *principal components scores*.
3. Each of the new variables (i.e., x_1^* and x_2^*) are linear combinations of the original variables and remain mean corrected. That is, their means are zero.

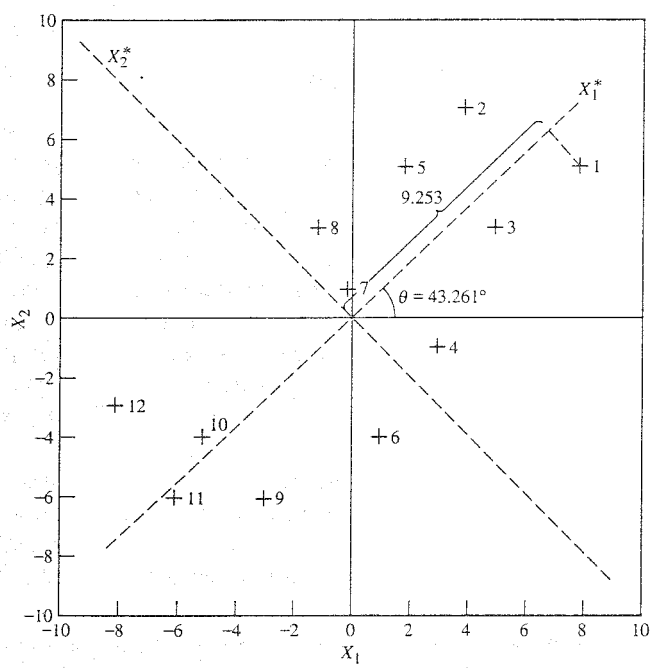


Figure 4.3 Plot of mean-corrected data and new axes

4. The total SS for x_1^* and x_2^* is 486 (i.e., 424.334 + 61.666) and is the same as the total SS for the original variables.
5. The variances of x_1^* and x_2^* are, respectively, 38.576 and 5.606. The total variance of the two variables is 44.182 (i.e., 38.576 + 5.606) and is the same as the total variance of x_1 and x_2. That is, the total variance of the data has not changed. Note that one would not expect the total variance (i.e., information) to change, as the orientation of the data points in the two-dimensional space has not changed.
6. The percentages of the total variance accounted for by x_1^* and x_2^* are, respectively, 87.31% (38.576/44.182) and 12.69% (5.606/44.182). The variance accounted for by the first new variable, x_1^*, is greater than the variance accounted for by any one of the original variables. The second new variable accounts for variance that has not been accounted for by the first new variable. The two new variables together account for all of the variance in the data.
7. The correlation between the two new variables is zero, i.e., x_1^* and x_2^* are uncorrelated.

The above geometrical illustration of principal components analysis can be easily extended to more than two variables. A data set consisting of p variables can be represented graphically in a p-dimensional space with respect to the original p axes or p new axes. The first new axis, X_1^*, results in a new variable, x_1^*, such that this new variable accounts for the maximum of the total variance. After this, a second axis, orthogonal to the first axis, is identified such that the corresponding new variable, x_2^*, accounts for the maximum of the variance that has not been accounted for by the first new variable, x_1^*, and x_1^* and x_2^* are uncorrelated. This procedure is carried on until all the p new axes have been identified such that the new variables, $x_1^*, x_2^*, \ldots, x_p^*$ account for successive maximum variances and the variables are uncorrelated.[2] Note that the maximum number of new variables (i.e., principal components) is equal to the number of original variables.

4.1.2 Principal Components Analysis as a Dimensional Reducing Technique

In the previous section it was seen that principal components analysis essentially reduces to identifying a new set of orthogonal axes. The principal components scores or the new variables were projections of points onto the axes. Now suppose that instead of using both of the original variables we use only one new variable, x_1^*, to represent most of the information contained in the data. Geometrically, this is equivalent to representing the data in a one-dimensional space. In the case of p variables one may want to represent the data in a lower m-dimensional space where m is much less than p. Representing data in a lower-dimensional space is referred to as *dimensional reduction*. Therefore, principal components analysis can also be viewed as a *dimensional reduction technique*.

The obvious question is: how well can the few new variable(s) represent the information contained in the data? Or geometrically, how well can we capture the configuration of the data in the reduced-dimensional space? Consider the plot of hypothetical data given in Panels I and II of Figure 4.4. Suppose we desire to represent the data in

[2]It should be noted that once the $p - 1$ axes have been identified, the identification of the pth axis will be fixed due to the condition that all the axes must be orthogonal.

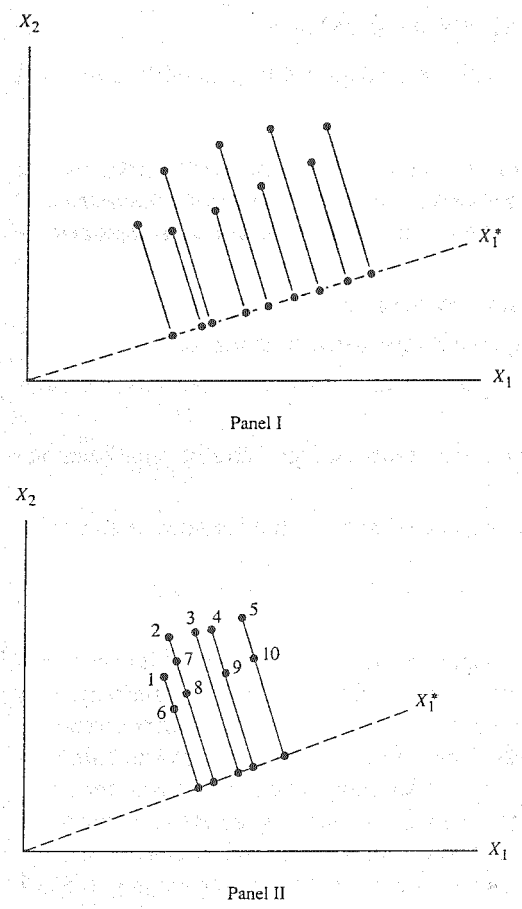


Figure 4.4 Representation of observations in lower-dimensional subspace.

only one dimension, given by the dotted axis representing the first principal component. As can be clearly seen, the one-dimensional representation of points in Panel I is much better than that of Panel II.[3] For example, in Panel II points 1 and 6; 2, 7, and 8; 4 and 9; and 5 and 10 cannot be distinguished from one another. In other words, the configuration of the observations in the one-dimensional subspace is much better in Panel I than in Panel II. Or, we can say that the data of Panel I can be represented by one variable with less loss of information as compared to the data set of Panel II.

Typically the sum of the variances of the new variables not used to represent the data is used as the measure for the loss of information resulting from representing the data in a lower-dimensional space. For example, if in Table 4.4 only x_1^* is used then the loss of information is the variance accounted for by the second variable (i.e., x_2^*) which is 12.69% (5.606/41.182) of the total variance. Whether this loss is substantial or not depends on the purpose or objective of the study. This point is discussed further in the later sections of the chapter.

[3]The representation of the observations in a space of a given dimension is obtained by the projection of the points onto the space. A one-dimensional space is a line, a two-dimensional space is a plane, a three-dimensional space is a hyperplane, and so on. For example, for a one-dimensional subspace the representation is obtained by projecting the observations onto the line representing the dimension. See the discussion in Chapter 2 regarding the projection of vectors onto subspaces.

4.1.3 Objectives of Principal Components Analysis

Geometrically, the objective of principal components analysis is to identify a new set of orthogonal axes such that:

1. The coordinates of the observations with respect to each of the axes give the values for the new variables. As mentioned previously, the new axes or the variables are called *principal components* and the values of the new variables are called *principal components scores*.
2. Each new variable is a linear combination of the original variables.
3. The first new variable accounts for the maximum variance in the data.
4. The second new variable accounts for the maximum variance that has not been accounted for by the first variable.
5. The third new variable accounts for the maximum variance that has not been accounted for by the first two variables.
6. The pth new variable accounts for the variance that has not been accounted for by the $p-1$ variables.
7. The p new variables are uncorrelated.

Now if a substantial amount of the total variance in the data is accounted for by a few (preferably far fewer) principal components or new variables, then the researcher can use these few principal components for interpretational purposes or in further analysis of the data instead of the original p variables. This would result in a substantial amount of data reduction if the value of p is large. Note that data reduction is not in terms of how much data has to be collected, as all the original p variables are needed to form the principal components scores; rather it is in terms of how many new variables are retained for further analysis. Hence, principal components analysis is commonly referred to as a data-reduction technique.

4.2 ANALYTICAL APPROACH

The preceding section provides a geometric view of principal components analysis. This section presents the algebraic approach to principal components analysis. The Appendix gives the mathematics of principal components analysis.

We can now formally state the objective of principal components analysis. Assuming that there are p variables, we are interested in forming the following p linear combinations:

$$\begin{aligned}
\xi_1 &= w_{11}x_1 + w_{12}x_2 + \cdots + w_{1p}x_p \\
\xi_2 &= w_{21}x_1 + w_{22}x_2 + \cdots + w_{2p}x_p \\
&\vdots \\
\xi_p &= w_{p1}x_1 + w_{p2}x_2 + \cdots + w_{pp}x_p.
\end{aligned} \tag{4.3}$$

where $\xi_1, \xi_2, \ldots, \xi_p$ are the p principal components and w_{ij} is the weight of the jth variable for the ith principal component.[4] The weights, w_{ij}, are estimated such that:

[4]To be consistent with the standard notation used in most statistics textbooks, the new variables or the principal components are denoted by Greek letters.

1. The first principal component, ξ_1, accounts for the maximum variance in the data, the second principal component, ξ_2, accounts for the maximum variance that has not been accounted for by the first principal component, and so on.
2. $$w_{i1}^2 + w_{i2}^2 + \cdots + w_{ip}^2 = 1 \qquad i = 1, \ldots, p \tag{4.4}$$
3. $$w_{i1}w_{j1} + w_{i2}w_{j2} + \cdots + w_{ip}w_{jp} = 0 \qquad \text{for all } i \neq j. \tag{4.5}$$

The condition given by Eq. 4.4 requires that the squares of the weights sum to one and is somewhat arbitrary. This condition is used to fix the scale of the new variables and is necessary because it is possible to increase the variance of a linear combination by changing the scale of the weights.[5] The condition given by Eq. 4.5 ensures that the new axes are orthogonal to each other. The mathematical problem is: how do we obtain the weights of Eq. 4.3 such that the conditions specified above are satisfied? This is essentially a calculus problem, the details of which are provided in the Appendix.

4.3 HOW TO PERFORM PRINCIPAL COMPONENTS ANALYSIS

A number of computer programs are available for performing principal components analysis. The two most widely used statistical packages are the Statistical Analysis System (SAS), and the Statistical Package for the Social Sciences (SPSS). In the following section we discuss the outputs obtained from SAS. The output from SPSS is very similar, and the reader is encouraged to obtain the corresponding output from SPSS and compare it with the SAS output. The data set of Table 4.1 is used to discuss the output obtained from SAS.

4.3.1 SAS Commands and Options

Table 4.5 gives the SAS commands necessary for performing principal components analysis. The PROC PRINCOMP command invokes the principal components analysis

Table 4.5 **SAS Statements**

```
DATA ONE;
TITLE PRINCIPAL COMPONENTS ANALYSIS FOR DATA OF TABLE 4.1;
INPUT X1 X2;
CARDS;
insert data here
PROC PRINCOMP COV OUT=NEW;
VAR X1 X2;
PROC PRINT;
VAR X1 X2 PRIN1 PRIN2;
PROC CORR;
VAR X1 X2 PRIN1 PRIN2;
```

[5]For example, one can increase the variance accounted for in the first principal component by a factor of 4 by assuming that $w_{11} = 2w_{11}, w_{12} = 2w_{12}$, and so on.

procedure. It has a number of options. Principal components analysis can be performed either on mean-corrected or standardized data. Each of these data sets can result in a different solution, which implies that the solution is not scale invariant. The solution depends upon the relative variances of the variables. A detailed discussion of the effects of standardization on principal components analysis results is provided later in the chapter. The COV option requests that mean-corrected data should be used. In other words, the covariance matrix will be used to estimate the weights of the linear combinations. The OUT= option is used to specify the name of a data set in which the original and the new variables are saved. The name of the data set specified is NEW. The PROC PRINT procedure gives a printout of the original and the new variables and the PROC CORR procedure gives means, standard deviations, and the correlation of the new and original variables.

4.3.2 Interpreting Principal Components Analysis Output

Exhibit 4.1 gives the resulting output. Following is a discussion of the various sections of the output. The numbers in square brackets correspond to the circled numbers in the exhibit. For convenience, the values from the exhibit reported in the text are rounded to three significant digits. Any discrepancies between the numbers reported in the text and the output are due to rounding errors.

Descriptive Statistics

This part of the output gives the basic descriptive statistics such as the mean and the standard deviation of the original variables. As can be seen, the means of the variables are 8.00 and 3.000 and the standard deviations are 4.805 and 4.592 [1]. The output also gives the covariance matrix [2]. From the covariance matrix, it can be seen that the total variance is 44.182, with x_1 accounting for almost 52.26% (i.e., 23.091/44.182) of the total variance in the data set. The covariance between the two variables can be converted to the correlation coefficient by dividing the covariance by the product of the respective standard deviations. The correlation between the two variables is 0.746 (i.e., correlation $= 16.455/(4.805 \times 4.592) = .746$).

Principal Components

The eigenvectors give the weights that are used for forming the equation (i.e., the principal component) to compute the new variables [3b]. The name, *eigenvector,* for the principal component is derived from the analytical procedure used for estimating the weights.[6] Therefore, the two new variables are:

$$\xi_1 = Prin1 = 0.728x_1 + 0.685x_2 \tag{4.6}$$

$$\xi_2 = Prin2 = -0.685x_1 + 0.728x_2 \tag{4.7}$$

where $Prin1$ and $Prin2$ are the new variables or linear combinations and x_1 and x_2 are the original mean-corrected variables. In principal components analysis terminology, $Prin1$ and $Prin2$ are normally referred to as principal components. Note that Eqs. 4.6 and 4.7 are the same as Eqs. 4.1 and 4.2. As can be seen, the sum of the squared weights of each principal component is one (i.e., $0.728^2 + 0.685^2 = 1$ and

[6] As discussed in the Appendix, the solution to principal components analysis is obtained by computing the eigenvalues and eigenvectors of the covariance matrix. The eigenvectors give the weights that can be used to form the new variables and the eigenvalues give the variances of the new variables.

Exhibit 4.1

Principal components analysis for data in Table 4.1

1 SIMPLE STATISTICS

	X1	X2
MEAN	8.00000	3.00000
ST DEV	4.80530	4.59248

2 COVARIANCES

	X1	X2
X1	23.09091	16.45455
X2	16.45455	21.09091

TOTAL VARIANCE=44.18182

3a

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	38.5758	32.9698	0.873115	0.87312
PRIN2	5.6060	.	0.126885	1.00000

3b EIGENVECTORS

	PRIN1	PRIN2
X1	0.728238	-.685324
X2	0.685324	0.728238

4

VARIABLE	N	MEAN	STD DEV
X1	12	8.000000	4.805300
X2	12	3.000000	4.592484
PRIN1	12	-8.697E-16	6.210943
PRIN2	12	0.000000	2.367700

5 PEARSON CORRELATION COEFFICIENTS

	X1	X2	PRIN1	PRIN2
X1	1.00000	0.74562	0.94126	-0.33768
X2	0.74562	1.00000	0.92684	0.37545
PRIN1	0.94126	0.92684	1.00000	0.00000
PRIN2	-0.33768	0.37545	0.00000	1.00000

6

OBS	X1	X2	PRIN1	PRIN2
1	16	8	9.2525	-1.8414
2	12	10	7.7102	2.3564
3	13	6	5.6972	-1.2419
4	11	2	1.4994	-2.7842
5	10	8	4.8831	2.2705
6	9	-1	-2.0131	-3.5983
7	8	4	0.6853	0.7282
8	7	6	1.3277	2.8700
9	5	-3	-6.2967	-2.3135
10	3	-1	-6.3825	0.5137
11	2	-3	-8.4814	-0.2575
12	0	0	-7.8819	3.2979

$(-0.685)^2 + 0.728^2 = 1$) and the sum of the cross products of the weights is equal to zero (i.e., $0.728 \times -(0.685) + 0.685 \times 0.728$).

Principal Components Scores

This part of the output gives the original variables and the principal components scores, obtained by using Eqs. 4.6 and 4.7 [6]. For example, principal components scores, *Prin*1 and *Prin*2, for the first observation are, respectively, 9.249 (i.e., $.728 \times (16 - 8) + .685 \times (8 - 3)$) and -1.840 (i.e., $-.685 \times (16 - 8) + .728 \times (8 - 3)$). Note that the principal components scores reported are the same as those in Table 4.4.

The standard deviations of *Prin*1 and *Prin*2 are 6.211 and 2.368, respectively [4]. Consequently, the variances accounted for by each principal component are, respectively, 38.576 (i.e., 6.211^2) and 5.606 (i.e., 2.368^2). The means of the principal components, within rounding error, are zero as these are linear combinations of mean-corrected data [4]. The *eigenvalues* reported in the output are the same as the variance accounted for by each new variable (i.e., principal component) [3a]. The total variance of the new variables is 44.182 which is the same as the original variables. However, the variance accounted for by the first new variable, *Prin*1, is 87.31% (i.e., 38.576/44.182) which is given in the proportion column. Thus, if we were to use only the first new variable, instead of the two original variables, we would be able to account for almost 87% of the variance of the original data.

Sometimes the principal components scores, *Prin*1 and *Prin*2, are standardized to a mean of zero and a standard deviation of one. Table 4.6 gives the standardized scores which can be obtained by dividing the principal components scores by the respective standard deviations. Or you could instruct SAS to report the standardized scores by changing the PROC PRINCOMP statement to PROC PRINCOMP COV STD OUT=NEW. Note that this command still requests principal components analysis on mean-corrected data. The only difference is that the STD option requests standardization of the principal components scores.

Loadings

This part of the output reports the correlation among the variables [5]. The correlation between the new variables, *Prin*1 and *Prin*2, is zero, implying that they are not correlated. The simple correlations between the original and the new variables, also called *loadings*, give an indication of the extent to which the original variables are influential or important in forming new variables. That is, the higher the loading the more influential the variable is in forming the principal components score and vice versa. For example, high correlations of 0.941 and 0.927 between *Prin*1 and x_1 and x_2, respectively, indicate that x_1 and x_2 are very influential in forming *Prin*1. As will be discussed later in the chapter, the loadings can be used to interpret the meaning of the principal components or the new variables. The loadings can also be obtained by using

Table 4.6 **Standardized Principal Components Scores**

Observation	*Prin*1	*Prin*2
1	1.490	−0.778
2	1.241	0.995
3	0.917	−0.525
4	0.241	−1.176
5	0.786	0.959
6	−0.324	−1.520
7	0.110	0.308
8	0.214	1.212
9	−1.014	−0.977
10	−1.028	0.217
11	−1.366	−0.109
12	−1.269	1.393

the following equation:

$$l_{ij} = \frac{w_{ij}}{\hat{s}_j}\sqrt{\lambda_i} \tag{4.8}$$

where l_{ij} is the loading of the jth variable for the ith principal components, w_{ij} is the weight of the jth variable for the ith principal components, λ_i is the eigenvalue (i.e., the variance) of the ith principal components, and $\hat{s}_j$ is the standard deviation of the jth variable.

4.4 ISSUES RELATING TO THE USE OF PRINCIPAL COMPONENTS ANALYSIS

We have seen that principal components analysis is the formation of new variables that are linear combinations of the original variables. However, as a data analytic technique, the use of principal components analysis raises a number of issues that need to be addressed. These issues are:

1. What effect does the type of data (i.e., mean-corrected or standardized data) have on principal components analysis?

Table 4.7 **Food Price Data**

	Average Price (in cents per pound)				
City	Bread	Burger	Milk	Oranges	Tomatoes
Atlanta	24.5	94.5	73.9	80.1	41.6
Baltimore	26.5	91.0	67.5	74.6	53.3
Boston	29.7	100.8	61.4	104.0	59.6
Buffalo	22.8	86.6	65.3	118.4	51.2
Chicago	26.7	86.7	62.7	105.9	51.2
Cincinnati	25.3	102.5	63.3	99.3	45.6
Cleveland	22.8	88.8	52.4	110.9	46.8
Dallas	23.3	85.5	62.5	117.9	41.8
Detroit	24.1	93.7	51.5	109.7	52.4
Honolulu	29.3	105.9	80.2	133.2	61.7
Houston	22.3	83.6	67.8	108.6	42.4
Kansas City	26.1	88.9	65.4	100.9	43.2
Los Angeles	26.9	89.3	56.2	82.7	38.4
Milwaukee	20.3	89.6	53.8	111.8	53.9
Minneapolis	24.6	92.2	51.9	106.0	50.7
New York	30.8	110.7	66.0	107.3	62.6
Philadelphia	24.5	92.3	66.7	98.0	61.7
Pittsburgh	26.2	95.4	60.2	117.1	49.3
St. Louis	26.5	92.4	60.8	115.1	46.2
San Diego	25.5	83.7	57.0	92.8	35.4
San Francisco	26.3	87.1	58.3	101.8	41.5
Seattle	22.5	77.7	62.0	91.1	44.9
Washington, DC	24.2	93.8	66.0	81.6	46.2

Source: Estimated Retail Food Prices by Cities, March 1973, U.S. Department of Labor, Bureau of Labor Statistics, pp. 1–8.

2. Is principal components analysis the appropriate technique for forming the new variables? That is, what additional insights or parsimony is achieved by subjecting the data to a principal components analysis?
3. How many principal components should be retained? That is, how many new variables should be used for further analysis or interpretation?
4. How do we interpret the principal components (i.e., the new variables)?
5. How can principal components scores be used in further analyses?

These issues will be discussed using the data in Table 4.7, which presents prices of food items in 23 cities. It should be noted that the preceding issues also suggest a procedure that one can follow to analyze data using principal components analysis.

4.4.1 Effect of Type of Data on Principal Components Analysis

Principal components analysis can be either done on mean-corrected or standardized data. Each data set could give a different solution depending upon the extent to which the variances of the variables differ. In other words, variances of the variables could have an effect on principal components analysis.

Assume that the main objective for the data given in Table 4.7 is to form a measure of the Consumer Price Index (*CPI*). That is, we would like to form a weighted sum of the various food prices that would summarize how expensive or cheap are a given city's food items. Principal components analysis would be an appropriate technique for developing such an index. Exhibit 4.2 gives the partial output obtained when the principal components procedure in SAS was applied to the mean-corrected data. The variances of the five food items are as follows [1]:

Food Item	Variance	Percent of Total Variance
Bread	6.284	1.688
Hamburger	57.077	15.334
Milk	48.306	12.978
Oranges	202.756	54.472
Tomatoes	57.801	15.528
Total	372.224	100.000

As can be seen, the price of oranges accounts for a substantial portion (almost 55%) of the total variance. Since there are five variables, a total of five principal components can be extracted. Let us assume that only one principal component is retained, and it is used as a measure of *CPI*.[7] Then, from the eigenvector, the first principal component, $Prin1$, is given by [2b]:

$$Prin1 = 0.028 * Bread + 0.200 * Burger + 0.042 * Milk + 0.939 * Oranges + 0.276 * Tomatoes, \quad (4.9)$$

and the eigenvalue indicates that the variance of $Prin1$ is 218.999, accounting for 58.84% of the total variance of the original data [2a]. Equation 4.9 indicates that the value of $Prin1$, though a weighted sum of *all the food prices*, is very much affected by

[7]The issue pertaining to the number of principal components to retain is discussed later.

Exhibit 4.2

Principal components analysis for data in Table 4.7

Simple Statistics

	BREAD	BURGER	MILK	ORANGES	TOMATOES
Mean	25.29130435	91.85652174	62.29565217	102.9913043	48.76521739
StD	2.50688380	7.55493975	6.95024383	14.2392515	7.60266752

1 Covariance Matrix

	BREAD	BURGER	MILK	ORANGES	TOMATOES
BREAD	6.2844664	12.9109684	5.7190514	1.3103755	7.2851383
BURGER	12.9109684	57.0771146	17.5075296	22.6918775	36.2947826
MILK	5.7190514	17.5075296	48.3058893	-0.2750395	13.4434783
ORANGES	1.3103755	22.6918775	-0.2750395	202.7562846	38.7624111
TOMATOES	7.2851383	36.2947826	13.4434783	38.7624111	57.8005534

2a Total Variance = 372.2243083

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	218.999	127.276	0.588351	0.58835
PRIN2	91.723	54.060	0.246419	0.83477
PRIN3	37.663	16.852	0.101183	0.93595
PRIN4	20.811	17.781	0.055909	0.99186
PRIN5	3.029	.	0.008138	1.00000

2b Eigenvectors

	PRIN1	PRIN2
BREAD	0.028489	0.165321
BURGER	0.200122	0.632185
MILK	0.041672	0.442150
ORANGES	0.938859	-.314355
TOMATOES	0.275584	0.527916

3

OBS	CITY	PRIN1	PRIN2
1	BALTIMORE	-25.3258	13.2784
2	LOS ANGELES	-22.6270	-3.1387
3	ATLANTA	-22.4763	10.0846
	.	.	.
21	PITTSBURGH	14.0411	-2.6890
22	BUFFALO	14.1399	-5.9650
23	HONOLULU	35.5971	14.7894

4 Pearson Correlation Coefficients

	BREAD	BURGER	MILK	ORANGES	TOMATOES
PRIN1	0.16818	0.39200	0.08873	0.97574	0.53642
PRIN2	0.63159	0.80141	0.60927	-0.21143	0.66503

the price of oranges. Values of *Prin*1 suggest that Honolulu is the most expensive city and Baltimore is the least expensive city [3].[8]

The main reason the price of oranges dominates the formation of *Prin*1 is that there exists a wide variation in the price of oranges across the cities (i.e., the variance of the price for oranges is very high compared to the variances of the prices of other food items).

[8]Note that the principal components scores are mean corrected, and since all the weights are positive a high score will imply that the food prices are high and vice versa.

In general, the weight assigned to a variable is affected by the relative variance of the variable. If we do not want the relative variance to affect the weights, then the data should be standardized so that the variance of each variable is the same (i.e., one). Exhibit 4.3 gives the SAS output for standardized data. Since the data are standardized, the variance of each variable is one and each variable accounts for 20% of the total variance. The first principal component, $Prin1$, accounts for 48.44% (i.e., 2.422/5) of the total variance [1], and as per the eigenvectors it is given as [2][9]

$$Prin1 = 0.496 * Bread + 0.576 * Burger + 0.340 * Milk + 0.225 * Oranges + 0.506 * Tomatoes. \quad (4.10)$$

We can see that the first principal component, $Prin1$, is a weighted sum of all the food prices and no one food item dominates the formation of the score. The value of $Prin1$ suggests that Honolulu is the most expensive city and the least expensive city now is Seattle, as compared to Baltimore when the data were not standardized [3]. Therefore, the weights that are used to form the index (i.e., the principal component) are affected by the relative variances of the variables.

Exhibit 4.3

Principal components analysis on standardized data

Correlation Matrix

	BREAD	BURGER	MILK	ORANGES	TOMATOES
BREAD	1.0000	0.6817	0.3282	0.0367	0.3822
BURGER	0.6817	1.0000	0.3334	0.2109	0.6319
MILK	0.3282	0.3334	1.0000	-.0028	0.2544
ORANGES	0.0367	0.2109	-.0028	1.0000	0.3581
TOMATOES	0.3822	0.6319	0.2544	0.3581	1.0000

1 Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
PRIN1	2.42247	1.31779	0.484494	0.48449
PRIN2	1.10467	0.36619	0.220935	0.70543
PRIN3	0.73848	0.24487	0.147696	0.85312
PRIN4	0.49361	0.25285	0.098722	0.95185
PRIN5	0.24077	.	0.048153	1.00000

2 Eigenvectors

	PRIN1	PRIN2
BREAD	0.496149	-.308620
BURGER	0.575702	-.043802
MILK	0.339570	-.430809
ORANGES	0.224990	0.796777
TOMATOES	0.506434	0.287028

3

OBS	CITY	PRIN1	PRIN2
1	SEATTLE	-2.09100	-0.36728
2	SAN DIEGO	-1.89029	-0.72501
3	HOUSTON	-1.28764	0.14847
.	.	.	.
21	BOSTON	2.24797	-0.07359
22	NEW YORK	3.69680	-0.25362
23	HONOLULU	4.07722	0.49398

4 Pearson Correlation Coefficients

	BREAD	BURGER	MILK	ORANGES	TOMATOES
PRIN1	0.77222	0.89604	0.52852	0.35018	0.78823
PRIN2	-0.32437	-0.04604	-0.45280	0.83744	0.30168

[9]Since the variables are standardized, standardized prices should be used for forming the principal components score.

The choice between the analysis obtained from mean-corrected and standardized data also depends on other factors. For example, in the present situation there is no compelling reason to believe that any one food item is more important than the other food items comprising a person's diet. Consequently, in formulating the index, the price of oranges should not receive an artificially higher weight due to the variation in its prices. Therefore, given the objective, standardized data should be used. In cases for which there is reason to believe that the variances of the variables do indicate the importance of a given variable, then mean-corrected data should be used. Since it is more appropriate to use standardized data for forming the *CPI*, all subsequent discussions will be for standardized data.

۱- هدف تشکیل ترکیبهای ناهمبسته یا متغیرهای باشد

۲- هدف کاهش متغیرها باید با تعداد کم PC

کاهش اطلاعات ناچیز باشد

4.4.2 Is Principal Components Analysis the Appropriate Technique?

Whether the data should or should not be subjected to principal components analysis primarily depends on the objective of the study. If the objective is to form uncorrelated linear combinations then the decision will depend on the interpretability of the resulting principal components. If the principal components cannot be interpreted then their subsequent use in other statistical techniques may not be very meaningful. In such a case one should avoid principal components analysis for forming uncorrelated variables.

On the other hand, if the objective is to reduce the number of variables in the data set to a few variables (principal components) that are linear combinations of the original variables, then it is imperative that the number of principal components be less than the number of original variables. In such a case principal components analysis should only be performed if the data can be represented by a *fewer* number of principal components *without a substantial loss of information.* But what do we mean by *without a substantial loss of information?* A geometric view of this notion was provided in Section 4.1.2 where it was mentioned that the notion of *substantial loss of information* depends on the purpose for which the principal components will be used.

Consider the case where scientists have available a total of 100 variables or pieces of information for making a launch decision for the space shuttle. It is found that five principal components account for 99% of all the variation in the 100 variables. However, in this case the scientists may consider the 1% of unaccounted variation (i.e., loss of information) as substantial, and thus the scientists may want to use *all* the variables for making a decision. In this case, the data cannot be represented in a reduced-dimensional space. On the other hand, if the 100 variables are prices of various food items then the five principal components accounting for 99% of the variance may be considered as very good because the 1% of unaccounted variation may not be substantial.

Is principal components analysis an appropriate technique for the data set given in Table 4.7? Keep in mind that the objective is to form consumer price indices. That is, the objective is data reduction. From Exhibit 4.3, the first two principal components, $Prin1$ and $Prin2$, account for about 71% of the total variance [1]. If we are willing to sacrifice 29% of the variance in the original data then we can use the first two principal components, instead of the original five variables, to represent the data set. In this case principal components analysis would be an appropriate technique. Note that we are using the amount of unexplained variance as a measure for *loss of information.*

There will be instances where it may not be possible to explain a substantial portion of the variance by only a few new variables. In such cases we may have to use the same number of principal components as the number of variables to account for a significant amount of variation. This normally happens when the variables are not correlated among themselves. For example, if the variables are orthogonal then each

principal component will account for the same amount of variance. In this case we have not really achieved any data reduction. On the other hand, if the variables are perfectly correlated among themselves then the first principal component will account for all of the variance in the data. That is, the greater the correlation among the variables the greater the data reduction we can achieve and vice versa.

This discussion suggests that principal components analysis is most appropriate if the variables are interrelated, for only then will it be possible to reduce a number of variables to a manageable few without much loss of information. If we cannot achieve the above objective, then principal components analysis may not be an appropriate technique. Formal statistical tests are available for determining if the variables are significantly correlated among themselves. The choice of test depends on the type of data that is used (i.e., mean-corrected or standardized data). Bartlett's test is one such test that can be used for standardized data. However, the tests, including the Bartlett's test, are sensitive to sample sizes in that for large sample sizes even small correlations are statistically significant. Therefore, the tests are not that useful in a practical sense and will not be discussed. For discussion of these tests see Green (1978) and Dillon and Goldstein (1984). In practice, researchers have used their own judgment in determining whether a "few" principal components have accounted for a "substantial" portion of the information or variance.

4.4.3 Number of Principal Components to Extract

Once it has been decided that performing principal components analysis is appropriate, the next obvious issue is determining the number of principal components that should be retained. As discussed earlier, the decision is dependent on how much information (i.e., unaccounted variance) one is willing to sacrifice, which, of course, is a judgmental question. Following are some of the suggested rules:

1. In the case of standardized data, retain only those components whose eigenvalues are greater than one. This is referred to as the eigenvalue-greater-than-one rule.
2. Plot the percent of variance accounted for by each principal component and look for an elbow. The plot is referred to as the scree plot. This rule can be used for both mean-corrected and standardized data.
3. Retain only those components that are statistically significant.

The eigenvalue-greater-than-one rule is the default option in most of the statistical packages, including **SAS** and **SPSS**. The rationale for this rule is that for standardized data the amount of variance extracted by each component should, at a minimum, be equal to the variance of at least one variable. For the data in Table 4.7, this rule suggests that two principal components should be retained as the eigenvalues of the first two components are greater than one [Exhibit 4.3: 1]. It should be noted that Cliff (1988) has shown that the eigenvalue-greater-than-one rule is flawed in the sense that, depending on various conditions, this heuristic or rule may lead to a greater or fewer number of retained principal components than are necessary and, therefore, should not be used blindly. It should be used in conjunction with other rules or heuristics.

The scree plot, proposed by Cattell (1966), is very popular. In this rule a plot of the eigenvalues against the number of components is examined for an "elbow." The number of principal components that need to be retained is given by the elbow. Panel I of Figure 4.5 gives the scree plot for the principal components solution using standardized data. From the figure it appears that two principal components should be extracted

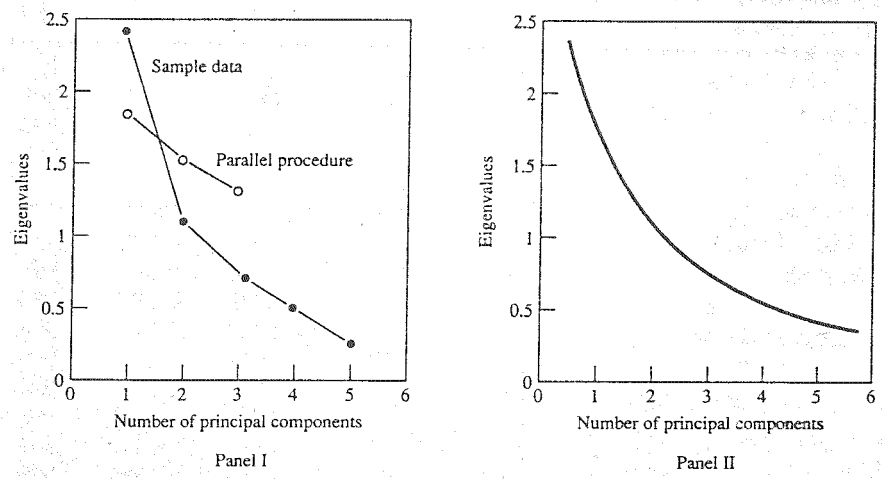


Figure 4.5 Scree plots. Panel I, Scree plot and plot of eigenvalues from parallel analysis. Panel II, Scree plot with no apparent elbow.

as that is where the elbow appears to be. It is obvious that a considerable amount of subjectivity is involved in identifying the elbow. In fact, in many instances the scree plot may be so smooth that it may be impossible to determine the elbow (see Panel II, Figure 4.5).

Horn (1965) has suggested a procedure, called parallel analysis, for overcoming the above difficulty when standardized data are used. Suppose we have a data set which consists of 400 observations and 20 variables. First, k multivariate normal random samples each consisting of 400 observations and 20 variables will be generated from an identity population correlation matrix.[10] The resulting data are subjected to principal components analysis. Since the variables are not correlated, each principal component would be expected to have an eigenvalue of 1.0. However, due to sampling error some eigenvalues will be greater than one and some will be less than one. Specifically, the first $p/2$ principal components will have an eigenvalue greater than one and the second set of $p/2$ principal components will have an eigenvalue of less than one. The average eigenvalues for each component over the k samples is plotted on the same graph containing the scree plot of the actual data. The cutoff point is assumed to be where the two graphs intersect.

It is, however, not necessary to run the simulation studies described above for standardized data.[11] Recently, Allen and Hubbard (1986) have developed the following regression equation to estimate the eigenvalues for random data for standardized data input:

$$\ln \lambda_k = a_k + b_k \ln(n-1) + c_k \ln\{(p-k-1)(p-k+2)/2\} + d_k \ln(\lambda_{k-1}) \quad (4.11)$$

where λ_k is the estimate for the kth eigenvalue, p is the number of variables, n is the number of observations, a_k, b_k, c_k, and d_k are regression coefficients, and $\ln \lambda_0$ is assumed to be 1. Table 4.8 gives the regression coefficients estimated using simulated data. Note from Eq. 4.11 that the last two eigenvalues cannot be estimated because the third term results in the logarithm of a zero or a negative value, which is undefined. However, this limitation does not hold for $p > 43$, for from Table 4.8 it can be seen that

[10] An identity correlation matrix represents the case where the variables are not correlated among themselves.

[11] For unstandardized data (i.e., covariance matrix) the above cumbersome procedure would have to be used.

Table 4.8 **Regression Coefficients for the Principal Components**

Root (k)	Number of Points[a]	a	b	c	d	R^2
1	62	.9794	−.2059	.1226	0.0000	.931
2	62	−.3781	.0461	.0040	1.0578	.998
3	62	−.3306	.0424	.0003	1.0805	.998
4	55	−.2795	.0364	−.0003	1.0714	.998
5	55	−.2670	.0360	−.0024	1.0899	.998
6	55	−.2632	.0368	−.0040	1.1039	.998
7	55	−.2580	.0360	−.0039	1.1173	.998
8	55	−.2544	.0373	−.0064	1.1421	.998
9	48	−.2111	.0329	−.0079	1.1229	.998
10	48	−.1964	.0310	−.0083	1.1320	.998
11	48	−.1858	.0288	−.0073	1.1284	.999
12	48	−.1701	.0276	−.0090	1.1534	.998
13	48	−.1697	.0266	−.0075	1.1632	.998
14	41	−.1226	.0229	−.0113	1.1462	.999
15	41	−.1005	.0212	−.0133	1.1668	.999
16	41	−.1079	.0193	−.0088	1.1374	.999
17	41	−.0866	.0177	−.0110	1.1718	.999
18	41	−.0743	.0139	−.0081	1.1571	.999
19	34	−.0910	.0152	−.0056	1.0934	.999
20	34	−.0879	.0145	−.0051	1.1005	.999
21	34	−.0666	.0118	−.0056	1.1111	.999+
22	34	−.0865	.0124	−.0022	1.0990	.999+
23	34	−.0919	.0123	−.0009	1.0831	.999+
24	29	−.0838	.0116	−.0016	1.0835	.999+
25	28	−.0392	.0083	−.0053	1.1109	.999+
26	28	−.0338	.0065	−.0039	1.1091	.999+
27	28	.0057	.0015	−.0049	1.1276	.999+
28	28	.0017	.0011	−.0034	1.1185	.999+
29	22	−.0214	.0048	−.0041	1.0915	.999+
30	22	−.0364	.0063	−.0030	1.0875	.999+
31	22	−.0041	.0022	−.0033	1.0991	.999+
32	22	.0598	−.0067	−.0032	1.1307	.999+
33	21	.0534	−.0062	−.0023	1.1238	.999+
34	16	.0301	−.0032	−.0027	1.0978	.999+
35	16	.0071	.0009	−.0038	1.0895	.999+
36	16	.0521	−.0052	−.0030	1.1095	.999+
37	16	.0824	−.0105	−.0014	1.1209	.999+
38	16	.1865	−.0235	−.0033	1.1567	.999+
39	10	.0075	.0009	−.0039	1.0773	.999+
40	10	.0050	−.0021	.0025	1.0802	.999+
41	10	.0695	−.0087	−.0016	1.0978	.999+
42	10	.0686	−.0086	−.0003	1.1004	.999+
43	10	.1370	−.0181	.0012	1.1291	.999+
44	10	.1936	−.0264	.0000	1.1315	.999+
45	10	.3493	−.0470	.0000	1.1814	.999
46	5	.1444	−.0185	.0000	1.1188	.999+
47	5	.0550	−.0067	.0000	1.0902	.999+
48	5	.1417	−.0189	.0000	1.1079	.999+

[a]The number of points used in the regression.

Source: Allen, S. J. and R. Hubbard (1986), "Regression Equations for the Latent Roots of Random Data Correlation Matrices with Unities on the Diagonal," *Multivariate Behavioral Research*, (21) 393–398.

$c_k = 0$ and, consequently, the third term is not necessary for estimating the eigenvalue. Using Eq. 4.11 and the coefficients from Table 4.8, the estimated value for $\ln \lambda_1$ is equal to

$$\begin{aligned} \ln \lambda_1 &= .9794 - .2059 \ln(23 - 1) + .1226 \ln\{(5 - 1 - 1)(5 - 1 + 2)/2\} \\ &= 0.61233 \end{aligned}$$

and, therefore, $\lambda_1 = 1.845$. Similarly, the reader can verify that the estimated values for λ_2 and λ_3 are, respectively, equal to 1.520 and 1.288. Figure 4.5 also shows the resulting plot. From the figure we can see that two principal components should be retained.

A statistical test that determines the statistical significance of the various principal components has been proposed. The test is a variation of the Bartlett's test used to determine if the correlations among the variables are significant. Consequently, the test has the same limitations—that is, it is very sensitive to sample sizes—and hence is very rarely used in practice.[12]

In practice, the most widely used procedures are the scree plot test, Horn's parallel procedure, and the rule of retaining only those components whose eigenvalues are greater than one. Simulation studies have shown that Horn's parallel procedure performed the best; consequently, we recommend its use. However, no one rule is best under all circumstances. One should take into consideration the purpose of the study, the type of data, and the trade-off between parsimony and the amount of variation in the data that the researcher is willing to sacrifice in order to achieve parsimony. Lastly, and more importantly, one should determine the interpretability of the principal components in deciding upon how many principal components should be retained.

4.4.4 Interpreting Principal Components

Since the principal components are linear combinations of the original variables, it is often necessary to interpret or provide a meaning to the linear combination. As mentioned earlier, one can use the loadings for interpreting the principal components. Consider the loadings for the first two principal components from Exhibit 4.3 [4] (i.e., when standardized data are used).

	Variables				
Loadings	**Bread**	**Hamburger**	**Milk**	**Oranges**	**Tomatoes**
Prin1	.772	.896	.529	.350	.788
Prin2	−.324	−.046	−.453	.837	.302

The higher the loading of a variable, the more influence it has in the formation of the principal component score and vice versa. Therefore, one can use the loadings to determine which variables are influential in the formation of principal components, and one can then assign a meaning or label to the principal component. But, what do we mean by influential? How high should the loading be before we can say that a given variable is influential in the formation of a principal component score? Unfortunately, there are no guidelines to help us in establishing how high is high. Traditionally, researchers have used a loading of .5 or above as the cutoff point. If we use .5 as the

[12] See Green (1978) for a discussion of this test.

cutoff value, then it can be said that the first principal component represents the price index for nonfruit items, and the second principal component represents the price of the fruit item (i.e., oranges). In other words, the first principal component is a measure of the prices of bread, hamburger, milk, and tomatoes across the cities and the second principal component is a measure of the price of oranges across the cities. Therefore, *Prin*1 can be labeled as the *CPI* of nonfruit items and *Prin*2 as the *CPI* of fruit items.

In many instances the retained principal components cannot be meaningfully interpreted. In such cases researchers have typically resorted to a rotation of the principal components. The PRINCOMP procedure does not have the option of rotating the principal components because, strictly speaking, the concept of rotation was primarily developed for factor analysis. If one desires to rotate the retained principal components, then one must use the PROC FACTOR procedure. Therefore, the concept of rotation is discussed in the next chapter.

4.4.5 Use of Principal Components Scores

The principal components scores can be plotted for further interpreting the results. For example, Figure 4.6 gives a plot of the first two principal components scores for standardized data. Based on a visual examination of the plot, one might argue that there are five groups or clusters of cities. The first cluster consists of cities that have average food prices for nonfruit items but higher prices for fruits; the second cluster consists of cities that have slightly lower prices for nonfruit items and average prices for fruits; the third cluster consists of cities with slightly higher prices for fruits and average prices for nonfruit items; the fourth cluster has high prices for nonfruit items and average prices for fruit items; and the fifth cluster has average prices for nonfruit items and low prices for fruits. Of course, this grouping or clustering scheme is visual and arbitrary. Formal clustering algorithms discussed in Chapter 7 could be used for grouping the cities with respect to the two principal components scores.

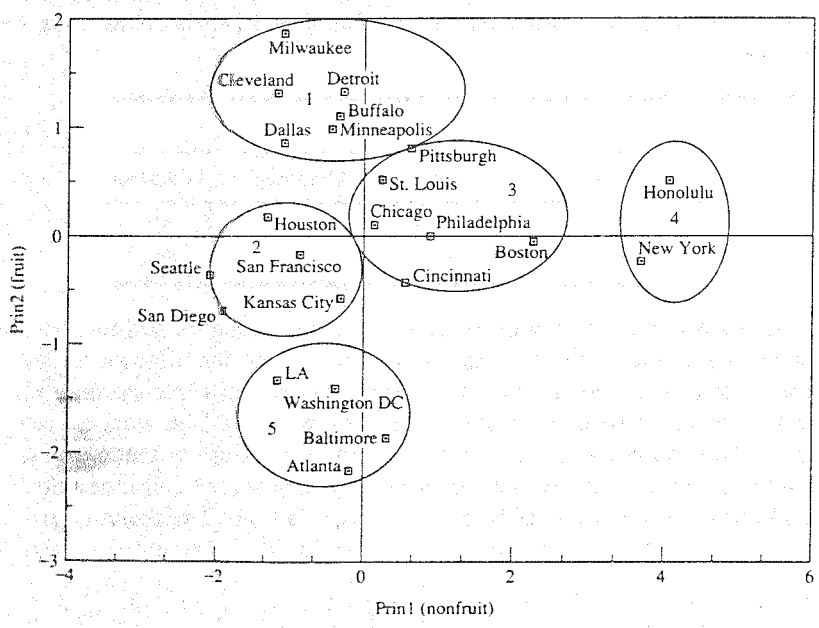


Figure 4.6 Plot of principal components scores.

The scores resulting from the principal components can also be used as input variables for further analyzing the data using other multivariate techniques such as cluster analysis, regression, and discriminant analysis. The advantage of using principal components scores is that the new variables are not correlated and the problem of multicollinearity is avoided. It should be noted, however, that although we may have "solved" the multicollinearity problem, a new problem can arise due to the inability to meaningfully interpret the principal components.

4.5 SUMMARY

This chapter provides a conceptual explanation of principal components analysis. The technique is described without the use of formal mathematics. The mathematical formulation of principal components analysis is given in the Appendix.

The main objective of principal components analysis is to form new variables that are linear combinations of the original variables. The new variables are referred to as the principal components and are uncorrelated with each other. Furthermore, the first principal component accounts for the maximum variance in the data, the second principal component accounts for the maximum of the variance that has not been accounted for by the first principal component, and so on. It is hoped that only a few principal components would be needed to account for most of the variance in the data. Consequently, the researcher needs to use only a few principal components rather than all of the variables. Therefore, principal components analysis is commonly classified as a data-reduction technique.

The results of principal components analysis can be affected by the type of data used (i.e., mean-corrected or standardized). If mean-corrected data are used then the relative variances of the variables have an effect on the weights used to form the principal components. Variables that have a high variance relative to other variables will receive a higher weight, and vice versa. To avoid the effect of the relative variance on the weights, one can use standardized data.

A number of statistical packages are available for performing principal components analysis. Hypothetical and actual data sets were used to demonstrate interpretation of the resulting output from SAS and to discuss various issues that arise when using principal components analysis. The next chapter discusses factor analysis. As was pointed out earlier, principal components analysis is often confused with factor analysis. In the next chapter we will provide a discussion of the similarities and the differences between the two techniques.

QUESTIONS

4.1 The following table provides six observations on variables x_1 and x_2:

Observation	x_1	x_2
1	1	1
2	5	4
3	4	3
4	1	2
5	2	1
6	4	5

(a) Compute the variance of each variable. What percentage of the total variance is accounted for by x_1 and x_2 respectively?

(b) Let X_1^* be any axis in a two-dimensional space making an angle of θ with X_1. Projection of the observations on X_1^* give the coordinates x_1^* of the observations with respect to X_1^*. Express x_1^* as a function of θ, x_1, and x_2.

(c) For what value of θ does x_1^* have the maximum variance? What percentage of the total variance is accounted for by x_1^*?

4.2 Given the covariance matrix

$$\Sigma = \begin{bmatrix} 8 & 0 & 1 \\ 0 & 8 & 3 \\ 1 & 3 & 5 \end{bmatrix}$$

(a) Compute the eigenvalues λ_1, λ_2, and λ_3 of Σ, and the eigenvectors γ_1, γ_2, and γ_3 of Σ.

Hint: You may use the PROC MATRIX or PROC IML procedures in SAS to compute the eigenvalues and eigenvectors.

(b) Show that $\lambda_1 + \lambda_2 + \lambda_3 = \text{tr}(\Sigma)$ where the trace of a matrix equals the sum of its diagonal elements.

(c) Show that $\lambda_1\lambda_2\lambda_3 = |\Sigma|$ where $|\Sigma|$ is the determinant of Σ.

(d) $\mathbf{x}'_1\mathbf{x}_2 = \mathbf{x}'_1\mathbf{x}_3 = \mathbf{x}'_2\mathbf{x}_3$. What does this imply?

4.3 Given

$$\bar{\mathbf{x}} = \begin{bmatrix} 12.45 \\ 1.35 \end{bmatrix} \qquad \hat{\mathbf{S}} = \begin{bmatrix} 65.41 & 4.57 \\ 4.57 & 1.27 \end{bmatrix}$$

(a) Use PROC IML to determine the sample principal components and their variances.

(b) Compute the loadings of the variables.

(c) What interpretation, if any, can you give to the first principal component? (Assume that x_1 = return on income and x_2 = earnings before interest and taxes.)

(d) Would the results change if correlation matrix is used to extract the principal components? Why? (Answer this question without computing the principal components.)

4.4 File FOODP.DAT gives the average price in cents per pound of five food items in 24 U.S. cities.[a]

(a) Using principal components analysis, define price index measure(s) based on the five food items.

(b) Identify the most and least expensive cities (based on the above price index measures). Do the most and least expensive cities change when standardized data are used as against mean-corrected data? Which type of data should be used to define price index measures? Why?

(c) Plot the data using principal components scores and identify distinct groups of cities. How are these groups different from each other?

4.5 The Personnel Department of a large multinational company commissioned a marketing research firm to undertake a study to measure the attitudes of junior executives employed by the company. As part of the study, the marketing research firm collected responses on 12 statements. Nineteen junior executives responded to the twelve statements on a five-point scale (1 = disagree strongly to 5 = agree strongly). The data collected are given in file PERS.DAT. The twelve statements are given in File PERS.DOC.

Use principal components analysis to analyze the data and help the marketing research firm identify key attitudes. How would you label these attitudes?

4.6 Consumers intending to purchase an automobile were asked to rate the following benefits desired by them in an automobile:

1. My car should have sleek, sporty looks.
2. My car should have dual air bags.
3. My car should be capable of accelerating to high speeds within seconds.

[a]U.S. Department of Labor, Bureau of Labor Statistics, Washington, D.C., May 1978.

4. My car should have luxurious upholstery.
5. I want excellent dealer service.
6. I want automatic transmission in my car.
7. I want my car to have high gas mileage.
8. I want power windows and power door locks in my car.
9. My car should be the fastest model in the market.
10. I want to impress my friends with the looks of my car.
11. My car should have air conditioning.
12. My car should have AM/FM radio and cassette player installed.
13. I want my car dealer to be located close to where I live.
14. I want tires that ensure safe driving under bad road conditions.
15. My car should have power brakes.
16. The exterior color of my car should be compatible with the upholstery color.
17. My car should have a powerful engine that provides fast acceleration.
18. My car should be equipped with safety belts.
19. My car should come with a service warranty that covers all the major parts.

Respondents indicated their agreement with the above statements using a five-point scale (1 = strongly disagree to 5 = strongly agree). The following table gives the loadings of the benefits on the principal components with eigenvalues greater than one.

	Loadings				
Benefits	**Prin1**	**Prin2**	**Prin3**	**Prin4**	**Prin5**
1	0.753	0.211	0.125	0.231	0.126
2	0.252	0.152	0.702	0.001	0.014
3	0.014	0.762	0.114	0.025	0.056
4	0.310	0.411	0.014	0.683	0.008
5	0.215	0.012	0.005	0.114	0.902
6	0.004	0.003	0.215	0.723	0.104
7	0.515	0.187	0.210	0.056	0.102
8	0.285	0.241	0.298	0.853	0.201
9	0.312	0.825	0.331	0.152	0.005
10	0.851	0.216	0.025	0.004	0.310
11	0.141	0.265	0.001	0.675	0.008
12	0.120	0.305	0.002	0.069	0.025
13	0.015	0.411	0.214	0.145	0.699
14	0.341	0.012	0.896	0.214	0.014
15	0.421	0.001	0.222	0.598	0.104
16	0.672	0.056	0.017	0.009	0.025
17	0.122	0.803	0.105	0.056	0.017
18	0.301	0.219	0.692	0.012	0.112
19	0.111	0.212	0.210	0.178	0.707

From the loadings given above identify the benefits that contribute significantly to each principal component and label the principal components. What therefore are the key dimensions that are considered by prospective car buyers? What is the correlation between these key dimensions?

4.7 File AUDIO.DAT gives the audiometric data[b] for 100 males, age 39. An audiometer is used to expose an individual to a signal of a given frequency with an increasing intensity until the signal is perceived. These threshold measurements are calibrated in units referred

[b]Jackson, J. Edward (1991). *A User's Guide to Principal Components.* New York: John Wiley & Sons, Table 5.1, pp. 107–109.

to as *decibel loss* in comparison to a reference standard for the instrument. Observations were obtained, one ear at a time, for the frequencies 500 Hz, 1000 Hz, 2000 Hz, and 4000 Hz. The limits of the instrument are −10 to 99 decibels. A negative value does not imply better than average hearing; the audiometer had a calibration "zero" and these observations are in relation to that. Perform a principal components analysis on the data. How many components should be retained? On what basis? What do the retained components represent?

4.8 Following the 1973–74 Arab oil embargo and the subsequent dramatic increase in oil prices, a study was conducted in three cities of a southern state to estimate the potential demand for mass transportation. The data from this survey are given in File MASST.DAT and a description of the data and the variables are provided in File MASST.DOC.

Perform principal components analysis on the variables $V_{19} - V_{38}$ (ignore the other variables for this question) to identify the key perceptions about the energy crisis. What do the retained components represent?

Appendix

We will show that principal components analysis reduces to finding the eigenstructure of the covariance matrix of the original data. Alternatively, principal components analysis can also be done by finding the singular value decomposition (SVD) of the data matrix or a spectral decomposition of the covariance matrix.

A4.1 EIGENSTRUCTURE OF THE COVARIANCE MATRIX

$X_{n \times p}$ ابعاد

نسبت میانگین نقطه چرخش

Let X be a p-component random vector where p is the number of variables. The covariance matrix, Σ, is given by $E(XX')$. Let $\gamma' = (\gamma_1\, \gamma_2\, \cdots\, \gamma_p)$ be a vector of weights to form the linear combination of the original variables, and $\xi = \gamma' X$ be the new variable, which is a linear combination of the original variables. The variance of the new variable is given by the $E(\xi\xi')$ and is equal to $E(\gamma' XX' \gamma)$ or $\gamma' \Sigma \gamma$. The problem now reduces to finding the weight vector, γ', such that the variance, $\gamma' \Sigma \gamma$, of the new variable is maximum over the class of linear combinations that can be formed subject to the constraint $\gamma' \gamma = 1$.

The solution to the maximization problem can be obtained as follows:

Let

$$Z = \gamma' \Sigma \gamma - \lambda(\gamma' \gamma - 1), \tag{A4.1}$$

where λ is the Lagrange multiplier. The p-component vector of the partial derivative is given by

$$\frac{\partial Z}{\partial \gamma} = 2\Sigma\gamma - 2\lambda\gamma. \tag{A4.2}$$

Setting the above vector of partial derivatives to zero results in the final solution. That is,

$$(\Sigma - \lambda I)\gamma = 0. \tag{A4.3}$$

For the above system of homogeneous equations to have a nontrivial solution the determinant of $(\Sigma - \lambda I)$ should be zero. That is,

$$|\Sigma - \lambda I| = 0. \tag{A4.4}$$

Equation A4.4 is a polynomial in λ of order p, and therefore has p roots. Let $\lambda_1 \geq, \lambda_2 \geq, \ldots, \lambda_p$ be the p roots. That is, Eq. A4.4 results in p values for λ, and each value is called the eigenvalue or root of the Σ matrix. Each value of λ results in a set of weights given by the p-component vector γ by solving the following equations:

$$(\Sigma - \lambda \mathbf{I})\gamma = \mathbf{0} \tag{A4.5}$$

$$\gamma'\gamma = 1. \tag{A4.6}$$

Therefore, the first eigenvector, γ_1, corresponding to the first eigenvalue, λ_1, is obtained by solving equations

$$(\Sigma - \lambda_1 \mathbf{I})\gamma_1 = \mathbf{0} \tag{A4.7}$$

$$\gamma_1'\gamma_1 = 1. \tag{A4.8}$$

Premultiplying Eq. A4.7 by γ_1' gives

$$\begin{aligned} \gamma_1'(\Sigma - \lambda_1 \mathbf{I})\gamma_1 &= \mathbf{0} \\ \gamma_1'\Sigma\gamma_1 &= \lambda_1\gamma_1'\gamma_1 \\ \gamma_1'\Sigma\gamma_1 &= \lambda_1 \end{aligned} \tag{A4.9}$$

as $\gamma_1'\gamma_1 = 1$. The left-hand side of Eq. A4.9 is the variance of the new variable, ξ_1, and is equal to the eigenvalue, λ_1. The first principal component, therefore, is given by the eigenvector, γ_1, corresponding to the largest eigenvalue, λ_1.

Let γ_2 be the second p-component vector of weights to form another linear combination. The next linear combination can be found such that the variance of $\gamma_2'\mathbf{X}$ is the maximum subject to the constraints $\gamma_1'\gamma_2 = 0$ and $\gamma_2'\gamma_2 = 1$. It can be shown that γ_2' is the eigenvector of λ_2, the second largest eigenvalue of Σ. Similarly, it can be shown that the remaining principal components, $\gamma_3', \gamma_4', \ldots, \gamma_p'$, are the eigenvectors corresponding to the eigenvalues, $\lambda_3, \lambda_4, \ldots, \lambda_p$, of the covariance matrix, Σ. Thus, the problem of finding the weights reduces to finding the eigenstructure of the covariance matrix. The eigenvectors give the vectors of weights and the eigenvalues represent the variances of the new variables or the principal components scores.

معمولاً بردار ویژه‌ها به روش SVD محاسبه می‌کنند.

A4.2 SINGULAR VALUE DECOMPOSITION

Singular value decomposition (SVD) expresses any $n \times p$ matrix (where $n \geq p$) as a triple product of three matrices, $\mathbf{P}$, $\mathbf{D}$, and $\mathbf{Q}$ such that

$$\mathbf{X} = \mathbf{PDQ}', \tag{A4.10}$$

where $\mathbf{X}$ is an $n \times p$ matrix of column rank r, $\mathbf{P}$ is an $n \times r$ matrix, $\mathbf{D}$ is an $r \times r$ diagonal matrix, and $\mathbf{Q}'$ is an $r \times p$ matrix. The matrices $\mathbf{P}$ and $\mathbf{Q}$ are orthonormal; that is,

$$\mathbf{P}'\mathbf{P} = \mathbf{I}, \tag{A4.11}$$

and

$$\mathbf{Q}'\mathbf{Q} = \mathbf{I}. \tag{A4.12}$$

The p column of $\mathbf{Q}'$ contain the eigenvectors of the $\mathbf{X}'\mathbf{X}$ matrix and the diagonals of the $\mathbf{D}$ matrix contain the square root of the corresponding eigenvalues of the $\mathbf{X}'\mathbf{X}$ matrix. Also, the eigenvalues of the matrices $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}\mathbf{X}'$ are the same.

A4.2.1 Singular Value Decomposition of the Data Matrix

Let $\mathbf{X}$ be an $n \times p$ data matrix. Since $\mathbf{X}$ is a data matrix it will be assumed that its rank is p (i.e., $r = p$) and consequently $\mathbf{Q}$ will be a square symmetric matrix. The columns of $\mathbf{Q}$ will give the

eigenvectors of the $\mathbf{X}'\mathbf{X}$ matrix and the diagonal values of the $\mathbf{D}$ matrix will give the square root of the corresponding eigenvalues of the $\mathbf{X}'\mathbf{X}$ matrix.

Let Ξ be an $n \times p$ matrix of the values of the new variables or principal components scores. Then:

$$\begin{aligned} \Xi &= \mathbf{XQ} \\ &= (\mathbf{PDQ}')\mathbf{Q} \\ &= \mathbf{PDQ}'\mathbf{Q} \\ &= \mathbf{PD}. \end{aligned} \tag{A4.13}$$

The covariance matrix, Σ_ξ, of the new variables is given by:

$$\begin{aligned} \Sigma_\xi = E(\Xi'\Xi) &= E[(\mathbf{PD})'(\mathbf{PD})] \\ &= E(\mathbf{D}'\mathbf{P}'\mathbf{PD}) \\ &= E(\mathbf{D}^2) \\ &= \frac{1}{n-1}\mathbf{D}^2. \end{aligned} \tag{A4.14}$$

Since $\mathbf{D}$ is a diagonal matrix the new variables are uncorrelated among themselves.

As can be seen from the preceding discussion, the SVD of the data matrix also gives the principal components analysis solution. The weights for forming the new variables are given by the matrix $\mathbf{Q}$, the principal components scores are given by $\mathbf{PD}$, and the variances of the new variables are given by $\mathbf{D}^2/(n-1)$.

A4.3 SPECTRAL DECOMPOSITION OF A MATRIX

The singular value decomposition of a square symmetric matrix is also called the spectral decomposition of a matrix. Any $p \times p$ square symmetric matrix $\mathbf{X}$ can be written as a product of two matrices, $\mathbf{P}$ and Λ, such that

$$\mathbf{X} = \mathbf{P}\Lambda\mathbf{P}', \tag{A4.15}$$

where $\mathbf{P}$ is a $p \times p$ square symmetric orthogonal matrix containing the eigenvectors of the $\mathbf{X}$ matrix, and the $p \times p$ diagonal matrix Λ contains the eigenvalues of the $\mathbf{X}$ matrix. Also, $\mathbf{P}'\mathbf{P} = \mathbf{PP}' = \mathbf{I}$.

A4.3.1 Spectral Decomposition of the Covariance Matrix

Since Σ is a square symmetric matrix, its spectral decomposition can be written as

$$\Sigma = \mathbf{P}\Lambda\mathbf{P}',$$

where Λ is a diagonal matrix whose elements are the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$ of the symmetric matrix, Σ, and $\mathbf{P}$ is a $p \times p$ orthogonal matrix whose jth column is the eigenvector corresponding to the jth eigenvalue, and so on.

Values of the new variables or principal components scores are given by the matrix $\Xi = \mathbf{XP}$ and the covariance matrix of the principal components scores is given by

$$\begin{aligned} \Sigma_\xi = E(\Xi'\Xi) &= E(\mathbf{XP})'(\mathbf{XP}) \\ &= E(\mathbf{P}'\mathbf{X}'\mathbf{XP}) \\ &= \mathbf{P}'\Sigma\mathbf{P}. \end{aligned} \tag{A4.16}$$

Substituting for the covariance matrix Σ we get

$$\begin{aligned}\Sigma_\xi &= \mathbf{P}'\mathbf{P}\Lambda\mathbf{P}'\mathbf{P} \\ &= \Lambda \end{aligned} \tag{A4.17}$$

as $\mathbf{P}'\mathbf{P} = \mathbf{I}$. Therefore, the new variables $\xi_1, \xi_2, \ldots, \xi_p$ are uncorrelated with variances equal to $\lambda_1, \lambda_2, \ldots, \lambda_p$, respectively.

Also, we can see that the trace of Σ is given by

$$\mathrm{tr}(\Sigma) = \Sigma_{j=1}^{p}\sigma_{jj}^2 \tag{A4.18}$$

where σ_{jj}^2 is the variance of the jth variable. The trace of Σ can also be represented as

$$\begin{aligned}\mathrm{tr}(\Sigma) &= \mathrm{tr}(\mathbf{P}\Lambda\mathbf{P}') \\ &= \mathrm{tr}(\mathbf{P}'\mathbf{P}\Lambda) \\ &= \mathrm{tr}(\Lambda) \\ &= \mathrm{tr}(\Sigma_\xi), \end{aligned} \tag{A4.19}$$

which is equal to the sum of the eigenvalues of the covariance matrix, Σ. The preceding results show that the total variance of the original variables is the same as the total variance of the new variables (i.e., the linear combinations).

In conclusion, principal components analysis reduces to finding the eigenvalues and eigenvectors of the covariance matrix, or finding the SVD of the original data matrix $\mathbf{X}$, or obtaining the spectral decomposition of the covariance matrix.

A4.4 ILLUSTRATIVE EXAMPLE

The PROC IML procedure in SAS can be used to obtain the eigenstructure, SVD, and the spectral decomposition of the appropriate matrices. The data in Table 4.1 are used to illustrate PROC IML. Table A4.1 gives the SAS commands for PROC IML. Most of the commands have been discussed in the Appendix of Chapter 3. These commands compute the means, mean-corrected data, and the covariance and correlation matrices. The CALL EIGEN(EVAL,EVEC,SIGMA) command requests the eigenstructure of the Σ matrix (i.e., SIGMA). The eigenvalues and eigenvectors, respectively, are stored in EVAL and EVEC. The CALL SVD(P,D,Q,XM) requests a singular-value decomposition on the XM matrix (i.e., mean-corrected data) and CALL SVD(P,LAMBDA,Q,SIGMA) requests spectral decomposition (i.e., singular-value decomposition) on the SIGMA matrix. The PRINT command requests the printing of the various matrices. The output is given in Exhibit A4.1. Comparing the output in Exhibit A4.1 to that in Exhibit 4.1, one can see that:

1. For the eigenstructure of the covariance matrix, EVAL gives the eigenvalues and EVEC gives the weights for forming the principal components scores.
2. For the singular-value decomposition of the mean-corrected data matrix, the columns of Q are the same as the weights for forming the principal components scores. Note that $\mathbf{D}^2/(n-1)$ gives the variances of the principal components scores, and the PD matrix gives the principal components scores.
3. For the singular-value decomposition of the covariance matrix (i.e., spectral decomposition), the columns of P give the weights and the LAMBDA matrix gives the variances of the principal components scores.

Table A4.1 **PROC IML Commands**

```
TITLE PROC IML COMMANDS FOR MATRIX MANIPULATIONS ON DATA IN
  TABLE 4.1;
OPTIONS NOCENTER;
DATA TEMP;
 INPUT X1 X2;
 CARDS;
 insert data here;
PROC IML;
USE TEMP;
READ ALL INTO X; * READ DATA INTO X MATRIX;
N=NROW(X); * N CONTAINS THE NUMBER OF OBSERVATIONS;
ONE=J(N,1,1); * 12X1 VECTOR CONTAINING ONES;
DF=N-1;
MEAN=(ONE`*X)/N; * MEAN MATRIX CONTAINS THE MEANS;
XM=X-ONE*MEAN;  * XM MATRIX CONTAINS THE MEAN-CORRECTED DATA;
SSCPM=XM`*XM;
SIGMA=SSCPM/DF;
D=DIAG(SIGMA);
XS=XM*SQRT(INV(D)); * XS MATRIX CONTAINS THE STANDARDIZED DATA;
R=XS`*XS/(N-1); * R IS THE CORRELATION MATRIX;
CALL EIGEN(EVAL,EVEC,SIGMA); *EIGENSTRUCTURE OF THE
  COVARIANCE MATRIX;
CALL SVD(P,D,Q,XM);  *SINGULAR VALUE DECOMPOSITION OF THE
  DATA MATRIX;
D=DIAG(D);
SCORES=P*D;   * COMPUTING THE PRINCIPAL COMPONENTS SCORES;
CALL SVD(P,LAMBDA,Q,SIGMA); *SPECTRAL DECOMPOSITION OF
  COVARIANCE MATRIX;
PRINT EVAL,EVEC;
PRINT Q,D,SCORES;
PRINT P,LAMBDA,Q;
```

Exhibit A4.1

PROC IML output

```
     EVAL
38.575813
5.6060049

     EVEC
0.7282381 -0.685324
0.6853242 0.7282381

        Q
0.7282381 -0.685324
0.6853242 0.7282381

        D
20.599368         0
        0 7.8527736

   SCORES
9.2525259 -1.841403
7.7102217 2.3563703
5.6971632 -1.241906
1.4993902 -2.784211
4.8830971 2.2705423
-2.013059 -3.598277
0.6853242 0.7282381
1.3277344 2.8700386
-6.296659 -2.313456
-6.382487 0.5136683
-8.481374 -0.257484
-7.881878  3.297879

        P
0.7282381 -0.685324
0.6853242 0.7282381

   LAMBDA
38.575813
5.6060049

        Q
0.7282381 -0.685324
0.6853242 0.7282381
```

CHAPTER 5

Factor Analysis

Consider each of the following situations.

- The marketing manager of an apparel firm wants to determine whether or not a relationship exists between patriotism and consumers' attitudes about domestic and foreign products.
- The president of a Fortune 500 firm wants to measure the firm's image.
- A sales manager is interested in measuring the sales aptitude of salespersons.
- Management of a high-tech firm is interested in measuring determinants of resistance to technological innovations.

Each of the above examples requires a scale, or an instrument, to measure the various constructs (i.e., attitudes, image, patriotism, sales aptitude, and resistance to innovation). These are but a few examples of the type of measurements that are desired by various business disciplines. Factor analysis is one of the techniques that can be used to develop scales to measure these constructs.

In this chapter we discuss factor analysis and illustrate the various issues using hypothetical data. The discussion is mostly analytical as the geometry of factor analysis is not as simple or straightforward as that of principal components analysis. Mathematical details are provided in the Appendix. Although factor analysis and principal components analysis are used for data reduction, the two techniques are clearly different. We also provide a discussion of the similarities between factor and principal components analysis, and between exploratory and confirmatory factor analysis. Confirmatory factor analysis is discussed in the next chapter.

5.1 BASIC CONCEPTS AND TERMINOLOGY OF FACTOR ANALYSIS

Factor analysis was originally developed to explain student performance in various courses and to understand the link between grades and intelligence. Spearman (1904) hypothesized that students' performances in various courses are intercorrelated and their intercorrelations could be explained by students' general intelligence levels. We will use a similar example to discuss the concept of factor analysis.

Suppose we have students' test scores (grades) for the following courses: Mathematics (M), Physics (P), Chemistry (C), English (E), History (H), and French (F). Further assume that students' performances in these courses are a function of their general

intelligence level, I. In addition, it can be hypothesized that students' aptitudes for the subject areas could be different. That is, a given student may have a greater aptitude for, say, math than French. Therefore, it can be assumed that a student's grade for any given course is a function of:

1. The student's general intelligence level; and
2. The student's aptitude for a given course (i.e., the specific nature of the subject area).

For example, consider the following equations:

$$\begin{aligned} M &= .80I + A_m; & P &= .70I + A_p \\ C &= .90I + A_c; & E &= .60I + A_e \\ H &= .50I + A_h; & F &= .65I + A_f. \end{aligned} \tag{5.1}$$

It can be seen from these equations that a student's performance on any given course, say math, is a linear function or combination of the general intelligence level, I, of the student, and his/her aptitude, A_m, for the specific subject, math. The coefficients (i.e., .8, .7, .9, .6, .5, and .65) of the above equations are called *pattern loadings*.

The relationship between grades and general intelligence level can also be depicted graphically as shown in Figure 5.1. In the figure, for any given jth variable the arrows from I and A_j to the variable indicate that the value of the variable is a function of I and A_j, and the variable is called *indicator* or *measure* of I. Note that Eq. 5.1 can be viewed as a set of regression equations where the grade of each subject is the dependent variable, the general intelligence level (I) is the independent variable, the unique factor (A_j) is the error term, and the pattern loadings are the regression coefficients. The variables can be considered as indicators of the construct I, which is responsible for the correlation among the indicators.[1] In other words, the various indicators (i.e., course grades) correlate among themselves because they share at least one common trait or feature, namely, level of raw intelligence. Since the general intelligence level construct is responsible for all of the correlation among the indicators and cannot be directly observed, it is referred to as *common* or *latent factor*, or as an *unobservable construct*.

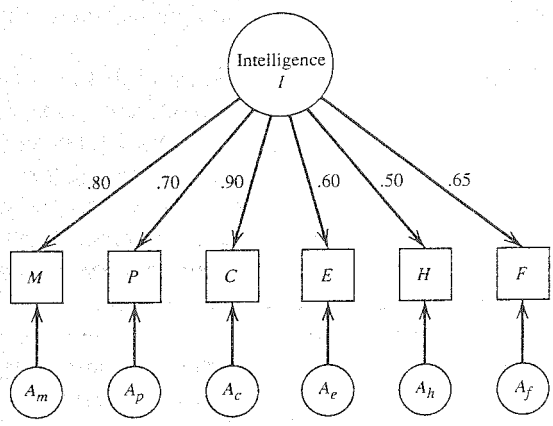


Figure 5.1 Relationship between grades and intelligence.

[1]Hereafter the terms *indicators* and *variables* will be used interchangeably.

It can be shown (see Eqs. A5.2, A5.3, and A5.4 of the Appendix) that:

1. The total variance of any indicator can be decomposed into the following two components:
   - Variance that is in common with general intelligence level, I, and is given by the square of the pattern loading; this part of the variance is referred to as the *communality* of the indicator with the common factor.
   - Variance that is in common with the specific factor, A_j, and is given by the variance of the variable minus the communality. This part of the variance is referred to as the *unique* or *specific* or *error* variance because it is unique to that particular variable.
2. The simple correlation between any indicator and the latent factor is called the *structure loading* or simply the *loading* of the indicator and is usually the same as the *pattern loading*.[2] (Further discussion of the differences between pattern and structural loading is provided in the next section and in Sections A5.2 and A5.5.1 of the Appendix.) The square of the structure loading is referred to as the shared variance between the indicator and the factor. That is, shared variance between an indicator and a factor is the indicator's communality with the factor. Often, the communality is used to assess the degree to which an indicator is a good or reliable measure of the factor. The greater the communality, the better the measure (i.e., reliable measure) and vice versa. Since communality is equal to the square of the structure loading, the structure loading can also be used to assess the degree to which a given indicator measures the construct.
3. The correlation between any two indicators is given by the product of their respective pattern loadings.

For the factor model depicted in Figure 5.1, Table 5.1 gives the communalities, unique variances, pattern and structure loadings, shared variances, and the correlation among the variables. The computations in Table 5.1 assume, without any loss of generality, that: (a) means of indicators, common factor I, and the unique factors are zero; (b) variances of the indicators and the common factor, I, are one; (c) correlations between the common factor, I, and the unique factors are zero; and (d) the correlations among the unique factors are zero.

From the above discussion, it is clear that correlations among the indicators are due to the common factor, I. For example, if the pattern loading of any one indicator is zero, then the correlations between this indicator and the remaining indicators will be zero. That is, there is one common factor, I, which links the indicators together and is, therefore, responsible for all of the correlations that exist among the indicators. Alternatively, if the effect of factor I is removed from the correlations, then the partial correlations will be zero. The correlation between M and P, for example, after the effect of factor I has been partialled out will be zero. Furthermore, it can be seen that not all of the indicator's variance is explained or accounted for by the common factor.

Since the common factor is unobservable, we cannot measure it directly; however, we can measure the indicators of the unobservable factor and compute the correlation

[2]For a one-factor model the structure and the pattern loadings are always the same. However, as discussed in later sections, this may not be true for models with two or more factors.

Table 5.1 **Communalities, Pattern and Structure Loadings, and Correlation Matrix for One-Factor Model**

Communalities

Variable	Communality	Error or Unique Variance	Pattern Loading	Structural Loading	Shared Variance
M	.640	.360	.800	.800	.640
P	.490	.510	.700	.700	.490
C	.810	.190	.900	.900	.810
E	.360	.640	.600	.600	.360
H	.250	.750	.500	.500	.250
F	.423	.577	.650	.650	.423
Total	2.973	3.027			2.973

Correlation Matrix for One-Factor Model

	M	P	C	E	H	F
M	1.000					
P	.56	1.000				
C	.72	.63	1.000			
E	.48	.42	.54	1.000		
H	.40	.35	.45	.30	1.000	
F	.52	.46	.59	.39	.33	1.000

matrix containing the correlations among the indicators. Now given the computed correlation matrix among the indicators, the purpose of factor analysis is to

1. Identify the common factor that is responsible for the correlations among the indicators; and
2. Estimate the pattern and structure loadings, communalities, shared variances, and the unique variances.

In other words, the objective of factor analysis is to obtain the structure presented in Figure 5.1 and Table 5.1 using the correlation matrix. That is, the correlation matrix is the input for the factor analysis procedure and the outputs are the entries in Table 5.1.

In the preceding example we had only one common factor explaining the correlations among the indicators. Factor models that use only one factor to explain the underlying structure or the correlations among the indicators are called *single-* or *one-factor models*. In the following section we discuss a two-factor model.

5.1.1 Two-Factor Model

It may not always be possible to completely explain the interrelationship among the indicators by just one common factor. There may be two or more latent factors or constructs that are responsible for the correlations among the indicators. For example, one could hypothesize that students' grades are a function of not one, but two latent constructs or factors. Let us label these two factors as Q and V.[3] The two-factor model is

[3]The reason for using these specific labels will become clear later.

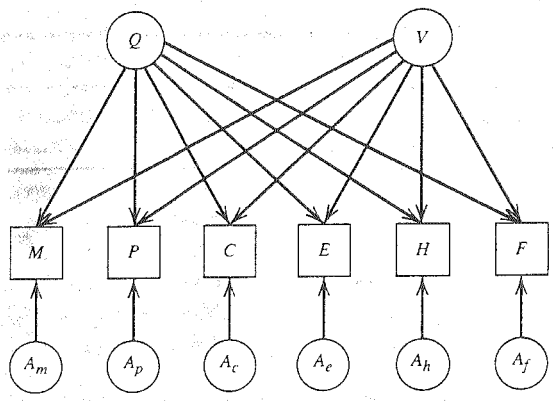


Figure 5.2 Two-factor model.

depicted in Figure 5.2 and can be represented by the following equations:

$$\begin{aligned} M &= .800Q + .200V + A_m; & P &= .700Q + .300V + A_p \\ C &= .600Q + .300V + A_c; & E &= .200Q + .800V + A_e \\ H &= .150Q + .820V + A_h; & F &= .250Q + .850V + A_f. \end{aligned} \tag{5.2}$$

In the above equations, a student's grade for any subject is a function or a linear combination of the two common factors, Q and V, and a unique factor. The two common factors are assumed to be uncorrelated. Such a model is referred to as an orthogonal factor model.

As shown in Eqs. A5.7, A5.9, and A5.13 of the Appendix:

1. Variance of any indicator can be decomposed into the following three components:
   - Variance that is in common with the Q factor and is equal to the square of its pattern loading. This variance is referred to as the indicator's communality with the common factor, Q.
   - Variance that is in common with the V factor and is equal to the square of its pattern loading. This variance is referred to as the indicator's communality with the common factor, V. The total variance of an indicator that is in common with both the latent factors, Q and V, is referred to as the total communality of the indicator.
   - Variance that is in common with the unique factor, and is equal to the variance of the variable minus the communality of the variable.
2. The coefficients of Eq. 5.2 are referred to as the pattern loadings, and the simple correlation between any indicator and the factor is equal to its structure loading. The shared variance between an indicator and a factor is equal to the square of its structure loading. As before, communality is equal to the shared variance. Notice that once again Eq. 5.2 represents a set of regression equations in which the grade of each subject is the dependent variable, V and Q are the independent variables, and the pattern loadings are the regression coefficients. Now in regression analysis the regression coefficients will be same as the simple correlations between the independent variables and the dependent variable only if the independent variables are uncorrelated among themselves. If, on the other hand, the independent variables are correlated among themselves then the regression coefficients will not be the same as the simple correlations between the independent variables and the dependent variable. Consequently, the pattern and structure loadings will only be the

same if the two factors are uncorrelated (i.e., if the factor model is orthogonal). This is further discussed in Sections A5.2 and A5.5.1 of the Appendix.

3. The correlation between any two indicators is equal to the sum of the products of the respective pattern loadings for each factor (see Eq. A5.13 of the Appendix). For example, the correlation between the math and history grades is given by

$$.800 \times .150 + .200 \times .820 = .284.$$

Note that now the correlation between the indicators is due to two common factors. If any given indicator is not related to the two factors (i.e., its pattern loadings are zero), then the correlation between this indicator and other indicators will be zero. In other words, correlations among the indicators are due to the two common factors, Q and V. Table 5.2 gives the communalities, unique variances, pattern and structure loadings,

Table 5.2 **Communalities, Pattern and Structure Loadings, and Correlation Matrix for Two-Factor Model**

Communalities

	Communalities			
Variable	Q	V	Total	Unique Variance
M	.640	.040	.680	.320
P	.490	.090	.580	.420
C	.360	.090	.450	.550
E	.040	.640	.680	.320
H	.023	.672	.695	.305
F	.063	.723	.786	.214
Total	1.616	2.255	3.871	2.129

Pattern and Structure Loadings and Shared Variance

	Pattern Loading		Structure Loading		Shared Variance	
Variable	Q	V	Q	V	Q	V
M	.800	.200	.800	.200	.640	.040
P	.700	.300	.700	.300	.490	.090
C	.600	.300	.600	.300	.360	.090
E	.200	.800	.200	.800	.040	.640
H	.150	.820	.150	.820	.023	.672
F	.250	.850	.250	.850	.063	.723
Total					1.616	2.255

Correlation Matrix

	M	P	C	E	H	F
M	1.000					
P	.620	1.000				
C	.540	.510	1.000			
E	.320	.380	.360	1.000		
H	.284	.351	.336	.686	1.000	
F	.370	.430	.405	.730	.735	1.000

and the correlation matrix. Note that the unique variance of each indicator/variable is equal to one minus the total communality.

Consequently, one can extend the objective of factor analysis to include the identification of the number of common factors required to explain the correlations among the indicators. Obviously, for the sake of parsimony, one would like to identify the least number of common factors that explain the maximum amount of correlation among the indicators.

In some instances researchers are also interested in obtaining values of the latent factors for each subject or observation. The values of the latent factors are called *factor scores*. Therefore, another objective of factor analysis is to estimate the factor scores.

5.1.2 Interpretation of the Common Factors

Having established that the correlations among the indicators are due to two common or latent factors, the next step is to interpret the two factors. From Table 5.2 it can be seen that the communalities or the shared variances of the variables E, H, and F with factor V are much greater than those with factor Q. Indeed, 90.24% $((.640+.672+.723)/2.255)$ of the total communality of V is due to variables E, H, and F. Therefore, one could argue that the common factor, V, measures subjects' verbal abilities. Similarly, one could argue that the common factor, Q, measures subjects' quantitative abilities because 92.20% $((.64 + .49 + .36)/1.616)$ of its communality is due to variables M, P, and C.

The above interpretation leads us to the following hypothesis or theory. Students' grades are a function of two common factors, namely quantitative and verbal abilities. The quantitative ability factor, Q, explains grades of such courses as math, physics, and chemistry and the verbal ability factor, V, explains grades of such courses as history, English, and French. Therefore, interpretation of the resulting factors can also be viewed as one of the important objectives of factor analysis.

5.1.3 More Than Two Factors

The preceding concept can be easily extended to a factor model that contains m factors. The m-factor model can be represented as:[4]

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \cdots + \lambda_{1m}\xi_m + \epsilon_1 \\ x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \cdots + \lambda_{2m}\xi_m + \epsilon_2 \\ &\vdots \\ x_p &= \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \cdots + \lambda_{pm}\xi_m + \epsilon_p. \end{aligned} \tag{5.3}$$

In these equations the intercorrelation among the p indicators is being explained by the m common factors. It is usually assumed that the number of common factors, m, is much less than the number of indicators, p. In other words, the intercorrelation among the p indicators is due to a small ($m < p$) number of common factors. The number of unique factors is equal to the number of indicators. If the m factors are not correlated the factor model is referred to as an orthogonal model, and if they are correlated it is referred to as an oblique model.

[4]To be consistent with the notation and the symbols used in standard textbooks, we use Greek letters to denote the unobservable constructs (i.e., the common factors), the unique factors, and the pattern loadings. Hence, in Eq. 5.3 the ξ's are the common factors, the λ's are the pattern loadings, and the ϵ's are the unique factors.

5.1.4 Factor Indeterminacy

The factor analysis solution is not unique due to two inherent indeterminacies: (1) factor indeterminacy due to the factor rotation problem; and (2) factor indeterminacy due to the estimation of communality problem. Each of these is discussed below.

Indeterminacy Due to the Factor Rotation Problem

Consider another two-factor model given by the following equations:

$$\begin{aligned} M &= .667Q - .484V + A_m; & P &= .680Q - .343V + A_p \\ C &= .615Q - .267V + A_c; & E &= .741Q + .361V + A_e \\ H &= .725Q + .412V + A_h; & F &= .812Q + .355V + A_f \end{aligned} \tag{5.4}$$

Table 5.3 gives the pattern and structure loadings, shared variances, communalities, unique variances, and the correlation matrix for the above factor model. Comparison of the results of Table 5.3 with those of Table 5.2 indicate that the loadings, shared variances, and communalities of each indicator are different. However, within rounding errors:

1. The total communalities of each variable are the same.
2. The unique variances of each variable are the same.
3. And the correlation matrices are identical.

It is clear that decomposition of the total communality of a variable into communalities of the variable with each factor is different for the two models; however, each model produces the *same correlations* between the indicators. That is, the factor solution is not unique. Indeed, one can decompose the total communality of a variable into the communality of that variable with each factor in an infinite number of ways, and each decomposition produces a different factor solution. Further, the interpretation of the factors for each factor solution might be different. For the factor model given by Eq. 5.4, factor Q can now be interpreted as a general intelligence factor because the communality of each variable is approximately the same. And factor V is interpreted as an aptitude factor that differentiates between quantitative and verbal ability of the subjects. This interpretation is reached because the communalities of each variable with the factor are about the same, but the loadings of variables M, P, and C are positive and the loadings for variables E, H, and F are negative. Furthermore, the general intelligence level factor accounts for almost 78.05% (3.019/3.868) of the total communality and the aptitude factor accounts for 21.95% of the total communality.

The preceding interpretation might give support to the following hypothesis: students' grades are, to a greater extent, a function of general or raw intelligence and, to a lesser extent, a function of the aptitude for the type of subject (i.e., quantitative or verbal).

The problem of obtaining multiple solutions in factor analysis is called the *factor indeterminacy due to rotation problem*, or simply the *factor rotation problem*. The question then becomes: which of the multiple solutions is the correct one? In order to obtain a unique solution, an additional constraint outside the factor model has to be imposed. This constraint pertains to providing a plausible interpretation of the factor model. For instance, for the two-factor solutions given by Eqs. 5.2 and 5.4, the solution that gives a theoretically more plausible or acceptable interpretation of the resulting factors would be considered to be the "correct" solution.

Table 5.3 **Communalities, Pattern and Structure Loadings, Shared Variances, and Correlation Matrix for Alternative Two-Factor Model**

Communalities

	Communalities			
Variable	Q	V	Total	Unique Variance
M	.445	.234	.679	.321
P	.462	.118	.580	.420
C	.378	.071	.449	.551
E	.549	.130	.679	.321
H	.526	.170	.696	.304
F	.659	.126	.785	.215
Total	3.019	.849	3.868	2.132

Pattern and Structure Loadings and Shared Variance

	Pattern Loading		Structure Loading		Shared Variance	
Variable	Q	V	Q	V	Q	V
M	.667	−.484	.667	−.484	.445	.234
P	.680	−.343	.680	−.343	.462	.118
C	.615	−.267	.615	−.267	.378	.071
E	.741	.361	.741	.361	.549	.130
H	.725	.412	.725	.412	.526	.170
F	.812	.355	.812	.355	.659	.126
Total					3.019	.849

Correlation Matrix

	M	P	C	E	H	F
M	1.000					
P	.620	1.000				
C	.540	.510	1.000			
E	.320	.380	.360	1.000		
H	.284	.351	.336	.686	1.000	
F	.370	.430	.405	.730	.735	1.000

Indeterminacy Due to the Estimation of Communality Problem

As will be seen later, in order to estimate the pattern and the structure loadings and the shared variance, an estimate of the communality of each variable is needed; however, in order to estimate the communality one needs estimates of the loadings. This circularity results in a second type of indeterminacy, referred to as the *indeterminacy due to the estimate of the communalities problem*, or simply as the *estimation of the communalities problem*. Indeed, many of the factor analysis techniques differ mainly with respect to the procedure used for estimating the communalities.

5.2 OBJECTIVES OF FACTOR ANALYSIS

As mentioned previously, the common factors are unobservable. However, we can measure their indicators and compute the correlation among the indicators. The objectives of factor analysis are to use the computed correlation matrix to:

1. Identify the smallest number of common factors (i.e., the most parsimonious factor model) that best explain or account for the correlations among the indicators.
2. Identify, via factor rotations, the most plausible factor solution.
3. Estimate the pattern and structure loadings, communalities, and the unique variances of the indicators.
4. Provide an interpretation for the common factor(s).
5. If necessary, estimate the factor scores.

That is, given the correlation matrices in Tables 5.1 and 5.2, estimate the corresponding factor structures depicted, respectively, in Figures 5.1 and 5.2 and provide a plausible interpretation of the resulting factors.

5.3 GEOMETRIC VIEW OF FACTOR ANALYSIS

The geometric illustration of factor analysis is not as straightforward as that of principal components analysis. However, it does facilitate the discussion of the indeterminacy problems discussed earlier. Consider the two-indicator, two-factor model given in Figure 5.3. The model can be represented as

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \epsilon_1 \\ x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \epsilon_2. \end{aligned}$$

Vectors $\mathbf{x}_1'$ and $\mathbf{x}_2'$ of n observations can be represented in an n-dimensional observation space. However, the two vectors will lie in a four-dimensional subspace defined by the orthogonal vectors ξ_1, ξ_2, ϵ_1, and ϵ_2. Specifically, $\mathbf{x}_1'$ will lie in the three-dimensional space defined by ξ_1, ξ_2, and ϵ_1 and $\mathbf{x}_2'$ will lie in the three-dimensional space defined by ξ_1, ξ_2, and ϵ_2. The objective of factor analysis is to identify these four vectors defining the four-dimensional subspace.

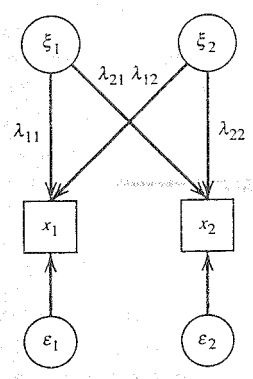


Figure 5.3 Two-indicator two-factor model.

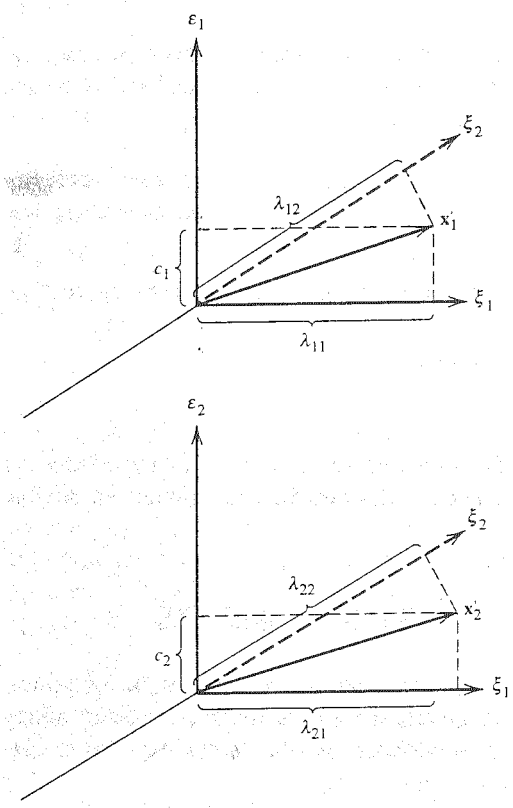


Figure 5.4 Indeterminacy due to estimates of communalities.

5.3.1 Estimation of Communalities Problem

As shown in Figure 5.4, let $\lambda_{11}, \lambda_{12}$, and c_1 be the projections of $\mathbf{x}_1'$ onto ξ_1, ξ_2, and ϵ_1, respectively, and $\lambda_{21}, \lambda_{22}$, and c_2 be the projections of $\mathbf{x}_2'$ onto ξ_1, ξ_2, and ϵ_2, respectively. From the Pythagorean theorem we know that

$$\|\mathbf{x}_1'\|^2 = \lambda_{11}^2 + \lambda_{12}^2 + c_1^2 \tag{5.5}$$

$$\|\mathbf{x}_2'\|^2 = \lambda_{21}^2 + \lambda_{22}^2 + c_2^2. \tag{5.6}$$

In these equations, $\lambda_{11}^2 + \lambda_{12}^2$ gives the communality of variable x_1, and $\lambda_{21}^2 + \lambda_{22}^2$ gives the communality of variable x_2. It is clear that the values of the communalities depend on the values of c_1^2 and c_2^2, or one can say that the value of c_1 depends on the values of λ_{11} and λ_{12} and the value of c_2 depends on the values of λ_{21} and λ_{22}. Therefore, in order to estimate the loadings one has to know the communalities of the variables or vice versa.

5.3.2 Factor Rotation Problem

Assuming that the axes ϵ_1 and ϵ_2 are identified and fixed (i.e., the communalities have been estimated), the vectors $\mathbf{x}_1'$ and $\mathbf{x}_2'$ can also be projected onto the two-dimensional subspace represented by ξ_1 and ξ_2. Figure 5.5 shows the resulting projection vectors, $\mathbf{x}_{1p}'$ and $\mathbf{x}_{2p}'$. The projection vectors, $\mathbf{x}_{1p}'$ and $\mathbf{x}_{2p}'$, can be further projected onto one-dimensional subspaces defined by vectors ξ_1 and ξ_2. Recall from Section 2.4.4 of

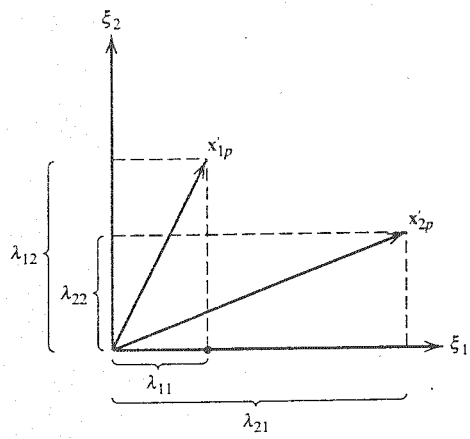


Figure 5.5 Projection of vectors onto a two-dimensional factor space.

Chapter 2 that the projection of a vector onto an axis gives the component of the point representing the vector with respect to that axis. These components (i.e., projections of the projection vectors) are the structure loadings and also the pattern loadings for orthogonal factor models. As shown in Figure 5.5, λ_{11} and λ_{12} are the structure loadings of $\mathbf{x}_1$ for ξ_1 and ξ_2, respectively, and λ_{21} and λ_{22} are the structure loadings of $\mathbf{x}_2$ for ξ_1 and ξ_2, respectively.

The square of the structure loadings gives the respective communalities. The communality of each variable is the sum of the communality of the variable with each of the two factors. That is, the communality for x_1 is equal to $\lambda_{11}^2 + \lambda_{12}^2$ and the communality for x_2 is equal to $\lambda_{21}^2 + \lambda_{22}^2$. From the Pythagorean theorem,

$$\|\mathbf{x}'_{1p}\|^2 = \lambda_{11}^2 + \lambda_{12}^2 \tag{5.7}$$

$$\|\mathbf{x}'_{2p}\|^2 = \lambda_{21}^2 + \lambda_{22}^2. \tag{5.8}$$

That is, lengths of the projection vectors give the communalities of the variables. The axes of Figure 5.5 can be rotated without changing the orientation or the length of the vectors $\mathbf{x}'_{1p}$ and $\mathbf{x}'_{2p}$ and hence the total communalities of the variables. The dotted axes in Figure 5.6 give one such rotation. It is clear from this figure that even though the total communality of a variable has not changed, the decomposition of the total communality will change. That is, decomposition of the total communality is arbitrary. This is also obvious from Eqs. 5.7 and 5.8. Each equation can be satisfied by an infinite number of values for the λ's. In other words, total communality of a variable can be

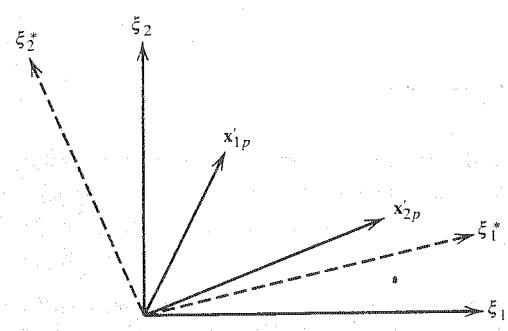


Figure 5.6 Rotation of factor solution.

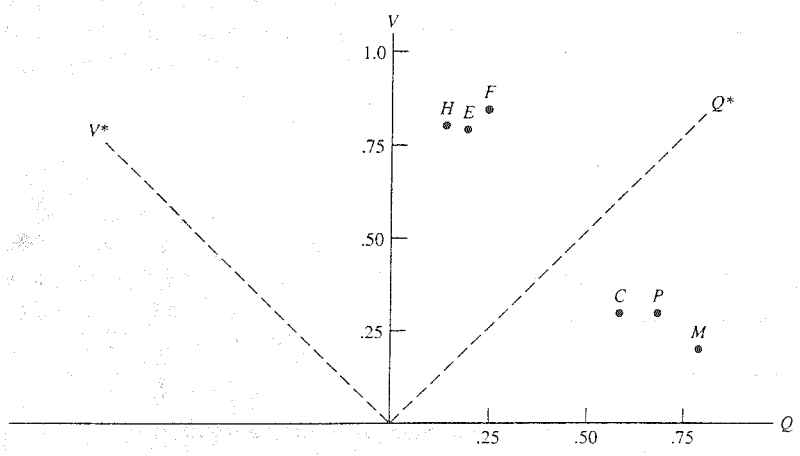


Figure 5.7 Factor solution.

decomposed into communality of the variable with each factor in an infinite number of ways. Each decomposition will result in a different factor solution. Therefore, as discussed in Section 5.1.4, one type of factor indeterminacy problem in factor analysis pertains to decomposition of the total communality, or indeterminacy due to the factor rotation problem. The factor solution given by Eq. 5.2 is plotted in Figure 5.7 where the loadings are the coordinates with respect to the Q and V axes. The factor solution given by Eq. 5.4 is equivalent to representing the loadings as coordinates with respect to axes Q^* and V^*. Note that the factor solution given by Eq. 5.4 can be viewed as a rotation problem because the two axes, Q and V, are rotated orthogonally to obtain a new set of axes, Q^* and V^*. Since we can have an infinite number of rotations, there will be an infinite number of factor solutions. The "correct" rotation is the one that gives the most plausible or acceptable interpretation of the factors.

5.3.3 More Than Two Factors

In the case of a p-indicator, m-factor model, the p vectors can be represented in an n-dimensional observation space. The p vectors will, however, lie in an $m+p$ dimensional subspace (i.e., m common factors and p unique factors). The objective once again is to identify the $m+p$ dimensions and the resulting communalities and error variances. Furthermore, the m vectors representing the m common factors can be rotated without changing the orientation of the p vectors. Of course the orientation of the m dimensions (i.e., decomposition of the total communalities) will have to be determined using other criteria.

5.4 FACTOR ANALYSIS TECHNIQUES

In this section we provide a nonmathematical discussion of the two most popular techniques: principal components factoring (PCF) and principal axis factoring (PAF) (see Harman 1976; Rummel 1970; and McDonald 1985 for a complete discussion of these and other techniques).[5] The correlation matrix given in Table 5.2 will be used for

[5]Factor analysis can be classified as exploratory or confirmatory. A discussion of the differences between the two types of factor analysis is provided later in the chapter. PCF and PAF are the most popular estimation techniques for exploratory factor analysis, and the maximum-likelihood estimation technique is the most popular technique for confirmatory factor analysis. A discussion of the maximum-likelihood estimation technique is provided in the next chapter.

illustration purposes and a sample size of n = 200 will be assumed. We will also assume that we have absolutely no knowledge about the factor model responsible for the correlations among the variables. Our objective, therefore, is to estimate the factor model responsible for the correlations among the variables.

5.4.1 Principal Components Factoring (PCF)

The first step is to provide an initial estimate of the communalities. In PCF it is assumed that the initial estimates of the communalities for all the variables are equal to one. Next, the correlation matrix with the estimated communalities in the diagonal is subjected to a principal components analysis. Exhibit 5.1 gives the SAS output for the principal components analysis. The six principal components can be represented as [2]:

$$\begin{aligned}
\xi_1 &= .368M + .391P + .372C + .432E + .422H + .456F \\
\xi_2 &= .510M + .409P + .383C - .375E - .421H - .329F \\
\xi_3 &= -.267M - .486P + .832C - .022E - .003H - .023F \\
\xi_4 &= .728M - .665P - .152C + .065E + .012H + .035F \\
\xi_5 &= .048M - .005P - .003C - .742E + .667H + .054F \\
\xi_6 &= .042M + .039P + .024C + .343E + .447H - .824F. \qquad (5.9)
\end{aligned}$$

The variances (given by the eigenvalues) of the six principal components, $\xi_1, \xi_2, \xi_3, \xi_4, \xi_5$, and ξ_6 are, respectively, 3.367, 1.194, .507, .372, .313, and .247 [1]. The above equations can be rewritten such that the principal components scores are standardized to have a variance of one. This can be done by dividing each ξ by its respective standard deviation. For example, for the first principal component

$$\frac{\xi_1}{\sqrt{3.367}} = .368M + .391P + .372C + .432E + .422H + .456F,$$

or

$$\xi_1 = .675M + .717P + .683C + .793E + .774H + .837F.$$

Exhibit 5.1

Principal components analysis for the correlation matrix of Table 5.2

1

	EIGENVALUE	DIFFERENCE	PROPORTION	CUMULATIVE
PRIN1	3.36689	2.17285	0.561149	0.56115
PRIN2	1.19404	0.68703	0.199007	0.76016
PRIN3	0.50701	0.13516	0.084501	0.84466
PRIN4	0.37185	0.05873	0.061974	0.90663
PRIN5	0.31312	0.06602	0.052186	0.95882
PRIN6	0.24709	.	0.041182	1.00000

2 EIGENVECTORS

	PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6
M	0.367802	0.509824	-.266979	0.727665	0.047857	0.041663
P	0.391381	0.409168	-.485915	-.664648	-.005389	0.038775
C	0.371982	0.382542	0.831629	-.152048	-.003335	0.023552
E	0.432206	-.374995	-.021560	0.065466	-.741529	0.343453
H	0.421900	-.421447	-.002701	0.011605	0.666963	0.446543
F	0.456476	-.328759	-.023047	0.034749	0.054439	-.823921

Standardizing each principal component results in the following equations

$$
\begin{aligned}
\xi_1 &= .675M + .717P + .683C + .793E + .774H + .837F \\
\xi_2 &= .557M + .447P + .418C - .410E - .461H - .359F \\
\xi_3 &= -.190M - .346P + .592C - .015E - .002H - .016F \\
\xi_4 &= .444M - .405P - .093C + .040E + .007H + .021F \\
\xi_5 &= .027M - .003P - .002C - .415E + .373H + .030F \\
\xi_6 &= .021M + .019P + .012C + .171E + .222H - .409F. \qquad (5.10)
\end{aligned}
$$

An alternative way of writing the preceding equations is to represent the indicators, M, P, C, E, H, and F, as functions of the six principal components, $\xi_1, \xi_2, \xi_3, \xi_4, \xi_5$, and ξ_6. It can be shown that Eq. 5.10 can be written as (see Section A5.6.1 of the Appendix):

$$
\begin{aligned}
M &= .675\xi_1 + .557\xi_2 - .190\xi_3 + .444\xi_4 + .027\xi_5 + .021\xi_6 \\
P &= .717\xi_1 + .447\xi_2 - .346\xi_3 - .405\xi_4 - .003\xi_5 + .019\xi_6 \\
C &= .683\xi_1 + .418\xi_2 + .592\xi_3 - .093\xi_4 - .002\xi_5 + .012\xi_6 \\
E &= .793\xi_1 - .410\xi_2 - .015\xi_3 + .040\xi_4 - .415\xi_5 + .171\xi_6 \\
H &= .774\xi_1 - .461\xi_2 - .002\xi_3 + .007\xi_4 + .373\xi_5 + .222\xi_6 \\
F &= .837\xi_1 - .359\xi_2 - .016\xi_3 + .021\xi_4 + .030\xi_5 - .409\xi_6. \qquad (5.11)
\end{aligned}
$$

Notice that the rows of Eq. 5.11 are the columns of Eq. 5.10 and vice versa.

The second step is to determine the number of principal components that need to be retained. As discussed in the previous chapter, the most popular rules are the eigenvalue-greater-than-one rule, the scree plot, and the parallel procedure. The eigenvalue-greater-than-one rule suggests that two principal components should be retained. Using Eq. 4.18, $\lambda_1 = 1.237, \lambda_2 = 1.105, \lambda_3 = 1.002$, and $\lambda_4 = 0.919$. Figure 5.8 gives the resulting scree plot and the plot of the eigenvalues from the parallel

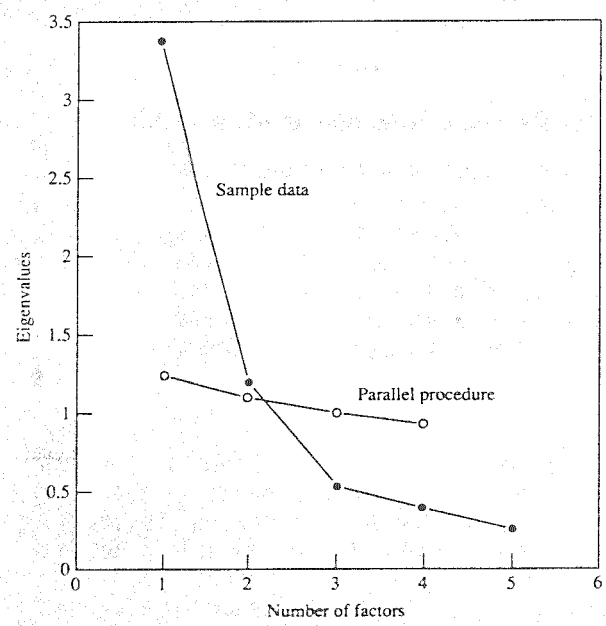


Figure 5.8 Scree plot and plot of eigenvalues from parallel analysis.

procedure. It is clear from the figure that two principal components should be retained. One way of representing the indicators as functions of two common factors and six unique factors is to modify Eq. 5.11 as follows:

$$\begin{aligned} M &= .675\xi_1 + .557\xi_2 + \epsilon_m \\ P &= .717\xi_1 + .447\xi_2 + \epsilon_p \\ C &= .683\xi_1 + .418\xi_2 + \epsilon_c \\ E &= .793\xi_1 - .410\xi_2 + \epsilon_e \\ H &= .774\xi_1 - .461\xi_2 + \epsilon_h \\ F &= .837\xi_1 - .359\xi_2 + \epsilon_f \end{aligned} \tag{5.12}$$

where

$$\begin{aligned} \epsilon_m &= -.190\xi_3 + .444\xi_4 + .027\xi_5 + .021\xi_6 \\ \epsilon_p &= -.346\xi_3 - .405\xi_4 - .003\xi_5 + .019\xi_6 \\ \epsilon_c &= .592\xi_3 - .093\xi_4 - .002\xi_5 + .012\xi_6 \\ \epsilon_e &= -.015\xi_3 + .040\xi_4 - .415\xi_5 + .171\xi_6 \\ \epsilon_h &= -.002\xi_3 + .007\xi_4 + .373\xi_5 + .222\xi_6 \\ \epsilon_f &= -.016\xi_3 + .021\xi_4 + .030\xi_5 - .409\xi_6. \end{aligned} \tag{5.13}$$

In Eq. 5.12, the principal components model has been modified to represent the original variables as the sum of two parts. The first part is a linear combination of the first two principal components, referred to as *common factors*. The second part is a sum of the remaining four components and represents the unique factor. The coefficients of Eq. 5.12 will be the pattern loadings, and because the factor model is orthogonal the pattern loadings are also the structure loadings. Table 5.4 gives a revised estimate of the communalities, and estimates of the loadings and the unique variances. The total

Table 5.4 **Summary of Principal Components Factor Analysis for the Correlation Matrix of Table 5.2**

Variable	Factor Loadings ξ_1	Factor Loadings ξ_2	Communalities	Specific Variance ϵ
M	.675	.557	.766	.234
P	.717	.447	.714	.286
C	.683	.418	.641	.359
E	.793	−.410	.797	.203
H	.774	−.461	.812	.188
F	.837	−.359	.829	.171

Notes:

1. Variance accounted for by factor ξ_1 is: 3.365 (i.e., $.675^2 + .717^2 + .683^2 + .774^2 + .837^2$).
2. Variance accounted for by factor ξ_2 is: 1.194 (i.e., $.557^2 + .447^2 + .418^2 + (-.410)^2 + (-.461)^2 + (-.359)^2$).
3. Total variance accounted for by factors ξ_1 and ξ_2 is: 4.559 (i.e., 3.365 + 1.194).
4. Total variance not accounted for by the common factors (i.e., specific variance) is: 1.441 (i.e., .234 + .286 + .359 + .203 + .188 + .171).
5. Total variance in the data is 6 (i.e., 4.559 + 1.441).

communality between all the variables and a factor is given by the eigenvalue of the factor, and is referred to as the variance explained or accounted for by the factor. That is, variances accounted for by the two factors, ξ_1 and ξ_2, are, respectively, 3.365 and 1.194. The total variance not accounted for by the common factors is the sum of the unique variances and is equal to 1.441. The amount of correlation among the indicators explained by or due to the two factors can be calculated by using the procedure described earlier in the chapter.

Table 5.5 gives the amount of correlation among the indicators that is due to the two factors and is referred to as the reproduced correlation matrix. The diagonal of the reproduced correlation matrix gives the communalities of each indicator. The table also gives the amount of correlation that is not explained by the two factors. This matrix is usually referred to as the residual correlation matrix because the diagonal contains the unique variances and the off-diagonal elements contain the differences between observed correlations and correlations explained by the estimated factor structure. Obviously, for a good factor model the residual correlations should be as small as possible. The residual matrix can be summarized by computing the square root of the average squared values of the off-diagonal elements. This quantity, known as the root mean square residual (RMSR), should be small for a good factor structure. The RMSR of the residual matrix is given by

$$RMSR = \sqrt{\frac{\sum_{i=1}^{p} \sum_{j=i}^{p} res_{ij}^2}{p(p-1)/2}} \tag{5.14}$$

Table 5.5 **Reproduced and Residual Correlation Matrices for PCF**

Reproduced Correlation Matrix

	M	P	C	E	H	F
M	.766	.733	.694	.307	.266	.365
P	.733	.714	.677	.385	.349	.440
C	.694	.677	.641	.370	.336	.422
E	.307	.385	.370	.797	.803	.811
H	.266	.349	.336	.803	.812	.813
F	.365	.440	.422	.811	.813	.829

Note: Communalities are on the diagonal.

Residual Correlation Matrix

	M	P	C	E	H	F
M	.234	−.113	−.154	.013	.018	.005
P	−.113	.285	−.167	−.005	.002	−.010
C	−.154	−.167	.359	−.010	.000	−.017
E	.013	−.005	−.010	.203	−.117	−.081
H	.018	.002	.000	−.117	.188	.079
F	.005	−.010	−.017	−.081	.078	.171

Note: Unique variances are on the diagonal.

Root mean square residual (RMSR) = .078.

where res_{ij} is the correlation between the ith and jth variables and p is the number of variables. The RMSR for the residual matrix given in Table 5.5 is equal to .078 which appears to be small implying a good factor solution.

It is clear that PCF is essentially principal components analysis where it is assumed that estimates of the communalities are one. That is, it is assumed that there are no unique factors and the number of components is equal to the number of variables. It is hoped that a few components would account for a major proportion of the variance in the data and these components are considered to be common factors. The variance that is in common between each variable and the common components is assumed to be the communality of the variable, and the variance of each variable that is in common with the remaining factors is assumed to be the error or unique variance of the variable. In the example presented here, the first two components are assumed to be the two common factors and the remaining components are assumed to represent the unique factors.

5.4.2 Principal Axis Factoring

In principal axis factoring (PAF) an attempt is made to estimate the communalities. An iterative procedure is used to estimate the communalities and the factor solution. The iterative procedure continues until the estimates of the communalities converge. The iteration process is described below.

Step 1. First, it is assumed that the prior estimates of the communalities are one. A PCF solution is then obtained. Based on the number of components (factors) retained, estimates of structure or pattern loadings are obtained which are then used to reestimate the communalities. The factor solution thus obtained has been described in the previous section.

Step 2. The maximum change in estimated communalities is computed. It is defined as the maximum difference between previous and revised estimates of the communality for each variable. For the solution given in the previous section, the maximum change in communality is for indicator C, and is equal to .359 (i.e., $1 - .641$). Note that it was assumed that the previous estimates of communalities are one.

Step 3. If the maximum change in communality is greater than a predetermined convergence criterion, then the original correlation matrix is modified by replacing the diagonals with the new estimated communalities. A new principal components analysis is done on the modified correlation matrix and the procedure described in Step 2 is repeated. Steps 2 and 3 are repeated until the change in the estimated communalities is less than the convergence criterion.

Table 5.6 gives the iteration history for PAF analysis of the correlation matrix given in Table 5.2. Assuming a convergence criterion of .001, nine iterations are required for the estimates of the communalities to converge. The solution after the first iteration has been discussed in the previous section. The solution in the second iteration is obtained by using the modified correlation matrix in which the diagonals contain the communalities estimated in the first iteration; solution for the third iteration is obtained by using the modified correlation matrix in which the diagonals contain the communalities obtained from the second iteration, and so on.

Table 5.6 **Iteration History for Principal Axis Factor Analysis**

		Communalities					
Iteration	Change	M	P	C	E	H	F
1	.359	.766	.714	.641	.797	.812	.829
2	.128	.698	.626	.513	.725	.744	.784
3	.042	.679	.598	.471	.698	.719	.774
4	.014	.675	.588	.457	.688	.708	.774
5	.005	.674	.585	.453	.684	.703	.776
6	.003	.675	.583	.451	.682	.700	.779
7	.002	.676	.582	.451	.681	.698	.781
8	.001	.677	.582	.451	.681	.697	.782
9	.001	.677	.581	.450	.680	.697	.783

Notes:

1. Maximum change in communality in iteration 1 is for variable C and is equal to .359 (i.e., 1 − .641).
2. Maximum change in communality in iteration 2 is also for variable C and is equal to .128 (i.e., .641 − .513).

5.4.3 Which Technique Is the Best?

In most cases, fortunately, there is very little difference between the results of PCF and PAF.[6] Therefore, in most cases it really does not matter which of the two techniques is used. However, there are conceptual differences between the two techniques.

In PCF it is assumed that the communalities are one and consequently no prior estimates of communalities are needed. This assumption, however, implies that a given variable is not composed of common and unique parts. The variance of a given variable is completely accounted for by the p principal components. It is, however, hoped that a few principal components would account for a major proportion of a variable's variance. These principal components are labeled as *common factors* and the accounted-for variance is labeled as the variable's communality. The remaining principal components are considered to be nuisance components and are lumped together into a single component labeled as the unique factor, and the variance in common with it is called the variable's unique or error variance. Therefore, strictly speaking, PCF is simply principal components analysis and not factor analysis.

PAF, on the other hand, implicitly assumes that a variable is composed of a common part and a unique part, and the common part is due to the presence of the common factors. The objectives are to first estimate the communalities and then identify the common factors responsible for the communalities and the correlation among the variables. That is, the PAF technique assumes an implicit underlying factor model. For this reason many researchers choose to use PAF.

5.4.4 Other Estimation Techniques

Other estimation techniques, besides the above two techniques, have also been proposed in the factor analysis literature. These techniques differ mainly with respect to how the communalities of the variables are estimated. We provide only a brief discussion

[6]Theoretically, the results will be identical if the true values of the communalities approach one.

of these techniques. The interested reader is referred to Harman (1976) and Rummel (1970) for further details.

Image Analysis

In image analysis, a technique proposed by Guttman (1953), the communality of a variable is ascribed a precise meaning. Communality of a variable is defined as the square of the multiple correlation obtained by regressing the variable on the remaining variables. That is, there is no indeterminacy due to the estimation of the communality problem. The squared multiple correlations are inserted in the diagonal of the correlation matrix and the off-diagonal values of the matrix are adjusted so that none of the eigenvalues are negative. Image factor analysis can be done using SAS and SPSS.

Alpha Factor Analysis

In alpha factor analysis it is assumed that the data are the population, and the variables are a sample from a population of variables. The objective is to determine if inferences about the factor solution using a sample of variables holds for the population of variables. That is, the objective is not to make statistical inferences, but to generalize the results of the study to a population of variables. Alpha factor analysis can be done using SAS and SPSS.

5.5 HOW TO PERFORM FACTOR ANALYSIS

A number of statistical packages such as SPSS and SAS can be used to perform factor analysis. We will use SAS to do a PAF analysis on the correlation matrix given in Table 5.2. For illustration purposes a sample size of n = 200 is assumed. Once again, it is assumed that we have no knowledge about the factor model that generated the correlation matrix. Table 5.7 gives the necessary SAS commands. Following is a brief discussion of the commands; however, the reader should consult the SAS manual for details.

The commands before the PROC FACTOR procedure are basic SAS commands for reading a correlation matrix. The METHOD option specifies that the analytic procedure PRINIT (which is PAF) should be used to extract the factors.[7] The ROTATE = V option

Table 5.7 **SAS Commands**

```
TITLE PRINCIPAL AXIS FACTORING FOR THE CORRELATION MATRIX
  OF TABLE 5.2;
DATA CORRMATR(TYPE=CORR);
  INPUT M P C E H F;
  _TYPE_='CORR';
CARDS;
insert correlation matrix here
;
PROC FACTOR METHOD=PRINIT ROTATE=V CORR MSA SCREE
  RESIDUALS PREPLOT PLOT;
  VAR M P C E H F;
```

[7]PRINIT stands for principal components analysis with iterations.

specifies that varimax rotation, which is explained in Section 5.6.6, should be used for obtaining a unique solution. CORR, MSA, SCREE, RESIDUALS, PREPLOT, and PLOT are the options for obtaining the desired output.

5.6 INTERPRETATION OF SAS OUTPUT

Exhibit 5.2 gives the SAS output for PAF analysis of the correlation matrix given in Table 5.2. The output is labeled to facilitate the discussion.

Exhibit 5.2

Principal axis factoring for the correlation matrix of Table 5.2

1 CORRELATIONS

	M	P	C	E	H	F
M	1.00000	0.62000	0.54000	0.32000	0.28400	0.37000
P	0.62000	1.00000	0.51000	0.38000	0.35100	0.43000
C	0.54000	0.51000	1.00000	0.36000	0.33600	0.40500
E	0.32000	0.38000	0.36000	1.00000	0.68600	0.73000
H	0.28400	0.35100	0.33600	0.68600	1.00000	0.73450
F	0.37000	0.43000	0.40500	0.73000	0.73450	1.00000

INITIAL FACTOR METHOD: ITERATED PRINCIPAL FACTOR ANALYSIS

2 PARTIAL CORRELATIONS CONTROLLING ALL OTHER VARIABLES

	M	P	C	E	H	F
M	1.00000	0.44624	0.30877	0.01369	-0.03195	0.06094
P	0.44624	1.00000	0.20253	0.05109	0.02594	0.09912
C	0.30877	0.20253	1.00000	0.04784	0.03159	0.08637
E	0.01369	0.05109	0.04784	1.00000	0.31767	0.41630
H	-0.03195	0.02594	0.03159	0.31767	1.00000	0.45049
F	0.06094	0.09912	0.08637	0.41630	0.45049	1.00000

3 KAISER'S MEASURE OF SAMPLING ADEQUACY: OVER-ALL MSA = 0.81299762

M	P	C	E	H	F
0.768873	0.812439	0.866916	0.831666	0.812326	0.796856

PRIOR COMMUNALITY ESTIMATES: ONE

4 PRELIMINARY EIGENVALUES: TOTAL = 6 AVERAGE = 1

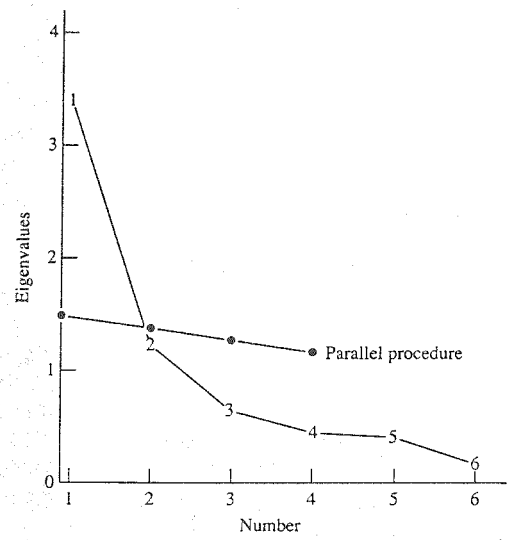
	1	2	3	4	5	6
EIGENVALUE	3.366893	1.194041	0.507006	0.371847	0.313119	0.247095
DIFFERENCE	2.172853	0.687035	0.135159	0.058728	0.066024	
PROPORTION	0.5611	0.1990	0.0845	0.0620	0.0522	0.0412
CUMULATIVE	0.5611	0.7602	0.8447	0.9066	0.9588	1.0000

(continued)

Exhibit 5.2 (continued)

4a 2 FACTORS WILL BE RETAINED BY THE MINEIGEN CRITERION

SCREE PLOT OF EIGENVALUES



5

ITER	CHANGE	COMMUNALITIES					
1	0.359385	0.76582	0.71564	0.64061	0.79685	0.81139	0.83061
2	0.127701	0.69839	0.62622	0.51291	0.72453	0.74431	0.78351
3	0.042178	0.67947	0.59762	0.47073	0.69818	0.71876	0.77359
4	0.013511	0.67488	0.58806	0.45722	0.68812	0.70800	0.77395
5	0.005153	0.67444	0.58455	0.45287	0.68398	0.70285	0.77646
6	0.002809	0.67510	0.58304	0.45140	0.68212	0.70004	0.77888
7	0.001871	0.67594	0.58224	0.45084	0.68120	0.69834	0.78075
8	0.001338	0.67671	0.58173	0.45059	0.68071	0.69725	0.78209
9	0.000928	0.67735	0.58136	0.45045	0.68043	0.69652	0.78302

CONVERGENCE CRITERION SATISFIED

6 EIGENVALUES OF THE REDUCED CORRELATION MATRIX: TOTAL = 3.86907

AVERAGE = 0.644845

	1	2	3	4	5	6
EIGENVALUE	3.028093	0.841027	0.001562	0.001118	-0.001222	-0.001508
DIFFERENCE	2.187066	0.839465	0.000444	0.002340	0.000285	
PROPORTION	0.7826	0.2174	0.0004	0.0003	-0.0003	-0.0004
CUMULATIVE	0.7826	1.0000	1.0004	1.0007	1.0004	1.0000

7 FACTOR PATTERN

	FACTOR1	FACTOR2
M	0.63584	0.52255
P	0.65784	0.38549
C	0.59812	0.30447
E	0.76233	-0.31509
H	0.74908	-0.36797
F	0.83129	-0.30329

(continued)

Exhibit 5.2 (continued)

VARIANCE EXPLAINED BY EACH FACTOR

FACTOR1	FACTOR2
3.028093	0.841027

8 FINAL COMMUNALITY ESTIMATES: TOTAL = 3.869120

M	P	C	E	H	F
0.677354	0.581356	0.450447	0.680426	0.696517	0.783020

9 RESIDUAL CORRELATIONS WITH UNIQUENESS ON THE DIAGONAL

	M	P	C	E	H	F
M	0.32265	0.00028	0.00059	-0.00007	-0.00001	-0.00008
P	0.00028	0.41864	-0.00084	-0.00003	0.00007	0.00006
C	0.00059	-0.00084	0.54955	-0.00003	-0.00000	0.00013
E	-0.00007	-0.00003	-0.00003	0.31957	-0.00099	0.00072
H	-0.00001	0.00007	-0.00000	-0.00099	0.30348	0.00020
F	-0.00008	0.00006	0.00013	0.00072	0.00020	0.21698

9a ROOT MEAN SQUARE OFF-DIAGONAL RESIDUALS: OVER-ALL = 0.00042458

M	P	C	E	H	F
0.000297	0.000397	0.000462	0.000548	0.000451	0.000345

10 PARTIAL CORRELATIONS CONTROLLING FACTORS

	M	P	C	E	H	F
M	1.00000	0.00076	0.00141	-0.00021	-0.00004	-0.00030
P	0.00076	1.00000	-0.00174	-0.00008	0.00020	0.00019
C	0.00141	-0.00174	1.00000	-0.00007	-0.00001	0.00039
E	-0.00021	-0.00008	-0.00007	1.00000	-0.00317	0.00275
H	-0.00004	0.00020	-0.00001	-0.00317	1.00000	0.00079
F	-0.00030	0.00019	0.00039	0.00275	0.00079	1.00000

ROOT MEAN SQUARE OFF-DIAGONAL PARTIALS: OVER-ALL = 0.00126957

M	P	C	E	H	F
0.000734	0.000860	0.001017	0.001878	0.001462	0.001301

(continued)

Exhibit 5.2 (continued)

11 PLOT OF FACTOR PATTERN FOR FACTOR1 AND FACTOR2

```
                                FACTOR1
                                   1

                                  .9
                          F
                                  .8
                        E D
                                  .7
                                        B   A
                                  .6  C

                                  .5

                                  .4

                                  .3

                                  .2
                                                              F
                                  .1                          A
                                                              C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                              O
                                 -.1                          R
                                                              2
                                 -.2

                                 -.3

                                 -.4

                                 -.5

                                 -.6

                                 -.7

                                 -.8

                                 -.9

                                  -1

M     =A   P     =B   C     =C   E     =D   H     =E   F     =F
```

(continued)

Exhibit 5.2 (continued)

ROTATION METHOD: VARIMAX

12 ORTHOGONAL TRANSFORMATION MATRIX

	1	2
1	0.76668	0.64202
2	-0.64202	0.76668

13a ROTATED FACTOR PATTERN

	FACTOR1	FACTOR2
M	0.15200	0.80886
P	0.25687	0.71790
C	0.26309	0.61744
E	0.78676	0.24786
H	0.81055	0.19881
F	0.83205	0.30118

VARIANCE EXPLAINED BY EACH FACTOR

FACTOR1	FACTOR2
2.126595	1.742525

FINAL COMMUNALITY ESTIMATES: TOTAL = 3.869120

M	P	C	E	H	F
0.677354	0.581356	0.450447	0.680426	0.696517	0.783020

13b SCORING COEFFICIENTS ESTIMATED BY REGRESSION

SQUARED MULTIPLE CORRELATIONS OF THE VARIABLES WITH EACH FACTOR

FACTOR1	FACTOR2
0.848437	0.769784

13c STANDARDIZED SCORING COEFFICIENTS

	FACTOR1	FACTOR2
M	-0.15607	0.53363
P	-0.06286	0.33912
C	-0.02954	0.21510
E	0.30267	-0.04518
H	0.34597	-0.09202
F	0.45373	-0.02863

(continued)

Exhibit 5.2 (continued)

```
ROTATION METHOD: VARIMAX

13d PLOT OF FACTOR PATTERN FOR FACTOR1  AND FACTOR2

                                FACTOR1
                                   1

                                  .9
                                            F
                                  .8     E D

                                  .7

                                  .6

                                  .5

                                  .4

                                  .3
                                                   C  B
                                  .2
                                                         A       F
                                  .1                             A
                                                                 C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                                 O
                                 -.1                             R
                                                                 2
                                 -.2

                                 -.3

                                 -.4

                                 -.5

                                 -.6

                                 -.7

                                 -.8

                                 -.9

                                  -1

M      =A    P      =B    C      =C    E      =D    H      =E    F      =F
```

5.6.1 Are the Data Appropriate for Factor Analysis?

The first decision the researcher faces is whether or not the data are appropriate for factor analysis. A number of measures are used for this purpose. This part of the output provides some of the measures. It should be noted that the measures to be discussed are basically heuristics or rules of thumb.

First, one can subjectively examine the correlation matrix. High correlations among the variables indicate that the variables can be grouped into homogeneous sets of variables such that each set of variables measures the same underlying constructs or dimensions. Low correlations among the variables indicate that the variables do not have much in common or are a group of heterogeneous variables. An examination of the correlation matrix in Exhibit 5.2 indicates that there are two groups or sets of variables that have high correlations among themselves [1]. In this sense, one could view factor analysis as a technique that tries to identify groups or clusters of variables such that variables in each group are indicators of a common trait or factor. This suggests that the correlation matrix is appropriate for factoring. However, visual examination of the correlation matrix for a large number of variables is almost impossible, and hence this rule may not be appropriate when there are many variables.

Second, one can examine the partial correlations controlling for all other variables. These correlations, also referred to as negative anti-image correlations, should be small for the correlation matrix to be appropriate for factoring. However, how small is "small" is essentially a judgmental question. It appears that the partial correlations are small, but one can easily take issue with this conclusion [2].

Third, one can examine Kaiser's measure of overall sampling adequacy and a measure of the sampling adequacy for each indicator. This measure, the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (Kaiser 1970), is a popular diagnostic measure. KMO provides a means to assess the extent to which the indicators of a construct belong together. That is, it is a measure of the homogeneity of variables. Although there are no statistical tests for the KMO measure, the following guidelines are suggested by Kaiser and Rice (1974).

KMO Measure	Recommendation
≥ .90	Marvelous
.80+	Meritorious
.70+	Middling
.60+	Mediocre
.50+	Miserable
Below .50	Unacceptable

Obviously a higher value of KMO is desired. It is suggested that the overall KMO measure should be greater than .80; however, a measure of above .60 is tolerable. The overall KMO measure can sometimes be increased by deleting the offending variables whose KMO value is low. An overall value of .813 for the KMO measure suggests that the correlation matrix is appropriate for factoring [3].

5.6.2 How Many Factors?

The next step is to determine the number of factors needed to explain correlations among the variables. The issue is very similar to determining the number of principal components that should be retained in principal components analysis. The most popular

heuristics are the eigenvalue-greater-than-one rule and the scree plot. Unless otherwise specified, SAS and SPSS use the eigenvalue-greater-than-one rule for extracting the number of factors. However, as suggested by Cliff (1988), caution is advised about relying exclusively on the eigenvalue-greater-than-one rule for determining the appropriate number of factors. Results of the simulation studies conducted by Zwick and Velicer (1986) found that the best-performing rules were the minimum average partial correlation (MAP), parallel analysis, and the scree plot. The MAP, however, mostly performed well for large numbers of indicators per factor.

Parallel analysis, discussed in the previous chapter, is recommended, along with the interpretability of the resulting factors for determining the number of factors. Indeed, interpretability of the factors should be one of the important criteria in determining the number of factors. The eigenvalues resulting from parallel analysis can be estimated using Eq. 4.18. The estimated eigenvalues are: $\lambda_1 = 1.237$; $\lambda_2 = 1.105$; $\lambda_3 = 1.002$; $\lambda_4 = .919$; $\lambda_5 = 0$; and $\lambda_6 = 0$.[8] These eigenvalues are plotted on the scree plot in Exhibit 5.2 [4a]. The scree plot and the parallel procedure plot suggest a two-factor solution.[9] Interpretation of the two extracted factors is provided later.

5.6.3 The Factor Solution

Next, the output gives the factor solution. The iteration history for the PAF [5] is the same, within rounding errors, as that given in Table 5.6. Note that nine iterations are required for the solution to converge. At each iteration, the output gives the communalities of each variable and the maximum change in the communality. Default values for the convergence criterion and the number of iterations in SAS are, respectively, .001 and 30. The user can increase the number of iterations if convergence is not achieved in 30 iterations. However, caution is advised when more iterations are required as that might suggest that the data may not be suitable for factor analysis.

The factor pattern matrix gives the pattern or structure loadings [7]. Note that the estimated pattern loadings are not the same as those reported in Table 5.2, due to the rotation problem described earlier.[10] As discussed previously, the square of the pattern loadings gives the variable's communality. For example, the communality of variable *M* with *Factor1* is .404 (i.e., $.636^2$) and with *Factor2* is .274 (i.e., $.523^2$), where .636 and .523 are pattern loadings of variable *M* with *Factor1* and *Factor2*, respectively [7]. The total communality of the variable will be .678 (i.e., .274 + .404). The output gives the total or final communalities of each variable which, within rounding error, are the same as those given in Table 5.2 and for the last iteration of Table 5.6 [8].

The sum of the squared pattern loadings for a given factor is the communality of all the variables with that factor and is given by the eigenvalue of the factor. The eigenvalues of the factors are reported in the output as the eigenvalues of the reduced (i.e., modified) correlation matrix [6]. Recall that the modified correlation matrix is one where the diagonals contain the estimated communalities. As discussed earlier, the factor solution is not unique because of the rotation problem. That is, the factor pattern loadings are not unique, and, therefore, the variance in common between the factor and the variables is also not unique. Consequently, the variance in common between the factor and the

[8] Actually, the estimated values for λ_5 and λ_6 are negative, implying that they are equal to zero.

[9] Note that this is consistent with the *a priori* knowledge that two factors are responsible for the correlation among the indicators.

[10] As the factor model is orthogonal, the pattern and the structure loadings are the same.

variables is not a very meaningful measure of factor importance unless constraints are imposed to obtain a unique solution. It should be emphasized here that the main objective of factor analysis is to explain the intercorrelations among the variables and not to account for the total variation in the data.

5.6.4 How Good Is the Factor Solution?

The next step is to assess the estimated factor solution. That is, how well can the factors account for the correlations among the indicators? The residual correlation matrix can be used for this purpose [9]. The residuals are all small and the RMSR is .0004, indicating that the final factor structure explains most of the correlations among the indicators. Comparison of this RMSR with the RMSR of .078 for the factor solution obtained from the PCF method suggests that the factor solution obtained from the PAF method does a better job of explaining the correlations among the variables than the factor solution from the PCF method. The RMSRs for each of the variables are also low [9a].

One can also examine the correlation among the indicators after the effect of the factors has been partialled out. It is obvious that for a good factor solution the resulting partial correlations should be close to zero, because once the effect of the common factors has been removed there is nothing to link the indicators. The overall RMSR for the partial correlations is .001 and is considered to be small [10]. To conclude, the RMSRs of the residual and the partial correlation matrices suggest that the estimated factor model is appropriate.

5.6.5 What Do the Factors Represent?

The next and perhaps the most important question is: What do the factors represent? In other words, what are the underlying dimensions that account for the correlation among the variables? Simply put, we have to attach labels or meanings to the factors. Variable loadings and researcher's knowledge about the variables are used for interpreting the factors. As discussed earlier, high loading of a variable on a factor indicates that there is much in common between the factor and the respective variable. Although there are no definite cutoff points to tell us how high is "high," it has been suggested that the loadings should at least be greater than .60, and many researchers have used cutoff values as low as .40. It can be clearly seen from the factor pattern matrix in Exhibit 5.2 that all the variables have high loadings on the first factor [7]. This suggests that the first factor might represent subjects' general intelligence levels.

None of the variables load highly on the second factor but there is a clear pattern to the signs of the loadings [7]. Loadings of variables M, P, and C have a positive sign and loadings of variables E, H, and F have a negative sign. One might hypothesize that the second factor distinguishes between courses that require quantitative ability from courses that require verbal ability. Therefore, the second factor might be labeled as the quantitative/verbal ability factor.

This interpretation of the factors can also be reached by plotting the variables in the factor space. The output provides a plot of the factor structure [11]. It is a plot of the variables in the factor space with the respective loadings as the coordinates and is very similar to the plot given in Figure 5.7. Note that indicators M, P, and C (labeled A, B, and C, respectively, by SAS) are close to each other, as are the indicators E, H, and F (labeled D, E, and F, respectively, by SAS). Both sets of variables are closer to *Factor1*

than *Factor2*; however, the projections of variables M, P, and C and variables E, H, and F on *Factor2* will have different signs (i.e., the loadings will have different signs). Therefore, as before, *Factor1* can be interpreted as a general factor and *Factor2* as a quantitative/verbal ability factor.

If the preceding interpretation of the factors does not appear plausible or theoretically defendable then one can seek alternative solutions that would result in a better interpretation of the factor model. And since the factor solution is not unique, one can obtain another factor solution by rotating the axes. The objective of rotation is to obtain another solution that will provide a "better" representation of the factor structure.[11] A number of analytical techniques have been developed to obtain a new set of axes that might provide a better interpretation of the factor structure. Most of these methods impose certain mathematical constraints on the rotation in order to obtain a unique solution.

5.6.6 Rotation

The objective of rotation is to achieve a simpler factor structure that can be meaningfully interpreted by the researcher. An orthogonal or an oblique rotation can be performed to achieve this objective. In the orthogonal rotation, which is the most popular, the rotated factors are orthogonal to each other; whereas in oblique rotation the rotated factors are not orthogonal to each other. The interpretation of the factor structure resulting from an oblique rotation is more complex than that resulting from orthogonal rotations. Since oblique rotations are not used commonly, they are discussed in the Appendix. Varimax and quartimax are the most popular types of orthogonal rotations. The factor structure was rotated using each rotation technique. Discussion and the results of varimax and quartimax rotation are described below.

Varimax Rotation

In the varimax rotation the major objective is to have a factor structure in which each variable loads highly on one and only one factor. That is, a given variable should have a high loading on one factor and near zero loadings on other factors. Such a factor structure will result in each factor representing a distinct construct.

The output gives the rotated factor solution. The transformation matrix gives the weights of the equations used to represent the coordinates with respect to the new axes [12]. For example, the following equations can be used to obtain the coordinates (loadings) of the variables with respect to the new axes (factors):

$$l_1^* = .767l_1 - .642l_2$$
$$l_2^* = .642l_1 + .767l_2,$$

where l_j and l_j^* are, respectively, loadings of the jth variable with respect to the old and rotated factors (axes). The output provides the rotated pattern loadings and a plot of the rotated factor structure [13a,d]. It can be clearly seen that the variables M, P, and C load highly on the second factor, and lie close to the axis representing *Factor2*. And variables E, H, and F load high on the first factor and lie close to the axis representing *Factor1*. Therefore, the first factor represents the verbal ability and the second

[11] Recall the rotation problem discussed in Sections 5.1.4 and 5.3.2.

factor represents the quantitative ability. The factor pattern loadings are very similar to those given in Table 5.2.[12] However, note that the communality estimates of each variable and, therefore, the estimates for total communality are the same as those for the unrotated solution.

The output also gives the standardized weights or scoring coefficients that can be used for computing the factor scores [13c]. The equations for computing the factor scores are

$$\xi_1 = -.156M - .063P - .030C + .303E + .346H + .454F$$
$$\xi_2 = .534M + .339P + .215C - .045E - .092H - .029F,$$

where ξ_1 and ξ_2, respectively, are *Factor1* and *Factor2*. A number of different approaches are used to estimate the factor coefficients. The multiple regression approach is one such approach, and is discussed in the Appendix.

From the above equations it can be seen that each factor is a linear combination of the variables. The squared multiple correlation of each equation represents the amount of variance that is in common between all the variables and the respective factor, and is used to determine the ability of the variables to measure or represent the respective factor. In other words, squared multiple correlation simply represents the extent to which the variables or indicators are good measures of a given construct. Obviously, the squared multiple correlations should be high. Many researchers have considered values greater than 0.60 as high; however, once again, how high is "high" is subject to debate. For the present example, squared multiple correlations of 0.848 and 0.770, respectively, for *Factor1* and *Factor2* seem to be high [13b].

Quartimax Rotation

The major objective of this rotation technique is to obtain a pattern of loadings such that:

- All the variables have a fairly high loading on one factor.
- Each variable should have a high loading on one other factor and near zero loadings on the remaining factors.

Obviously such a factor structure will represent one factor that might be considered as an overall factor and other factors that might be specific constructs. Thus, quartimax rotation will be most appropriate when the researcher suspects the presence of a general factor. Varimax rotation destroys or suppresses the general factor and should not be used when the presence of a general factor is suspected.

Quartimax rotation of the factor solution can be obtained by specifying ROTATE = QUARTIMAX in the corresponding SAS command given in Table 5.7. Exhibit 5.3 gives only that portion of the SAS output containing quartimax rotation results. The quartimax rotation gives an interpretation of the factor structure similar to that of varimax rotation. However, this may not be true for other data sets. In general, one should use the rotation that results in a meaningful factor structure consistent with theoretical expectations. Again, note that the communality estimates of the variables are not affected.

[12]They are not exactly the same due to the indeterminacy problem.

5.7 AN EMPIRICAL ILLUSTRATION

Consider the following example.[13] The product manager of a consumer packaged goods firm is interested in identifying the major underlying factors or dimensions that consumers use to evaluate various detergents in the marketplace. These factors are assumed to be latent; however, management believes that the various attributes or properties of detergents are indicators of these underlying factors. Factor analysis can be used to identify these underlying factors. A study is conducted in which 143 respondents rated three brands of detergents on 12 product attributes using a five-point semantic differential scale. Following is an example of a semantic differential scale to elicit subjects' response for the detergent's ability to get dirt out.

Gets dirt out — — — — — Does not get dirt out

Table 5.8 gives the list of 12 product attributes and Table 5.9 gives the correlation matrix among the twelve attributes.

Exhibit 5.3

Quartimax rotation

```
ROTATION METHOD: QUARTIMAX

ORTHOGONAL TRANSFORMATION MATRIX

            1          2

 1       0.77365    0.63361
 2      -0.63361    0.77365

ROTATED FACTOR PATTERN

            FACTOR1    FACTOR2

M           0.16082    0.80715
P           0.26469    0.71505
C           0.26982    0.61453
E           0.78942    0.23925
H           0.81267    0.18995
F           0.83529    0.29207

VARIANCE EXPLAINED BY EACH FACTOR

  FACTOR1    FACTOR2
  2.150071   1.719049

FINAL COMMUNALITY ESTIMATES: TOTAL =    3.869120

          M          P          C          E          H          F
  0.677354   0.581356   0.450447   0.680426   0.696517   0.783020
```

(continued)

[13]This example is adapted from Urban and Hauser (1993).

Exhibit 5.3 (continued)

```
ROTATION METHOD: QUARTIMAX

PLOT OF FACTOR PATTERN FOR FACTOR1  AND FACTOR2

                                FACTOR1
                                   1

                                  .9
                                          F
                                  .8    E D

                                  .7

                                  .6

                                  .5

                                  .4

                                  .3
                                                C  B
                                  .2
                                                      A       F
                                  .1                          A
                                                              C
-1 -.9-.8-.7-.6-.5-.4-.3-.2-.1  0 .1 .2 .3 .4 .5 .6 .7 .8 .9 1.0T
                                                              O
                                 -.1                          R
                                                              2
                                 -.2

                                 -.3

                                 -.4

                                 -.5

                                 -.6

                                 -.7

                                 -.8

                                 -.9

                                  -1

M     =A    P     =B    C     =C    E     =D    H     =E    F     =F
```

Table 5.8 **List of Attributes**

V1:	Gentle to natural fabrics
V2:	Won't harm colors
V3:	Won't harm synthetics
V4:	Safe for lingerie
V5:	Strong, powerful
V6:	Gets dirt out
V7:	Makes colors bright
V8:	Removes grease stains
V9:	Good for greasy oil
V10:	Pleasant fragrance
V11:	Removes collar soil
V12:	Removes stubborn stains

An examination of the correlation matrix indicates that high correlations exist among the variables, ranging from a low of .17 to a high of .72. Because of the large number of variables, further examination of the correlation is not feasible.

In order to show the type of output that results from SPSS, we will use the PAF procedure in SPSS to extract the factor structure. The PAF procedure is the same as the PRINIT procedure in SAS. Table 5.10 gives the SPSS commands. Following is a brief discussion of the commands. The commands before the FACTOR command are the basic SPSS commands for reading the correlation matrix. The VARIABLES subcommand specifies the list of variables from which variables for conducting factor analysis are selected. The ANALYSIS subcommand specifies the variables that should be used for factor analysis. The EXTRACTION subcommand specifies the extraction procedure to be used. The PRINT and the PLOT subcommands specify the printed output that is desired. The ROTATION subcommand specifies the type of rotation that should be used to obtain a unique solution. Exhibit 5.4 gives the partial SPSS output. The following section discusses the various parts of the indicated output.

5.7.1 Identifying and Evaluating the Factor Solution

An overall KMO measure of .90 is quite high suggesting that the data are appropriate for factor analysis [1]. SPSS provides the Bartlett's test, which is a statistical test to assess whether or not the correlation matrix is appropriate for factoring. The Bartlett's test examines the extent to which the correlation matrix departs from orthogonality. An orthogonal correlation matrix will have a determinant of one, indicating that the variables are not correlated. On the other hand, if there is a perfect correlation between two or more variables the determinant will be zero. For the present data set, the Bartlett's test statistic is highly significant ($p < .00000$), implying that the correlation matrix is not orthogonal (i.e., the variables are correlated among themselves) and is, therefore, appropriate for factoring [1]. However, as discussed in Chapter 4, the Bartlett's test is rarely used because it is sensitive to sample size; that is, for large samples one is liable to conclude that the correlation matrix departs from orthogonality even when the correlations among the variables are small. Overall, though, it appears that the data are appropriate for factoring.

Table 5.9 **Correlation Matrix for Detergent Study**

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
V1	1.00000	0.41901	0.51840	0.56641	0.18122	0.17454	0.23034	0.30647	0.24051	0.21192	0.27443	0.20694
V2	0.41901	1.00000	0.57599	0.49886	0.18666	0.24648	0.22907	0.22526	0.21967	0.25879	0.32132	0.25853
V3	0.51840	0.57599	1.00000	0.64325	0.29080	0.34428	0.41083	0.34028	0.32854	0.38828	0.39433	0.36712
V4	0.56641	0.49886	0.64325	1.00000	0.38360	0.39637	0.37699	0.40391	0.42337	0.36564	0.33691	0.36734
V5	0.18122	0.18666	0.29080	0.38360	1.00000	0.57915	0.59400	0.67623	0.69269	0.43873	0.55485	0.65261
V6	0.17454	0.24648	0.34428	0.39637	0.57915	1.00000	0.57756	0.70103	0.62280	0.62174	0.59855	0.57845
V7	0.23034	0.22907	0.41083	0.37699	0.59400	0.57756	1.00000	0.67682	0.68445	0.54175	0.78361	0.63889
V8	0.30647	0.22526	0.34028	0.40391	0.67623	0.70103	0.67682	1.00000	0.69813	0.68589	0.71115	0.71891
V9	0.24051	0.21967	0.32854	0.42337	0.69269	0.62280	0.68445	0.69813	1.00000	0.58579	0.64637	0.69111
V10	0.21192	0.25879	0.38828	0.36564	0.43873	0.62174	0.54175	0.68589	0.58579	1.00000	0.62250	0.63494
V11	0.27443	0.32132	0.39433	0.33691	0.55485	0.59855	0.78361	0.71115	0.64637	0.62250	1.00000	0.63973
V12	0.20694	0.25853	0.36712	0.36734	0.65261	0.57845	0.63889	0.71891	0.69111	0.63494	0.63973	1.00000

Table 5.10 **SPSS Commands**

```
MATRIX DATA VARIABLES=V1 TO V12/CONTENTS=CORR/N=143/FORMAT=FULL
BEGIN DATA
insert data here
END DATA
FACTOR
 /MATRIX=IN(COR=*)
 /ANALYSIS=V1, TO, V12
 /EXTRACTION=PAF
 /ROTATION=VARIMAX
 /PRINT=INITIAL EXTRACTION ROTATION REPR KMO
 /PLOT=EIGEN ROTATION(1,2)
FINISH
```

The estimated eigenvalues (see Eq. 4.18) for the parallel procedure are plotted on the scree plot [2a]. The eigenvalue-greater-than-one rule, the scree plot, and the parallel procedure suggest that there are two factors [2, 2a]. Unless otherwise specified the SPSS program also uses the eigenvalue-greater-than-one rule for extracting the number of factors. A total of 7 iterations were required for the PAF solution to converge [2a].

Instead of the RMSR, SPSS indicates how many residual correlations are above .05. Out of a possible 66 residual correlations only 9 or 13% are greater than .05, suggesting that the factor structure adequately accounts for the correlation among the variables [3].[14] It should be noted that there are no hard and fast rules regarding how many should be less than .05 for a good factor solution. Furthermore, the cutoff value (of .05) itself is a rule of thumb and is subject to debate. Overall, though, it appears that the two-factor model extracted is doing an adequate job in accounting for the correlations among the twelve attributes.

5.7.2 Interpreting the Factor Structure

As pointed out earlier, the most important step in factor analysis is to provide an interpretation of the extracted factor structure. For the purpose of this example, it is hypothesized that consumers are essentially using orthogonal dimensions to evaluate various detergents in the marketplace. Consequently, varimax rotation is employed to provide a simple structure. As can be seen from the rotated factor loading matrix and the plot, the first four attributes load highly on the second factor and the remaining eight variables load highly on the first factor [4, 5]. The first factor, therefore, may be labeled as efficacy or ability of the detergent to do its job (i.e., clean the clothes) and the second factor can be labeled as mildness (the mildness quality of detergents).

5.8 FACTOR ANALYSIS VERSUS PRINCIPAL COMPONENTS ANALYSIS

Although factor analysis and principal components analysis are typically labeled as data-reduction techniques, there are significant differences between the two techniques. The objective of principal components analysis is to reduce the number of variables to

[14]The number of pairwise correlations among p variables is given by $p(p-1)/2$.

Exhibit 5.4

SPSS output for detergent study

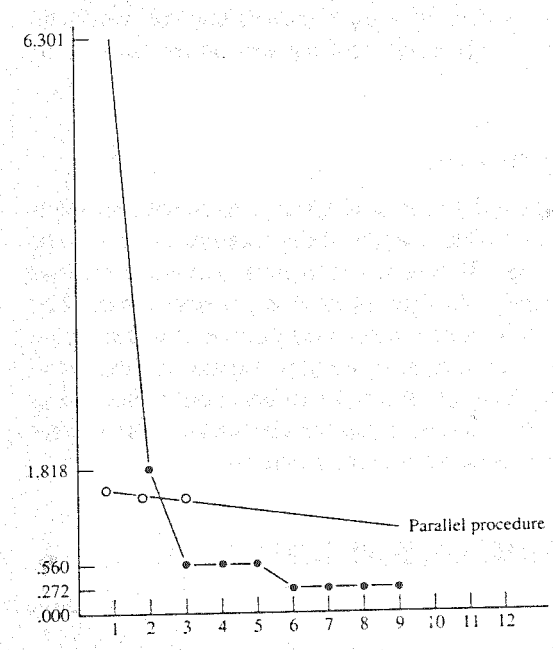
1 KAISER-MEYER-OLKIN MEASURE OF SAMPLING ADEQUACY = .90233
BARTLETT TEST OF SPHERICITY = 1091.5317, SIGNIFICANCE = .00000

EXTRACTION 1 FOR ANALYSIS 1, PRINCIPAL AXIS FACTORING (PAF)

2 INITIAL STATISTICS:

VARIABLE	COMMUNALITY	*	FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
V1	.42052	*	1	6.30111	52.5	52.5
V2	.39947	*	2	1.81757	15.1	67.7
V3	.56533	*	3	.66416	5.5	73.2
V4	.56605	*	4	.57155	4.8	78.0
V5	.60467	*	5	.55995	4.7	82.6
V6	.57927	*	6	.44517	3.7	86.3
V7	.69711	*	7	.41667	3.5	89.8
V8	.74574	*	8	.32554	2.7	92.5
V9	.66607	*	9	.27189	2.3	94.8
V10	.59287	*	10	.25690	2.1	96.9
V11	.71281	*	11	.19159	1.6	98.5
V12	.64409	*	12	.17789	1.5	100.0

2a



3 THERE ARE 9 (13.0%) RESIDUALS (ABOVE DIAGONAL) THAT ARE > 0.05
0 VARIMAX CONVERGED IN 3 ITERATIONS.

(continued)

Exhibit 5.4 (continued)

4 ROTATED FACTOR MATRIX:

	FACTOR 1	FACTOR 2
V1	.12289	.65101
V2	.13900	.64781
V3	.24971	.78587
V4	.29387	.74118
V5	.73261	.15469
V6	.73241	.20401
V7	.77455	.22464
V8	.85701	.20629
V9	.80879	.19538
V10	.69326	.23923
V11	.77604	.25024
V12	.79240	.19822

5 HORIZONTAL FACTOR 1 VERTICAL FACTOR 2

```
                                I
                                I
                                I
                                I       3
                                I         4
                                I
                                I   2
                                I
                                I
                                I
                                I
                                I
                                I                       11
                                I                  10 612 8
                                I                     5
                                I
                                I
- - - - - - - - - - - - - - - - - + - - - - - - - - - - - - - - - - - - -
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
                                I
```

a few components such that each component forms a new variable and the number of retained components explains the maximum amount of variance in the data. The objective of factor analysis, on the other hand, is to search or identify the underlying factor(s) or latent constructs that can explain the intercorrelation among the variables. There are two major differences. First, principal components analysis places emphasis on explaining the variance in the data; the objective of factor analysis is to explain the correlation among the indicators. Second, in principal components analysis the variables form an index (e.g., Consumer Price Index, Dow Jones Industrial Average). For example, in the following equation

$$\xi_1 = w_{11}x_1 + w_{12}x_2 + \cdots + w_{1p}x_p,$$

the ξ_1 component is formed by the variables $x_1, x_2, \ldots, x_p$. The variables are called formative indicators of the component as the index is formed by the variables. In factor analysis, on the other hand, the variables or indicators reflect the presence of unobservable construct(s) or factor(s). For example, in the following equations:

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \cdots + \lambda_{1m}\xi_m + \epsilon_1 \\ x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \cdots + \lambda_{2m}\xi_m + \epsilon_2 \\ &\vdots \\ x_p &= \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \cdots + \lambda_{pm}\xi_m + \epsilon_p, \end{aligned}$$

where the variables, $x_1, x_2, \ldots, x_p$, are functions of the latent construct(s) or factor(s), $\xi_1, \xi_2, \ldots, \xi_m$, and the unique factors. In other words, they reflect the presence of the unobservable or the latent constructs (i.e., the factor(s)) and hence the variables are called reflective indicators.

5.9 EXPLORATORY VERSUS CONFIRMATORY FACTOR ANALYSIS

In an exploratory factor analysis the researcher has little or no knowledge about the factor structure. For example, consider the case where the researcher is interested in measuring the excellence of a given firm. Suppose the researcher has no knowledge regarding: (1) the number of factors or dimensions of excellence; (2) whether these dimensions are orthogonal or oblique; (3) the number of indicators of each factor; and (4) which indicators represent which factor. In other words, there is very little theory that can be used for answering the above questions. In such a case, the researcher may collect data and explore or search for a factor structure or theory which can explain the correlations among the indicators. Such an analysis is called *exploratory factor analysis.*

Confirmatory factor analysis, on the other hand, assumes that the factor structure is known or hypothesized *a priori*. For example, consider the factor structure given in Figure 5.9. Excellence is hypothesized as a general factor with eight subdimensions or subfactors. Each of these subdimensions is measured by its respective indicators. The indicators are measures of one and only one factor. In other words, the complete factor structure along with the respective indicators and the nature of the pattern loadings is specified *a priori*. The objective is to empirically verify or confirm the factor structure. Such an analysis is referred to as *confirmatory factor analysis*. Confirmatory factor analysis is discussed in the next chapter.

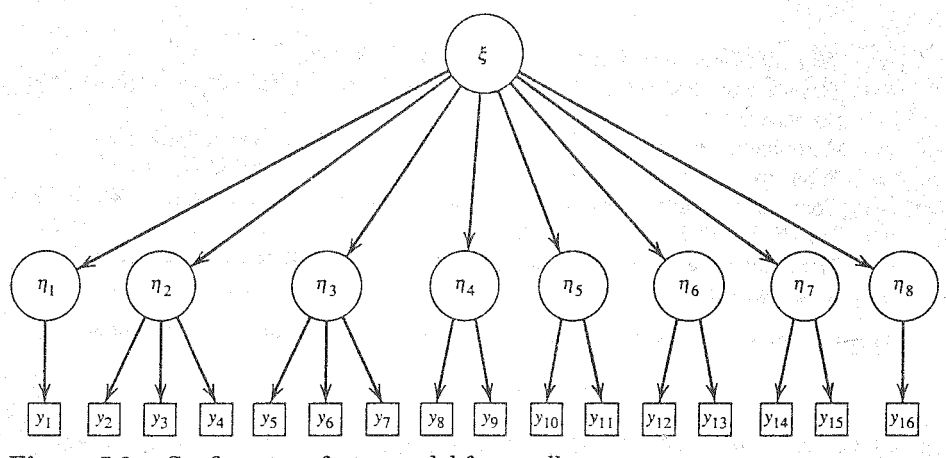


Figure 5.9 Confirmatory factor model for excellence.

5.10 SUMMARY

In the behavioral and social sciences, researchers need to develop scales for the various unobservable constructs such as attitudes, image, intelligence, personality, and patriotism. Factor analysis is a technique that can be used to develop such scales. Factor analysis is also useful for understanding the underlying reasons for the correlations among the variables.

The two most popular factor analysis techniques are principal components factoring (PCF) and principal axis factoring (PAF). In PCF, factors are extracted by assuming that the communalities of all the variables are one. That is, it is assumed that the error or unique variance is zero. For this reason many researchers do not consider PCF to be a true factor analysis technique. On the other hand, in PAF first an estimate of the communalities is obtained followed by an estimate of the factor solution. An iterative procedure is used to estimate the communalities and the factor solution. The iterative procedure terminates when the estimate of communalities converges.

Although factor analysis and principal components analysis appear to be related, they are conceptually two different techniques. In principal components analysis, one is interested in forming a composite index of a number of variables. There is no theory or reason as to why the different variables comprising the index should be correlated. Factor analysis, on the other hand, posits that any correlation among the indicators or variables is due to the common factors. That is, the common factors are responsible for any correlation that might exist among the indicators.

A distinction was also made between exploratory factor analysis and confirmatory factor analysis. Exploratory factor analysis is used when there is very little knowledge about the underlying structure of the factor model. On the other hand, in confirmatory factor analysis the main objective is to empirically confirm or verify a given factor model. Confirmatory factor analysis is discussed in the next chapter.

QUESTIONS

5.1 Consider the five-indicator single-factor model represented by the following equations:

$$\begin{aligned} V &= 0.65F_1 + U_V \\ W &= 0.84F_1 + U_W \\ X &= 0.70F_1 + U_X \\ Y &= 0.32F_1 + U_Y \\ Z &= 0.28F_1 + U_Z. \end{aligned}$$

The usual assumptions hold for this model; i.e., means of indicators, common factor, and unique factors are zero, indicators and the common factor have unit variances, etc

(a) What are the pattern loadings for indicators V, X, and Z?
(b) Show a graphical representation of the model. Indicate the pattern loadings in your representation.
(c) Compute the communalities of the indicators with the common factor F_1.
(d) What are the unique variances associated with each indicator?
(e) Compute the correlations between the following sets of indicators: (i) V, W; (ii) W, X; (iii) W, Z; (iv) Y, Z.
(f) What is the shared variance between each indicator and the common factor?
(g) What percentage of the total shared variance is due to indicators V, W, and X?

5.2 Consider the two-indicator two-factor model represented by the following equations:

$$\begin{aligned} A &= 0.85F_1 + 0.12F_2 + U_A \\ B &= 0.74F_1 + 0.07F_2 + U_B \\ C &= 0.67F_1 + 0.18F_2 + U_C \\ D &= 0.21F_1 + 0.93F_2 + U_D \\ E &= 0.05F_1 + 0.77F_2 + U_E \\ F &= 0.08F_1 + 0.62F_2 + U_F. \end{aligned}$$

The usual assumptions hold for the above model. Also, assume that the common factors F_1 and F_2 are uncorrelated.
(a) What are the pattern loadings of indicators A, C, and E on the factors F_1 and F_2?
(b) What are the structure loadings of indicators A, C, and E on the factors F_1 and F_2?
(c) Compute the correlations between the following sets of indicators: (i) A, B; (ii) C, D; (iii) E, F.
(d) What percentage of the variance of indicators A, C, and F is not accounted for by the common factors F_1 and F_2?
(e) Identify sets of indicators that share more than 90% of the total shared variance with each common factor. Which indicators should therefore be used to interpret each common factor?

5.3 Repeat parts (a), (b), (c), and (d) of Question 5.2, taking into account the fact that the correlation between the common factors F_1 and F_2 is given by $Corr(F_1, F_2) = \phi_{12} = 0.20$.

5.4 The correlation matrix for a hypothetical data set is given in Table Q5.1.

Table Q5.1

	X_1	X_2	X_3	X_4
X_1	1.000			
X_2	0.690	1.000		
X_3	0.280	0.255	1.000	
X_4	0.350	0.195	0.610	1.000

The following estimated factor loadings were extracted by the principal axis factoring procedure:

Variable	F_1	F_2
X_1	0.80	0.20
X_2	0.70	0.15
X_3	0.10	0.90
X_4	0.20	0.70

Compute and discuss the following: (a) Specific variances; (b) Communalities; (c) Proportion of variance explained by each factor; (d) Estimated or reproduced correlation matrix; and (e) Residual matrix.

5.5 Consider the following two-factor orthogonal models:

Model 1

$$
\begin{aligned}
X_1 &= 0.558F_1 + 0.615F_2 + U_1 \\
X_2 &= 0.604F_1 + 0.748F_2 + U_2 \\
X_3 &= 0.469F_1 + 0.556F_2 + U_3 \\
X_4 &= 0.818F_1 - 0.411F_2 + U_4 \\
X_5 &= 0.866F_1 - 0.466F_2 + U_5 \\
X_6 &= 0.686F_1 - 0.461F_2 + U_6
\end{aligned}
$$

Model 2

$$
\begin{aligned}
X_1 &= 0.104F_1 + 0.824F_2 + U_1 \\
X_2 &= 0.065F_1 + 0.959F_2 + U_2 \\
X_3 &= 0.065F_1 + 0.725F_2 + U_3 \\
X_4 &= 0.906F_1 + 0.134F_2 + U_4 \\
X_5 &= 0.977F_1 + 0.116F_2 + U_5 \\
X_6 &= 0.827F_1 + 0.016F_2 + U_6
\end{aligned}
$$

(a) Show that these two models provide an illustration of the factor indeterminacy problem. In other words, show that although the loadings and shared variances of each indicator are different for the two models, the total communalities of each indicator, the unique variances, and the correlation matrices of the indicators are the same for the two models.

(b) In what way(s) is the interpretation of the common factors in the two models different?

5.6 Plot the factor solution given in Model 1 of Question 5.5. Your horizontal and vertical axes are F_1 and F_2, respectively. Now rotate the axes in a clockwise direction by an angle $\theta = 35°$. Label the new axes F_1^* and F_2^*. Plot the factor solution given in Model 2 of Question 5.5, using the new axes F_1^* and F_2^*.

(a) How does the location of the points X_1, X_2, X_3, X_4, X_5, and X_6 change from the first to the second plot?

(b) Use your answer in part (a) to show that the factor indeterminacy problem is essentially a factor rotation problem.

5.7 What is the conceptual difference between factor analysis and principal components analysis?

FOR QUESTIONS 5.8 TO 5.13 EITHER THE DATA OR A CORRELATION MATRIX OF THE DATA IS PROVIDED ALONG WITH A DESCRIPTION OF THE DATA. IN EACH CASE DO THE FOLLOWING:

1. Factor analyze the data (or correlation matrix) and identify the smallest number of common factors that best account for the variance in the data.
2. Using factor rotations identify the most plausible factor solution.
3. Label the identified factors suitably and interpret/discuss the factors in light of the description of the data.
4. Do not forget to examine: (i) if the data are appropriate for factor analysis and (ii) the "goodness" of the factor solution.

5.8 File PHYSATT.DAT gives the correlation matrix of data on eight attributes of 293 male athletes representing various sports in a university. The eight attributes are: (1) Height;

(2) Weight; (3) Width of the shoulders; (4) Length of the legs; (5) Time taken to run a mile; (6) Time taken to run up 10 flights of stairs; (7) Number of push-ups completed in 5 minutes; (8) Heart rate after running a mile.

5.9 File TEST.DAT gives the correlation matrix of data on 12 tests conducted on 412 high school students. The 12 tests were as follows: (1) Differentiation of bright from dark objects; (2) Counting; (3) Differentiation of parallel from nonparallel lines; (4) Simple decoding speed; (5) Completion of words/sentences; (6) Comprehension; (7) Reading; (8) General awareness; (9) Arithmetic computations; (10) Permutations-combinations; (11) Routine task; (12) Repetitive task.

5.10 Analyze the audiometric data given in file AUDIO.DAT. Refer to Question 4.8 for a description of the data.

5.11 File BANK.DAT gives the correlation matrix of data from a customer satisfaction survey undertaken by ABC Savings Bank for their EasyBuy credit card. 540 respondents indicated their level of agreement/disagreement and level of satisfaction/dissatisfaction to 15 statements/services given in file BANK.DOC.

Note: The correlation matrix is based on fictitious data.

5.12 Analyze the mass transportation data given in file MASST.DAT (for this analysis use only variables ID and V_{19} to V_{38}; ignore the remaining variables). Refer to file MASST.DOC for a description of the data.

5.13 File NUT.DAT gives the data from a survey undertaken to determine the attitudes and opinions of 254 respondents toward nutrition. One section of the survey requested respondents to provide their views on 46 statements dealing with daily activities. File NUT.DOC gives a list of the statements. Respondents used a 5-point scale (1 = strongly disagree to 5 = strongly agree) to indicate their views.

5.14 File SOFTD.DAT gives data from a survey undertaken to determine consumer perceptions of six competing brands of soft drinks. The brands rated were as follows: (1) Pepsi Cola (regular); (2) Coke (regular); (3) Gatorade; (4) Allsport; (5) Lipton original tea; (6) Nestea.

Note: Although national brand names have been used, the data are fictitious.

Respondents used a 7-point scale (1 = strongly disagree to 7 = strongly agree) to indicate their level of agreement/disagreement with the 10 statements given in file SOFTD.DOC (in each of the statements substitute "Brand X" with the brands listed above).

Use factor analysis to identify and label the smallest number of factors that best account for the variance in the data. Also, use the factor scores to plot and interpret perceptual map(s) of the six brands.

Appendix

A5.1 ONE-FACTOR MODEL

Consider the following equations representing a p-indicator one-factor model:

$$\begin{aligned} x_1 &= \lambda_1\xi + \epsilon_1 \\ x_2 &= \lambda_2\xi + \epsilon_2 \\ &\vdots \\ x_p &= \lambda_p\xi + \epsilon_p \end{aligned} \tag{A5.1}$$

where $x_1, x_2, \ldots, x_p$ are indicators of the common factor ξ; $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the pattern loadings; and $\epsilon_1, \epsilon_2, \ldots, \epsilon_p$ are unique factors. Without loss of generality it is assumed that:

1. Means of indicators, common factor, and unique factors are zero.
2. Variances of indicators and common factor are one. That is, the indicators and the common factor are standardized.
3. The unique factors are not correlated among themselves or with the common factor. That is, $E(\xi_i \epsilon_j) = 0$ and $E(\epsilon_i \epsilon_j) = 0$.

The variance of any indicator, x_j, is given by:

$$\begin{aligned} E(x_j^2) &= E[(\lambda_j \xi + \epsilon_j)^2] \\ &= \lambda_j^2 E(\xi^2) + E(\epsilon_j^2) + 2E(\lambda_j \xi \epsilon_j) \\ Var(x_j) &= \lambda_j^2 + Var(\epsilon_j). \end{aligned} \tag{A5.2}$$

As can be seen from Eq. A5.2, the total variance of any indicator can be decomposed into the following two components:

1. Variance that is in common with the factor and is equal to the square of its pattern loading.
2. Variance that is in common with the unique factor, ϵ_j.

The correlation between any indicator, x_j, and the factor, ξ, is given by:

$$\begin{aligned} E(x_j \xi) &= E[(\lambda_j \xi + \epsilon_j)\xi] \\ &= \lambda_j E(\xi^2) + E(\xi \epsilon_j) \\ &= \lambda_j. \end{aligned} \tag{A5.3}$$

That is, the correlation between any indicator and the factor is equal to its pattern loading. This correlation is referred to as the *structure loading* of an indicator. The square of the structure loading gives the shared variance between the indicator and the factor.

The correlation between any two indicators, say x_j and x_k, is given by:

$$\begin{aligned} E(x_j x_k) &= E[(\lambda_j \xi + \epsilon_j)(\lambda_k \xi + \epsilon_k)] \\ &= \lambda_j \lambda_k E(\xi^2) + \lambda_j E(\xi \epsilon_k) + \lambda_k E(\xi \epsilon_j) + E(\epsilon_j \epsilon_k) \\ &= \lambda_j \lambda_k. \end{aligned} \tag{A5.4}$$

That is, the correlation between any two variables is given by the product of the respective pattern loadings.

A5.2 TWO-FACTOR MODEL

Consider a p-indicator two-factor model given by the following equations:

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \epsilon_1 \\ x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \epsilon_2 \\ &\vdots \\ x_p &= \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \epsilon_p. \end{aligned} \tag{A5.5}$$

For ease of notation we drop the indicator subscript, p. The variance of any variable x is given by:

$$\begin{aligned} E(x^2) &= E(\lambda_1 \xi_1 + \lambda_2 \xi_2 + \epsilon)^2 \\ &= \lambda_1^2 E(\xi_1^2) + \lambda_2^2 E(\xi_2^2) + E(\epsilon^2) \\ &\quad + 2\lambda_1 \lambda_2 E(\xi_1 \xi_2) + 2\lambda_1 E(\xi_1 \epsilon) + 2\lambda_2 E(\xi_2 \epsilon) \\ Var(x) &= \lambda_1^2 + \lambda_2^2 + Var(\epsilon) + 2\lambda_1 \lambda_2 \phi, \end{aligned} \tag{A5.6}$$

where ϕ is the correlation between factors ξ_1 and ξ_2. For an orthogonal factor model Eq. A5.6 reduces to

$$Var(x) = \lambda_1^2 + \lambda_2^2 + Var(\epsilon), \tag{A5.7}$$

as $\phi = 0$. As can be clearly seen, the variance of any indicator can be decomposed into the following components:

1. Variance that is in common with the first factor, ξ_1, and is equal to the square of the pattern loading, λ_1.
2. Variance that is in common with the second factor, ξ_2, and is equal to the square of the pattern loading, λ_2.
3. Variance that is in common with ξ_1 and ξ_2 due to the joint effect of the two factors, and is equal to twice the product of the respective pattern loadings and the correlation among the factors. For an orthogonal factor model this component is zero since the correlation, ϕ, is equal to zero.

   The sum of the preceding three components is referred to as the total communality of the indicator with the factors.
4. Variance that is in common with the unique factor.

The correlation between any indicator and any factor, say ξ_1, is given by:

$$\begin{aligned} E(x\xi_1) &= E[(\lambda_1\xi_1 + \lambda_2\xi_2 + \epsilon)\xi_1] \\ &= \lambda_1 E(\xi_1^2) + \lambda_2 E(\xi_1\xi_2) + E(\epsilon\xi_1) \\ Cor(x\xi_1) &= \lambda_1 + \lambda_2\phi. \end{aligned} \tag{A5.8}$$

That is, the correlation between any variable and the factor (i.e., the structure loading) is given by its pattern loading *plus* the product of the pattern loading for the second factor and the correlation between the two factors. For an orthogonal factor model, Eq. A5.8 can be written as

$$Cor(x\xi_1) = \lambda_1. \tag{A5.9}$$

It is obvious that for an orthogonal factor model the structure loading is equal to its pattern loading and is commonly referred to as the loading.

The shared variance between the factor and an indicator is obtained by squaring Eq. A5.8. That is,

$$\begin{aligned} \text{Shared Variance} &= (\lambda_1 + \lambda_2\phi)^2 \\ &= \lambda_1^2 + \lambda_2^2\phi^2 + 2\lambda_1\lambda_2\phi. \end{aligned} \tag{A5.10}$$

For an orthogonal factor model the above equation can be rewritten as:

$$\text{Shared Variance} = \lambda_1^2. \tag{A5.11}$$

As can be clearly seen, the shared variance of an orthogonal factor model is equal to the square of the respective loading and is the same as the communality. On the other hand, shared variance of an oblique factor model is not the same as the communality.

The correlation between any two indicators, say x_j and x_k, is given by:

$$\begin{aligned} E(x_jx_k) &= E[(\lambda_{j1}\xi_1 + \lambda_{j2}\xi_2 + \epsilon_j)(\lambda_{k1}\xi_1 + \lambda_{k2}\xi_2 + \epsilon_k)] \\ &= \lambda_{j1}\lambda_{k1}E(\xi_1^2) + \lambda_{j2}\lambda_{k2}E(\xi_2^2) + E(\epsilon_j\epsilon_k) \\ &\quad + \lambda_{j1}\lambda_{k2}E(\xi_1\xi_2) + \lambda_{j2}\lambda_{k1}E(\xi_1\xi_2) + \lambda_{j1}E(\xi_1\epsilon_k) \\ &\quad + \lambda_{j2}E(\xi_2\epsilon_k) + \lambda_{k1}E(\xi_1\epsilon_j) + \lambda_{k2}E(\xi_2\epsilon_j) \\ Cor(x_jx_k) &= \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2} + (\lambda_{j1}\lambda_{k2} + \lambda_{j2}\lambda_{k1})\phi. \end{aligned} \tag{A5.12}$$

For an orthogonal factor model, Eq. A5.12 can be written as:

$$Cor(x_j x_k) = \lambda_{j1}\lambda_{k1} + \lambda_{j2}\lambda_{k2}. \tag{A5.13}$$

A5.3 MORE THAN TWO FACTORS

Consider a p-indicator m-factor model given by the following equations:

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \lambda_{12}\xi_2 + \cdots + \lambda_{1m}\xi_m + \epsilon_1 \\ x_2 &= \lambda_{21}\xi_1 + \lambda_{22}\xi_2 + \cdots + \lambda_{2m}\xi_m + \epsilon_2 \\ &\vdots \\ x_p &= \lambda_{p1}\xi_1 + \lambda_{p2}\xi_2 + \cdots + \lambda_{pm}\xi_m + \epsilon_p, \end{aligned} \tag{A5.14}$$

where $x_1, x_2, \ldots, x_p$ are indicators of the m factors, λ_{pm} is the pattern loading of the pth variable on the mth factor, and ϵ_p is the unique factor for the pth variable.

Eq. A5.14 can be represented in matrix form as:

$$\mathbf{x} = \Lambda\xi + \epsilon, \tag{A5.15}$$

where $\mathbf{x}$ is a $p \times 1$ vector of variables, Λ is a $p \times m$ matrix of factor pattern loadings, ξ is an $m \times 1$ vector of unobservable factors, and ϵ is a $p \times 1$ vector of unique factors. Equation A5.15 is the basic factor analysis equation. It will be assumed that the factors are not correlated with the error components, and without loss of generality it will be assumed that the means and variances of variables and factors are zero and one, respectively. The correlation matrix, $\mathbf{R}$, of the indicators is given by:[1]

$$\begin{aligned} E(\mathbf{x}\mathbf{x}') &= E[(\Lambda\xi + \epsilon)(\Lambda\xi + \epsilon)'] \\ &= E[(\Lambda\xi + \epsilon)(\xi'\Lambda' + \epsilon')] \\ &= E(\Lambda\xi\xi'\Lambda') + E(\epsilon\epsilon') \\ \mathbf{R} &= \Lambda\Phi\Lambda' + \Psi, \end{aligned} \tag{A5.16}$$

where $\mathbf{R}$ is the correlation matrix of the observables, Λ is the pattern loading matrix, Φ is the correlation matrix of the factors, and Ψ is a diagonal matrix containing the unique variances. The communalities are given by the diagonal of the $\mathbf{R} - \Psi$ matrix. The off-diagonals of the $\mathbf{R}$ matrix give the correlation among the indicators. Λ, Φ, and Ψ matrices are referred to as parameter matrices of the factor analytic model, and it is clear that the correlation matrix of the observables is a function of the parameters. The objective of factor analysis is to estimate the parameter matrices given the correlation matrix.

For an orthogonal factor model, Eq. A5.16 can be rewritten as

$$\mathbf{R} = \Lambda\Lambda' + \Psi. \tag{A5.17}$$

If no *a priori* constraints are imposed on the parameter matrices then we have exploratory factor analysis; *a priori* constraints imposed on the parameter matrices result in a confirmatory factor analysis.

The correlation between the indicators and the factors is given by:

$$\begin{aligned} E(\mathbf{x}\xi') &= E[(\Lambda\xi + \epsilon)\xi'] \\ &= \Lambda E(\xi\xi') + E(\epsilon\xi') \\ \mathbf{A} &= \Lambda\Phi, \end{aligned} \tag{A5.18}$$

[1]Since the data are standardized, the correlation matrix is the same as the covariance matrix.

where A gives the correlation between indicators and the factors. For an orthogonal factor model,

$$A = \Lambda. \tag{A5.19}$$

Again, it can be clearly seen that for an orthogonal factor model the pattern loadings are equal to structure loadings and are commonly referred to as the loadings of the variables.

A5.4 FACTOR INDETERMINACY

In exploratory factor analysis the factor solution is not unique. A number of different factor pattern loadings and factor correlations will produce the same correlation matrix for the indicators. Mathematically it is not possible to differentiate between the alternative factor solutions, and this is referred to as the factor indeterminacy problem. Factor indeterminacy results from two sources: the first pertains to the estimation of the communalities and the second is the problem of factor rotation. Each is described below.

A5.4.1 Communality Estimation Problem

Equation A5.17 can be rewritten as

$$\Lambda\Lambda' = R - \Psi. \tag{A5.20}$$

This is known as the fundamental factor analysis equation. Note that the right-hand side of the equation gives the correlation matrix with the communalities in the diagonal. Estimates of the factor loadings (i.e., Λ) are obtained by computing the eigenstructure of the $R - \Psi$ matrix. However, the estimate of Ψ is obtained by solving the following equation:

$$\Psi = R - \Lambda\Lambda'. \tag{A5.21}$$

That is, the solution of Eq. A5.20 requires the solution of Eq. A5.21, but the solution of Eq. A5.21 requires solution of Eq. A5.20. It is this circularity that leads to the estimation of communalities problem.

A5.4.2 Factor Rotation Problem

Once the communalities are known or have been estimated, the parameter matrices of the factor model can be estimated. However, one can obtain a number of different estimates for Λ and Φ matrices. Geometrically, this is equivalent to rotating the factor axes in the factor space without changing the orientation of the vectors representing the variables. For example, suppose we have any orthonormal matrix C such that $C'C = CC' = I$. Rewrite Eq. A5.16 as

$$\begin{aligned} R &= \Lambda CC'\Phi CC'\Lambda' + \Psi \\ &= \Lambda^*\Phi^*\Lambda^{*\prime} + \Psi, \end{aligned} \tag{A5.22}$$

where $\Lambda^* = \Lambda C$ and $\Phi^* = C'\Phi C$. As can be seen, the factor pattern matrix and the correlation matrix of factors can be changed by the transformation matrix, C, without affecting the correlation matrix of the observables. And, an infinite number of transformation matrices can be obtained, each resulting in a different factor analytic model. Geometrically, the effect of multiplying the Λ matrix by the transformation matrix, C, is to rotate the factor axes without changing the orientation of the indicator vectors. This source of factor indeterminacy is referred to as the *factor rotation* problem. One has to specify certain constraints in order to obtain a unique estimate of the transformation matrix, C. Some of the constraints commonly used are discussed in the following section.

A5.5 FACTOR ROTATIONS

Rotations of the factor solution are the common type of constraints placed on the factor model for obtaining a unique solution. There are two types of factor rotation techniques: orthogonal and oblique. Orthogonal rotations result in orthogonal factor models, whereas oblique rotations result in oblique factor models. Both types of rotation techniques are discussed below.

A5.5.1 Orthogonal Rotation

In an orthogonal factor model it is assumed that $\Phi = \mathbf{I}$. Orthogonal rotation technique involves the identification of a transformation matrix C such that the new loading matrix is given by $\Lambda^* = \Lambda C$ and

$$\mathbf{R} = \Lambda^* \Lambda^{*'}.$$

The transformation matrix is estimated such that the new loadings result in an interpretable factor structure. Quartimax and varimax are the most commonly used orthogonal rotation techniques for obtaining the transformation matrix.

Quartimax Rotation

As discussed in the chapter, the objective of quartimax rotation is to identify a factor structure such that all the indicators have a fairly high loading on the same factor; in addition, each indicator should load on one other factor and have near zero loadings on the remaining factors. This objective is achieved by maximizing the variance of the loadings across factors, subject to the constraint that the communality of each variable is unchanged. Thus, suppose for any given variable i, we define

$$Q_i = \frac{\sum_{j=1}^{m}(\lambda_{ij}^2 - \lambda_{i.}^2)^2}{m}, \tag{A5.23}$$

where Q_i is the variance of the communalities (i.e., square of the loadings) of variable i, λ_{ij}^2 is the squared loading of the ith variable on the jth factor, $\lambda_{i.}^2$ is the average squared loading of the ith variable, and m is the number of factors. The preceding equation can be rewritten as

$$Q_i = \frac{m\sum_{j=1}^{m}\lambda_{ij}^4 - (\sum_{j=1}^{m}\lambda_{ij}^2)^2}{m^2}. \tag{A5.24}$$

The total variance of all the variables is given by:

$$Q = \sum_{i=1}^{p} Q_i = \sum_{i=1}^{p}\left[\frac{m\sum_{j=1}^{m}\lambda_{ij}^4 - (\sum_{j=1}^{m}\lambda_{ij}^2)^2}{m^2}\right]. \tag{A5.25}$$

For quartimax rotation the transformation matrix, C, is found such that Eq. A5.23 is maximized subject to the condition that the communality of each variable remains the same. Note that once the initial factor solution has been obtained, the number of factors, m, remains constant. Furthermore, the second term in the equation, $\sum_{j=1}^{m}\lambda_{ij}^2$, is the communality of the variable and, hence, it will also be a constant. Therefore, maximization of Eq. A5.23 reduces to maximizing the following equation:

$$Q = \sum_{i=1}^{p}\sum_{j=1}^{m}\lambda_{ij}^4. \tag{A5.26}$$

In most cases, prior to performing rotation the loadings of each variable are normalized by dividing the loading of each variable by the total communality of the respective variable.

Varimax Rotation

As discussed in the chapter, the objective of varimax rotation is to determine the transformation matrix, C, such that any given factor will have some variables that will load very high on it and some that will load very low on it. This is achieved by maximizing the variance of the squared loading across variables, subject to the constraint that the communality of each variable is unchanged. That is, for any given factor

$$\begin{aligned} V_j &= \frac{\sum_{i=1}^{p}(\lambda_{ij}^2 - \bar{\lambda}_j^2)^2}{p} \\ &= \frac{p\sum_{i=1}^{p}\lambda_{ij}^4 - (\sum_{i=1}^{p}\lambda_{ij}^2)^2}{p^2}, \end{aligned} \tag{A5.27}$$

where V_j is the variance of the communalities of the variables within factor j and $\bar{\lambda}_j^2$ is the average squared loading for factor j. The total variance for all the factors is then given by

$$\begin{aligned} V &= \sum_{j=1}^{m} V_j \\ &= \sum_{j=1}^{m}\left(\frac{p\sum_{i=1}^{p}\lambda_{ij}^4 - (\sum_{i=1}^{p}\lambda_{ij}^2)^2}{p^2}\right) \\ &= \frac{\sum_{j=1}^{m}\sum_{i=1}^{p}\lambda_{ij}^4}{p} - \frac{\sum_{j=1}^{m}(\sum_{i=1}^{p}\lambda_{ij}^2)^2}{p^2}. \end{aligned} \tag{A5.28}$$

Since the number of variables remains the same, maximizing the preceding equation is the same as maximizing

$$pV = \sum_{j=1}^{m}\sum_{i=1}^{p}\lambda_{ij}^4 - \frac{\sum_{j=1}^{m}(\sum_{i=1}^{p}\lambda_{ij}^2)^2}{p}. \tag{A5.29}$$

The orthogonal matrix, C, is obtained such that Eq. A5.29 is maximized, subject to the constraint that the communality of each variable remains the same.

Other Orthogonal Rotations

It is clear from the preceding discussion that quartimax rotation maximizes the total variance of the loadings row-wise and varimax maximizes it column-wise. It is therefore possible to have a rotation technique that maximizes the weighted sum of row-wise and column-wise variance. That is, maximize

$$Z = \alpha Q + \beta pV, \tag{A5.30}$$

where Q is given by Eq. A5.26 and pV is given by Eq. A5.29. Now consider the following equation:

$$\sum_{j=1}^{m}\sum_{i=1}^{p}\lambda_{ij}^4 - \gamma\frac{\sum_{j=1}^{m}(\sum_{i=1}^{p}\lambda_{ij}^2)^2}{p} \tag{A5.31}$$

where $\gamma = \beta/(\alpha + \beta)$.

Different values of γ result in different types of rotation. Specifically, the above criterion reduces to a quartimax rotation if $\gamma = 0$ (i.e., $\alpha = 1; \beta = 0$), reduces to a varimax rotation if $\gamma = 1$ (i.e., $\alpha = 0; \beta = 1$), reduces to an equimax rotation if $\gamma = m/2$, and reduces to a biquartimax if $\gamma = 0.5$ (i.e., $\alpha = 1; \beta = 1$).

Empirical Illustration of Varimax Rotation

Because varimax is one of the most popular rotation techniques, we will provide an illustrative example. Table A5.1 gives the unrotated factor pattern loading matrix obtained from Exhibit 5.2 [7]. Assume that the factor structure is rotated counterclockwise by $\theta°$. As discussed in Section 2.7 of Chapter 2, the coordinates, a_1^* and a_2^*, with respect to the new axes will be

$$\left(a_1^* \; a_2^*\right) = (a_1 \; a_2)\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

or

$$\mathbf{a}^* = \mathbf{aC}, \tag{A5.32}$$

where C is an orthonormal transformation matrix. Table A5.1 gives the new pattern loadings for a counterclockwise rotation of, say, 350°. As can be seen, the communality of the variables does not change. Also, the total column-wise variance of the squared loadings is 0.056. Table A5.2 gives the column-wise variance for different angles of rotation and shows that the maximum column-wise variance is achieved for a counterclockwise rotation of 320.057°. Table A5.3 gives the resulting loadings and the transformation matrix. Note that the loadings and the transformation matrix given in Table A5.3 are the same as those reported in Exhibit 5.2 [13a,12].

Table A5.1 **Varimax Rotation of 350°**

	Unrotated Structure			Rotated Structure		
Variable	Factor1	Factor2	Communality	Factor1	Factor2	Communality
M	.636	.523	.677	.535	.625	.677
P	.658	.385	.581	.581	.494	.581
C	.598	.304	.450	.536	.404	.450
E	.762	−.315	.680	.805	−.178	.680
H	.749	−.368	.697	.802	−.232	.697
F	.831	−.303	.783	.871	−.154	.783

Transformation Matrix

$$C = \begin{bmatrix} .985 & .174 \\ -.174 & .985 \end{bmatrix}$$

Table A5.2 **Variance of Loadings for Varimax Rotation**

	Variance of Loadings Squared		
Rotation (deg)	Factor1	Factor2	Total
350	.038	.018	.056
340	.066	.038	.104
330	.087	.054	.142
320.057	.092	.058	.149
320	.092	.058	.149
310	.077	.047	.124
300	.051	.027	.078
290	.023	.009	.032
280	.005	.003	.008

Table A5.3 **Varimax Rotation of 320.057°**

	Unrotated Structure			Rotated Structure		
Variable	Factor1	Factor2	Communality	Factor1	Factor2	Communality
M	.636	.523	.677	.152	.809	.677
P	.658	.385	.581	.257	.718	.581
C	.598	.304	.450	.263	.617	.450
E	.762	−.315	.680	.787	.248	.680
H	.749	−.368	.697	.811	.199	.697
F	.831	−.303	.783	.832	.301	.783

Transformation Matrix

$$C = \begin{bmatrix} .767 & .642 \\ -.642 & .767 \end{bmatrix}$$

Oblique Rotation

In oblique rotation the axes are not constrained to be orthogonal to each other. In other words, it is assumed that the factors are correlated (i.e., $\Phi \neq I$). The pattern loadings and structure loadings will not be the same, resulting in two loading matrices that need to be interpreted. The projection of vectors or points onto the axes, which will give the loadings, can be determined in two different ways. In Panel I of Figure A5.1 the projection is obtained by dropping lines parallel to the axes. These projections give the pattern loadings (i.e., λ's). The square of the pattern loading gives the unique contribution that the factor makes to the variance of an indicator.

In Panel II of Figure A5.1 projections are obtained by dropping lines perpendicular to the axes. These projections give the structure loadings. As seen previously, structure loadings are the simple correlations among the indicators and the factors. The square of the structure loading

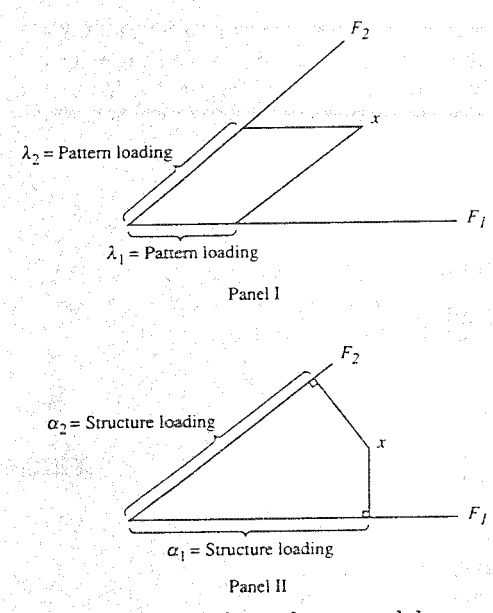


Figure A5.1 Oblique factor model.

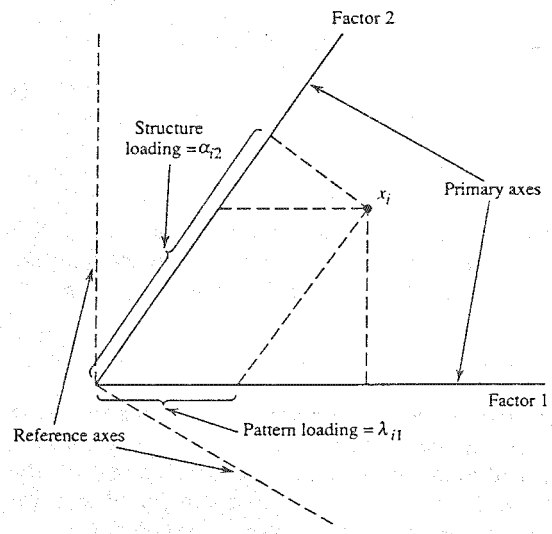


Figure A5.2 Pattern and structure loadings.

of a variable for any given factor measures the *variance accounted for in the variable jointly by the respective factor and the interaction effects of the factor with other factors.* Consequently, structure loadings are not very useful for interpreting the factor structure. It has been recommended that the pattern loadings should be used for interpreting the factors.

The coordinates of the vectors or points can be given with respect to another set of axes, obtained by drawing lines through the origin perpendicular to the oblique axes. In order to differentiate the two sets of axes, the original set of oblique axes is called the primary axes and the new set of oblique axes is called the reference axes. Figure A5.2 gives the two sets of axes. It can be clearly seen from the figure that the pattern loadings of the primary axes are the same as the structure loadings of the reference axes, and vice versa. Therefore, one can either interpret the pattern loadings of the primary axes or the structure loadings of the reference axes.

Interpretation of an oblique factor model is not very clear cut; therefore oblique rotation techniques are not very popular in behavioral and social sciences. We will not provide a mathematical discussion of oblique rotation techniques; however, the interested reader is referred to Harman (1976), Rummel (1970), and McDonald (1985) for further details.

A5.6 FACTOR EXTRACTION METHODS

A number of factor extraction methods have been proposed for exploratory factor analysis. We will only discuss some of the most popular ones. For other methods not discussed the interested reader is referred to Harman (1976), Rummel (1970), and McDonald (1985).

A5.6.1 Principal Components Factoring (PCF)

PCF assumes that the prior estimates of communality are one. The correlation matrix is then subjected to a principal components analysis. The principal components solution is given by

$$\xi = \Lambda \mathbf{x} \tag{A5.33}$$

where ξ is a $p \times 1$ vector of principal components, Λ is a $p \times p$ matrix of weights to form the principal components, and $\mathbf{x}$ is a $p \times 1$ vector of p variables. The weight matrix, Λ, is an

orthonormal matrix. That is, $\Lambda'\Lambda = \Lambda\Lambda' = \mathrm{I}$. Premultiplying Eq. A5.33 by Λ' results in

$$\Lambda'\xi = \Lambda'\Lambda\mathrm{x}, \tag{A5.34}$$

or

$$\mathrm{x} = \Lambda'\xi. \tag{A5.35}$$

As can be seen above, variables can be written as functions of the principal components. PCF assumes that the first m principal components of the ξ matrix represent the m common factors and the remaining $p - m$ principal components are used to determine the unique variance.

A5.6.2 Principal Axis Factoring (PAF)

PAF essentially reduces to PCF with iterations. In the first iteration the communalities are assumed to be one. The correlation matrix is subjected to a PCF and the communalities are estimated. These communalities are substituted in the diagonal of the correlation matrix. The modified correlation matrix is subjected to another PCF. The procedure is repeated until the estimates of communality converge according to a predetermined convergence criterion.

A5.7 FACTOR SCORES

Unlike principal components scores, which are computed, the factor scores have to be estimated. Multiple regression is one of the techniques that has been used to estimate the factor score coefficients. For example, the factor score for individual i on a given factor j can be represented as

$$\hat{F}_{ij} = \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}, \tag{A5.36}$$

where $\hat{F}_{ij}$ is the estimated factor score for factor j for individual i, $\hat{\beta}_p$ is the estimated factor score coefficient for variable p, and x_{ip} is the pth observed variable for individual i. This equation can be represented in matrix form as

$$\hat{\mathrm{F}} = \mathrm{X}\hat{\mathrm{B}}, \tag{A5.37}$$

where $\hat{\mathrm{F}}$ is an $n \times m$ matrix of m factor scores for the n individuals, X is an $n \times p$ matrix of observed variables, and $\hat{\mathrm{B}}$ is a $p \times m$ matrix of estimated factor score coefficients. For standardized variables

$$\hat{\mathrm{F}} = \mathrm{Z}\hat{\mathrm{B}}, \tag{A5.38}$$

Eq. A5.38 can be written as

$$\frac{1}{n}\mathrm{Z}'\hat{\mathrm{F}} = \frac{1}{n}\mathrm{Z}'\mathrm{Z}\hat{\mathrm{B}}$$

or

$$\Lambda = \mathrm{R}\hat{\mathrm{B}} \tag{A5.39}$$

as

$$\frac{1}{n}(\mathrm{Z}'\mathrm{Z}) = \mathrm{R} \quad \text{and} \quad \frac{1}{n}\mathrm{Z}'\hat{\mathrm{F}} = \Lambda.$$

Therefore, the estimated factor score coefficient matrix is given by

$$\hat{\mathrm{B}} = \mathrm{R}^{-1}\Lambda \tag{A5.40}$$

and the estimated factor scores by

$$\hat{F} = ZR^{-1}\Lambda. \tag{A5.41}$$

It should be noted from Eq. A5.41 that the estimated factor score is a function of the original standardized variables and the loading matrix. Due to the factor indeterminacy problem a number of loading matrices are possible, each resulting in a separate set of factor scores. In other words, the factor scores are not unique. For this reason many researchers hesitate to use the factor scores in further analysis. For further details on the indeterminacy of factor scores see McDonald and Mulaik (1979).

CHAPTER 6

Confirmatory Factor Analysis

In exploratory factor analysis the structure of the factor model or the underlying theory is not known or specified *a priori*; rather, data are used to help *reveal* or identify the structure of the factor model. Thus, exploratory factor analysis can be viewed as a technique to aid in theory building. In confirmatory factor analysis, on the other hand, the precise structure of the factor model, which is based on some underlying theory, is hypothesized. For example, suppose that based on previous research it is hypothesized that a construct or factor to measure consumers' ethnocentric tendencies is a one-dimensional construct with 17 indicators or variables as its measures.[1] That is, a one-factor model of consumer ethnocentric tendencies with 17 indicators is hypothesized. Now suppose we collect data using these 17 indicators. The obvious question is: How well do the empirical data conform to the hypothesized factor model of consumer ethnocentric tendencies? That is, how well do the data fit the model? In other words, we want to do an empirical confirmation of the hypothesized factor model and, as such, confirmatory factor analysis can be viewed as a technique for theory testing (i.e., hypotheses testing).

In this chapter we discuss confirmatory factor analysis and LISREL, which is one of the many software packages available for estimating the parameters of a hypothesized factor model. The LISREL program is available in SPSS. For detailed discussion of confirmatory factor analysis and LISREL the reader is referred to Long (1983) and Hayduk (1987).

6.1 BASIC CONCEPTS OF CONFIRMATORY FACTOR ANALYSIS

In this section we use one-factor models and a correlated two-factor model to discuss the basic concepts of confirmatory factor models. However, we first provide a brief discussion regarding the type of matrix (i.e., covariance or correlation matrix) that is normally employed for exploratory and confirmatory factor analysis.

6.1.1 Covariance or Correlation Matrix?

Exploratory factor analysis typically uses the correlation matrix for estimating the factor structure because factor analysis was initially developed to explain correlations among the variables. Consequently, the covariance matrix has been used rarely in exploratory

[1]This example is based on Shimp and Sharma (1986).

factor analysis. Indeed, the factor analysis procedure in SPSS does not even give the option of using a covariance matrix. Recall that correlations measure covariations among the variables for standardized data, and the covariances measure covariations among the variables for mean-corrected data. Therefore, the issue regarding the use of correlation or covariance matrices reduces to the type of data used (i.e., mean-corrected or standardized). Just as in principal components analysis, the results of PCF and PAF are not scale invariant. That is, PCF and PAF factor analysis results for a covariance matrix could be very different from those obtained by using the correlation matrix. Traditionally, researchers have used correlation matrices for exploratory factor analysis.

Most of the confirmatory factor models are scale invariant. That is, the results are the same irrespective of whether a covariance or a correlation matrix is used. However, since theoretically the maximum likelihood procedure for confirmatory factor analysis is derived for covariance matrices, it is recommended that one should always employ the covariance matrix. Therefore, in subsequent discussions we will use covariances rather than correlations.

6.1.2 One-Factor Model

Consider the one-factor model depicted in Figure 6.1. Assume that $p = 2$; that is, a one-factor model with two indicators is assumed. As discussed in Chapter 5, the factor model given in Figure 6.1 can be represented by the following set of equations

$$x_1 = \lambda_1\xi + \delta_1; \qquad x_2 = \lambda_2\xi + \delta_2. \tag{6.1}$$

The covariance matrix, Σ, among the variables is given by

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix}. \tag{6.2}$$

Assuming that the variance of the latent factor, ξ, is one, the error terms (δ) and the latent construct are uncorrelated, and the error terms are uncorrelated with each other, the variances and covariances of the indicators are given by (see Eqs. A5.2 and A5.4 in the Appendix to Chapter 5)

$$\begin{aligned} \sigma_1^2 &= \lambda_1^2 + V(\delta_1); \qquad \sigma_2^2 = \lambda_2^2 + V(\delta_2) \\ \sigma_{12} &= \sigma_{21} = \lambda_1\lambda_2. \end{aligned} \tag{6.3}$$

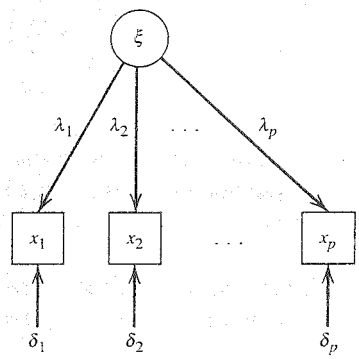


Figure 6.1 One-factor model.

In these equations, λ_1, λ_2, $V(\delta_1)$, and $V(\delta_2)$ are the model parameters, and it is obvious that the elements of the covariance matrix are functions of the model parameters. Let us define a vector, $\boldsymbol{\theta}$, that contains the model parameters; that is, $\boldsymbol{\theta}' = [\lambda_1, \lambda_2, V(\delta_1), V(\delta_2)]$. Substituting Eq. 6.3 into Eq. 6.2 we get

$$\boldsymbol{\Sigma}(\boldsymbol{\theta}) = \begin{pmatrix} \lambda_1^2 + V(\delta_1) & \lambda_1\lambda_2 \\ \lambda_1\lambda_2 & \lambda_2^2 + V(\delta_2) \end{pmatrix} \tag{6.4}$$

where $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the covariance matrix that would result for the parameter vector $\boldsymbol{\theta}$. Note that each parameter vector will result in a unique covariance matrix.

The problem in confirmatory factor analysis essentially reduces to estimating the model parameters (i.e., estimate $\boldsymbol{\theta}$) given the sample covariance matrix, $\mathbf{S}$. Let $\hat{\boldsymbol{\theta}}$ be the vector containing the parameter estimates. Now, given the parameter estimates, one can compute the estimated covariance matrix using Eq. 6.3. Let $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})$ be the estimated covariance matrix. The parameter estimates are obtained such that $\mathbf{S}$ is as close as possible to $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})$ (i.e., $\mathbf{S} = \hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})$). Hereafter, we will use $\hat{\boldsymbol{\Sigma}}$ to denote $\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\theta}})$.

In the two-indicator model discussed above we had three equations, one for each of the nonduplicated elements of the covariance matrix (i.e., σ_1^2, σ_2^2, and $\sigma_{12} = \sigma_{21}$).[2] But, there are four parameters to be estimated: λ_1, λ_2, $V(\delta_1)$, and $V(\delta_2)$. That is, the two-indicator factor model given in Figure 6.1 is underidentified as there are more parameters to be estimated than there are unique equations. In other words, in underidentified models the number of parameters to be estimated is greater than the number of unique pieces of information (i.e., unique elements) in the covariance matrix. An underidentified model can only be estimated if certain constraints or restrictions are placed on the parameters. For example, a unique solution may be obtained for the two-indicator model by assuming that $\lambda_1 = \lambda_2$ or $V(\delta_1) = V(\delta_2)$.

Now consider the model with three indicators. That is, $p = 3$ in Figure 6.1. Following is the set of equations linking the elements of the covariance matrix to the model parameters:

$$\begin{aligned} \sigma_1^2 &= \lambda_1^2 + V(\delta_1); & \sigma_2^2 &= \lambda_2^2 + V(\delta_2); & \sigma_3^2 &= \lambda_3^2 + V(\delta_3) \\ \sigma_{12} &= \lambda_1\lambda_2; & \sigma_{13} &= \lambda_1\lambda_3; & \sigma_{23} &= \lambda_2\lambda_3. \end{aligned}$$

We now have six equations and six parameters to be estimated. This model, therefore, is just-identified and will result in an exact solution.

Next, consider the four-indicator model (i.e., $p = 4$ in Figure 6.1). The following set of ten equations links the elements of the covariance matrix to the parameters of the model:

$$\begin{aligned} \sigma_1^2 &= \lambda_1^2 + V(\delta_1); & \sigma_2^2 &= \lambda_2^2 + V(\delta_2); & \sigma_3^2 &= \lambda_3^2 + V(\delta_3); & \sigma_4^2 &= \lambda_4^2 + V(\delta_4) \\ \sigma_{12} &= \lambda_1\lambda_2; & \sigma_{13} &= \lambda_1\lambda_3; & \sigma_{14} &= \lambda_1\lambda_4 \\ \sigma_{23} &= \lambda_2\lambda_3; & \sigma_{24} &= \lambda_2\lambda_4; & \sigma_{34} &= \lambda_3\lambda_4. \end{aligned}$$

The four-indicator model is overidentified as there are ten equations and only eight parameters to be estimated, resulting in two overidentifying equations—the difference between the number of nonduplicated elements (i.e., equations) of the covariance matrix and the number of parameters to be estimated.

Thus, factor models are under-, just-, or overidentified. Obviously, an underidentified model cannot be estimated and, furthermore, a unique solution does not exist

[2]In general, the number of nonduplicated elements of the covariance matrix will be equal to $[p(p+1)]/2$, where p is the number of indicators.

for an underidentified model. A just-identified model, though estimatable, is not very informative as an exact solution exists for any sample covariance matrix. That is, the fit between the estimated and the sample covariance matrix will always be perfect (i.e., $\hat{\Sigma} = S$), and therefore it is not possible to determine if the model fits the data. On the other hand, the overidentified model will, in general, not result in a perfect fit. The fit of some models might be better than the fit of other models, thus making it possible to assess the fit of the model to the data. The overidentifying equations are the degrees of freedom for hypothesis testing. In the case of the four-indicator model there are two degrees of freedom because there are two overidentifying equations. For a p-indicator model, there will be $p(p+1)/2 - q$ overidentifying equations or degrees of freedom where q is the number of parameters to be estimated.

6.1.3 Two-Factor Model with Correlated Constructs

Consider the two-factor model shown in Figure 6.2 and represented by the following equations:

$$x_1 = \lambda_1\xi_1 + \delta_1; \qquad x_2 = \lambda_2\xi_1 + \delta_2$$
$$x_3 = \lambda_3\xi_2 + \delta_3; \qquad x_4 = \lambda_4\xi_2 + \delta_4.$$

Notice that the two-factor model hypothesizes that x_1 and x_2 are indicators of ξ_1, and x_3 and x_4 are indicators of ξ_2. Furthermore, it hypothesizes that the two factors are correlated. Thus, the exact nature of the two-factor model is hypothesized *a priori*. No such *a priori* hypotheses for factor models discussed in the previous chapter are made. This is one of the major differences between confirmatory factor analysis and the exploratory factor analysis discussed in Chapter 5.

The following set of equations gives the relationship between model parameters and the elements of the covariance matrix.

$$\sigma_1^2 = \lambda_1^2 + V(\delta_1); \qquad \sigma_2^2 = \lambda_2^2 + V(\delta_2); \qquad \sigma_3^2 = \lambda_3^2 + V(\delta_3); \qquad \sigma_4^2 = \lambda_4^2 + V(\delta_4)$$
$$\sigma_{12} = \lambda_1\lambda_2; \qquad \sigma_{13} = \lambda_1\lambda_3\phi; \qquad \sigma_{14} = \lambda_1\lambda_4\phi$$
$$\sigma_{23} = \lambda_2\lambda_3\phi; \qquad \sigma_{24} = \lambda_2\lambda_4\phi; \qquad \sigma_{34} = \lambda_3\lambda_4,$$

where ϕ is the covariance between the two latent constructs. There are ten equations and nine parameters to be estimated (four loadings, four unique-factor variances, and the covariance between the two latent factors) resulting in one degree of freedom.

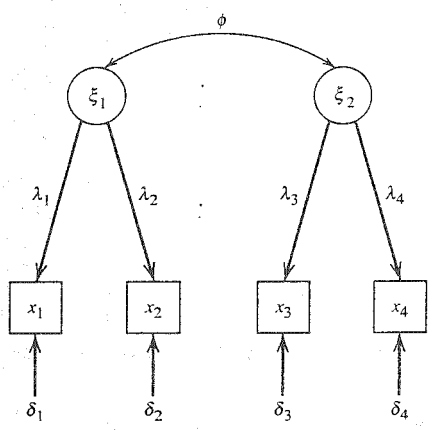


Figure 6.2 Two-factor model with correlated constructs.

6.2 OBJECTIVES OF CONFIRMATORY FACTOR ANALYSIS

The objectives of confirmatory factor analysis are:

- Given the sample covariance matrix, to estimate the parameters of the hypothesized factor model
- To determine the fit of the hypothesized factor model. That is, how close is the estimated covariance matrix, $\hat{\Sigma}$, to the sample covariance matrix, S?

The parameters of confirmatory factor models can be estimated using the maximum likelihood estimation technique. Section A6.2 of the Appendix contains a brief discussion of this technique, which facilitates hypotheses testing for model fit and significance tests for the parameter estimates. The maximum likelihood estimation technique, which assumes that the data come from a multivariate normal distribution, is employed by a number of computer programs such as EQS in **BMDP** (Bentler 1982), LISREL in **SPSS** (Jöreskog and Sörbom 1989), and CALIS in **SAS** (SAS 1993). Stand-alone PC versions of LISREL and EQS are also available. In the following section we discuss LISREL, as it is the most widely used program.

6.3 LISREL

LISREL (an acronym for *li*near *s*tructural *rel*ations) is a general-purpose program for estimating a variety of covariance structure models, with confirmatory factor analysis being one of them. We begin by first discussing the terminology used by the LISREL program.

6.3.1 LISREL Terminology

Consider the p-indicator one-factor model depicted in Figure 6.1. The model can be represented by the following equations:

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \delta_1 \\ x_2 &= \lambda_{21}\xi_1 + \delta_2 \\ &\vdots \\ x_p &= \lambda_{p1}\xi_1 + \delta_p. \end{aligned}$$

These equations can be represented as:

$$\begin{pmatrix} x_1 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} \lambda_{11} \\ \vdots \\ \lambda_{p1} \end{pmatrix} \xi_1 + \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_p \end{pmatrix},$$

where λ_{ij} is the loading of the ith indicator on the jth factor, ξ_j is the jth construct or factor, δ_i is the unique factor (commonly referred to as the error term for the ith indicator), and $i = 1, \ldots, p$ and $j = 1, \ldots, m$. Note that p is the number of indicators and m is the number of factors, which is one in the present case.

The preceding equations can be written in matrix form as

$$\mathbf{x} = \Lambda_x \boldsymbol{\xi} + \boldsymbol{\delta} \tag{6.5}$$

where $\mathbf{x}$ is a $p \times 1$ vector of indicators, Λ_x is a $p \times m$ matrix of factor loadings, ξ is an $m \times 1$ vector of latent constructs (factors), and δ is a $p \times 1$ vector of errors (i.e., unique factors) for the p indicators.[3] The covariance matrix for the indicators is given by (see Eq. A5.16 in the Appendix to Chapter 5)

$$\Sigma = \Lambda_x \Phi \Lambda_x' + \Theta_\delta \tag{6.6}$$

where Λ_x is a $p \times m$ parameter matrix of factor loadings, Φ is an $m \times m$ parameter matrix containing the variances and covariances of the latent constructs, and Θ_δ is a $p \times p$ parameter matrix of the variances and covariances of the error terms. Table 6.1 gives the symbols that LISREL uses to represent the various parameter matrices (i.e., Λ_x, Φ, and Θ_δ).

The parameters of the factor model can be fixed, free, and/or constrained. Free parameters are those that are to be estimated. Fixed parameters are those that are not estimated; their values are provided, i.e., fixed, at the value specified by the researcher. Constrained parameters are estimated; however, their values are constrained to be equal to other free parameters. For example, one could hypothesize that all the indicators are measured with the same amount of error. In this case, the variances of the errors for all the indicators would be constrained to be equal. Use of constrained parameters is discussed in Section 6.5.

In the following section we illustrate the use of LISREL to estimate the parameter matrices of confirmatory factor models. The correlation matrix given in Table 5.2, which is reproduced in Table 6.2, will be used. A one-factor model with six indicators is hypothesized, and our objective is to test the model using sample data. In order to convert the correlation matrix into the covariance matrix, we arbitrarily assume that the standard deviation of each variable is two.

Table 6.1 **Symbols Used by LISREL To Represent Parameter Matrices**

Parameter Matrix	LISREL Symbol	Order
Λ_x	LX	$p \times m$
Φ	PHI	$m \times m$
Θ_δ	TD	$p \times p$

Table 6.2 **Correlation Matrix**

	M	P	C	E	H	F
M	1.000					
P	0.620	1.000				
C	0.540	0.510	1.000			
E	0.320	0.380	0.360	1.000		
H	0.284	0.351	0.336	0.686	1.000	
F	0.370	0.430	0.405	0.730	0.735	1.000

[3]Henceforth we refer to the unique factors as errors, as this is the term used in confirmatory factor models to represent the unique factors.

6.3.2 LISREL Commands

Table 6.3 gives the commands. Commands before the LISREL command are standard SPSS commands for reading the correlation matrix and the standard deviations, and for converting the correlation matrix into a covariance matrix. The remaining commands are LISREL commands, which are briefly described below. The reader is strongly advised to refer to the LISREL manual (Jöreskog and Sörbom 1989) for a detailed discussion of these commands.

1. The TITLE command is the first command and is used to give a title to the model being analyzed.
2. The DATA command gives information about the input data. Following are the various options specified in the data command:
   (a) The NI option specifies the total number of indicators or variables in the model, which in the present case is equal to 6.

Table 6.3 **LISREL Commands for the One-Factor Model**

```
TITLE LISREL in SPSS
MATRIX DATA VARIABLES=M P C E H F/CONTENTS=CORR STDDEV/N=200
BEGIN DATA
insert correlation matrix here
2 2 2 2 2 2
END DATA
MCONVERT
LISREL
 /TITLE "ONE FACTOR MODEL"
 /DATA NI=6 NO=200 MA=CM
 /LABELS
 /'M' 'P' 'C' 'E' 'H' 'F'
 /MODEL NX=6 NK=1 TD=SY LX=FU PHI=SY
 /LK
 /'IQ'
 /PA LX
 /0
 /1
 /1
 /1
 /1
 /1
 /PA PHI
 /1
 /PA TD
 /1
 /0 1
 /0 0 1
 /0 0 0 1
 /0 0 0 0 1
 /0 0 0 0 0 1
 /VALUE 1.0 LX(1,1)
 /OUTPUT TV RS MI SS SC TO
FINISH
```

(b) The NO option specifies the number of observations used to compute the sample covariance matrix.

(c) The MA option specifies whether the correlation or covariance matrix is to be used for estimating model parameters. MA=KM implies that the correlation matrix should be used and MA=CM implies that the covariance matrix should be used. It is usually recommended that the covariance matrix be used, as the maximum likelihood estimation procedure is derived for the covariance matrix.

3. The LABELS command is optional and is used to assign labels to the indicators. In the absence of the LABELS command the variables are labeled as VAR1, VAR2, and so on. The labels are read in free format and are enclosed in single quotation marks. The labels cannot be longer than eight characters.

4. The MODEL command specifies model details. Following are the options for the MODEL command:

   (a) NX specifies the number of indicators for the factor model. In this case the number of indicators specified in the NX and NI commands are the same. However, this is not always true as LISREL is a general-purpose program for analyzing a variety of models. This distinction will become clear in a later chapter where we discuss the use of LISREL to analyze structural equation models.

   (b) NK specifies the number of factors in the model.

   (c) TD=SY specifies that the $p \times p$ Θ_δ matrix is symmetric. LX=FU specifies that the Λ_x is a $p \times m$ full matrix and PHI=SY specifies that the $m \times m$ Φ matrix is symmetric.

5. The LK command is optional and is used to assign labels to the latent constructs. As can be seen, the label 'IQ' is assigned to the latent construct.

These commands are followed by pattern matrices, which are used to specify which parameters are fixed and which are free. The elements of the pattern matrices contain zeroes and ones. A zero indicates that the corresponding parameter is fixed and a one indicates that it is free, i.e., the parameter is to be estimated. All fixed parameters, unless otherwise specified by alternative commands, are fixed to have a value of zero. The pattern matrices are as follows (once again, the pattern matrices are read in free format):

6. The PA LX command is used to specify the structure or the pattern of the LX (i.e., Λ_x) matrix. It can be seen that in the present case all the loadings except LX(1,1) (i.e., λ_{11}) are to be estimated. The reason for fixing λ_{11} will be discussed later.

7. The PA PHI command specifies the structure for the PHI matrix. For the present model, the only element of this matrix is specified as free, i.e., it is to be estimated.

8. The PA TD command specifies the structure for the covariance matrix of the error terms. Note that the variances of all the error terms are to be estimated, and it is assumed that the covariances among the errors are zero.

9. The VALUE command is used to specify alternative values for the fixed parameters. In the present case the value of the fixed parameter, LX(1,1), is set to 1.0. Values for all other fixed parameters are set to zero.

10. The OUTPUT command specifies the types of output desired. The following output is requested:

    TV the t-values for each parameter estimate.

    RS the residual matrix (i.e., $\mathbf{S} - \hat{\Sigma}$).

MI the modification indices.
SS the standardized solution.
SC the completely standardized solution.
TO requests that an 80-column format be used for printing the output.

Note that all the parameters of this model, except LX(1,1), are free. That is, the loading for one of the indicators is fixed to one. This is done for a specific reason. Most of the latent constructs such as attitudes, intelligence, resistance to innovation, and excellence do not have a natural measurement scale. Therefore, we have to define the metric or the scale for the latent construct. Usually the scale of the latent construct is defined such that it is the same as that of one of the indicators used to measure the construct. This is done by fixing the loading of one of its indicators to one.[4] For example, if λ_{11} is fixed to one then the equation linking x_1 and ξ_1 will be

$$x_1 = \xi_1 + \delta_1.$$

Since δ_1 is assumed to be random error and its expected value is equal to zero, the scale of ξ_1 will be the same as that of x_1.

6.4 INTERPRETATION OF THE LISREL OUTPUT

Exhibit 6.1 gives the LISREL output. As before, the output is labeled for discussion purposes, with bracketed numbers in the text corresponding to the circled numbers in the exhibit.

6.4.1 Model Information and Parameter Specifications

This part of the output simply lists information about the model and the requested output as specified in the LISREL commands [1]. Also, the matrix to be analyzed is printed [2]. The parameter specification section indicates which parameters are free (i.e., to be estimated) and which are fixed [3]. A fixed parameter is indicated by an entry of zero in the corresponding element of the pattern matrix. All the free parameters are numbered sequentially. Note that a total of 12 parameters are to be estimated: five loadings, the variance of the latent construct, and variances of the six error terms.

6.4.2 Initial Estimates

This part of the output gives the initial estimates obtained using the two-stage least-squares (TSLS) approach. These estimates are used as starting values by the maximum likelihood procedure and are usually not interpreted [4]. Since the maximum likelihood estimation technique uses an iterative procedure, it is quite possible that the solution may not converge in the default number of iterations, which is 250.[5] LISREL does give the option of increasing the number of iterations; however, caution is advised against

[4]If all the parameters are freed, the model cannot be estimated as no scale is defined for the latent construct. That is, the scale for each latent construct must be defined. Scales for the latent constructs are typically defined by fixing the value of one of its indicators to one, or by fixing the variance of the latent construct to one.

[5]The default number of iterations could be different for different programs and for different versions of LISREL.

Exhibit 6.1

LISREL output for the one-factor model

```
(1) 1TITLE ONE FACTOR MODEL
  0                          NUMBER OF INPUT VARIABLES  6
  0                          NUMBER OF Y - VARIABLES    0
  0                          NUMBER OF X - VARIABLES    6
  0                          NUMBER OF ETA - VARIABLES  0
  0                          NUMBER OF KSI - VARIABLES  1
  0                          NUMBER OF OBSERVATIONS   200
(2) 1TITLE ONE FACTOR MODEL
  0        COVARIANCE MATRIX TO BE ANALYZED
  0                 M        P        C        E        H        F
  +           -------- -------- -------- -------- -------- --------
          M     4.000
          P     2.480    4.000
          C     2.160    2.040    4.000
          E     1.280    1.520    1.440    4.000
          H     1.136    1.404    1.344    2.744    4.000
          F     1.480    1.720    1.620    2.920    2.940    4.000
(3) 1TITLE ONE FACTOR MODEL
  0PARAMETER SPECIFICATIONS
  0        LAMBDA X
  0                IQ
  +           --------
          M          0
          P          1
          C          2
          E          3
          H          4
          F          5
  0        PHI
  0                IQ
  +           --------
         IQ          6
  0        THETA DELTA
  0                 M        P        C        E        H        F
  +           -------- -------- -------- -------- -------- --------
          M          7
          P          0        8
          C          0        0        9
          E          0        0        0       10
          H          0        0        0        0       11
          F          0        0        0        0        0       12
(4) 1TITLE ONE FACTOR MODEL
  0INITIAL ESTIMATES (TSLS)
  0        LAMBDA X
  0                IQ
  +           --------
          M      1.000
          P      0.991
          C      0.855
          E      0.688
          H      0.658
          F      0.758
```

(continued)

Exhibit 6.1 (continued)

```
0        PHI
0                IQ
+           --------
       IQ      2.636
0        THETA DELTA
0                M          P          C          E          H          F
+           --------   --------   --------   --------   --------   --------
        M      1.364
        P      0.000      1.410
        C      0.000      0.000      2.073
        E      0.000      0.000      0.000      2.751
        H      0.000      0.000      0.000      0.000      2.860
        F      0.000      0.000      0.000      0.000      0.000      2.487
0        SQUARED MULTIPLE CORRELATIONS FOR X - VARIABLES
0                M          P          C          E          H          F
+           --------   --------   --------   --------   --------   --------
            0.659      0.648      0.482      0.312      0.285      0.378
0        TOTAL COEFFICIENT OF DETERMINATION FOR X - VARIABLES IS  0.860

5a 1TITLE ONE FACTOR MODEL
   0LISREL ESTIMATES (MAXIMUM LIKELIHOOD)
0        LAMBDA X
0                IQ
+           --------
        M      1.000
        P      1.134
        C      1.073
        E      1.786
        H      1.770
        F      1.937
5b 0        PHI
0                IQ
+           --------
       IQ      0.836
5c 0        THETA DELTA
0                M          P          C          E          H          F
+           --------   --------   --------   --------   --------   --------
        M      3.164
        P      0.000      2.925
        C      0.000      0.000      3.037
        E      0.000      0.000      0.000      1.334
        H      0.000      0.000      0.000      0.000      1.381
        F      0.000      0.000      0.000      0.000      0.000      0.863
5d 0        SQUARED MULTIPLE CORRELATIONS FOR X - VARIABLES
0                M          P          C          E          H          F
+           --------   --------   --------   --------   --------   --------
            0.209      0.269      0.241      0.667      0.655      0.784
5e 0        TOTAL COEFFICIENT OF DETERMINATION FOR X - VARIABLES IS  0.895

6 0         CHI-SQUARE WITH   9 DEGREES OF FREEDOM =   113.02 (P = .000)
0                   GOODNESS OF FIT INDEX =0.822
              ADJUSTED GOODNESS OF FIT INDEX =0.584
                   ROOT MEAN SQUARE RESIDUAL =     0.507
```

(continued)

Exhibit 6.1 (continued)

```
7a 1TITLE ONE FACTOR MODEL
 0        FITTED COVARIANCE MATRIX
 0                M        P        C        E        H        F
 +          --------  -------- -------- -------- -------- --------
        M      4.000
        P      0.948    4.000
        C      0.897    1.018    4.000
        E      1.493    1.693    1.602    4.000
        H      1.480    1.678    1.588    2.643    4.000
        F      1.620    1.837    1.738    2.892    2.866    4.000
7b 0        FITTED RESIDUALS
 0                M        P        C        E        H        F
 +          --------  -------- -------- -------- -------- --------
        M      0.000
        P      1.532    0.000
        C      1.263    1.022    0.000
        E     -0.213   -0.173   -0.162    0.000
        H     -0.344   -0.274   -0.244    0.101    0.000
        F     -0.140   -0.117   -0.118    0.028    0.074    0.000

7c 0        STANDARDIZED RESIDUALS
 0                M        P        C        E        H        F
 +          --------  -------- -------- -------- -------- --------
        M      0.000
        P      7.402    0.000
        C      5.966    5.063    0.000
        E     -1.747   -1.503   -1.370    0.000
        H     -2.741   -2.310   -2.002    1.941    0.000
        F     -1.689   -1.510   -1.479    0.952    2.416    0.000

8 1TITLE ONE FACTOR MODEL
 -T-VALUES
 0        LAMBDA X
 0                IQ
 +          --------
        M      0.000
        P      5.210
        C      5.046
        E      6.393
        H      6.375
        F      6.533
 0        PHI
 0                IQ
 +          --------
       IQ      3.234
 0        THETA DELTA
 0                M        P        C        E        H        F
 +          --------  -------- -------- -------- -------- --------
        M      9.661
        P      0.000    9.537
        C      0.000    0.000    9.598
        E      0.000    0.000    0.000    7.410
        H      0.000    0.000    0.000    0.000    7.556
        F      0.000    0.000    0.000    0.000    0.000    5.423
```

(continued)

Exhibit 6.1 (continued)

```
9  1TITLE ONE FACTOR MODEL
   -STANDARDIZED SOLUTION
   0        LAMBDA X
   0                IQ
   +           --------
          M      0.914
          P      1.037
          C      0.981
          E      1.633
          H      1.618
          F      1.771
   0        PHI
   0                IQ
   +           --------
         IQ      1.000
10a 1TITLE ONE FACTOR MODEL
   -COMPLETELY STANDARDIZED SOLUTION
   0        LAMBDA X
   0                IQ
   +           --------
          M      0.457
          P      0.518
          C      0.491
          E      0.816
          H      0.809
          F      0.886

10b 0        PHI
   0                IQ
   +           --------
         IQ      1.000
10c 0        THETA DELTA
   0                M          P          C          E          H          F
   +           --------   --------   --------   --------   --------   --------
          M      0.791
          P      0.000      0.731
          C      0.000      0.000      0.759
          E      0.000      0.000      0.000      0.333
          H      0.000      0.000      0.000      0.000      0.345
          F      0.000      0.000      0.000      0.000      0.000      0.216
11 1TITLE ONE FACTOR MODEL
   -MODIFICATION INDICES AND ESTIMATED CHANGE
   0NO NON-ZERO MODIFICATION INDICES FOR LAMBDA X
   0NO NON-ZERO MODIFICATION INDICES FOR PHI
   0        MODIFICATION INDICES FOR THETA DELTA
   0                M          P          C          E          H          F
   +           --------   --------   --------   --------   --------   --------
          M      0.000
          P     54.791      0.000
          C     35.598     25.630      0.000
          E      3.054      2.259      1.878      0.000
          H      7.511      5.337      4.007      3.766      0.000
          F      2.854      2.280      2.187      0.906      5.839      0.000
   0      MAXIMUM MODIFICATION INDEX IS   54.79 FOR ELEMENT (2, 1)
          OF THETA DELTA
```

arbitrarily increasing the number of iterations. A more prudent approach is to consider the factors that could lead to nonconvergence of solutions and rectify them.

One of the common reasons for nonconvergence is related to the start values used. Since the maximum likelihood value is an iterative procedure, in some models the iterative procedure could become sensitive to the start values employed. LISREL gives the user the option of specifying different start values. The researcher could use several start values to see if convergence can be achieved and if the solutions obtained by using different start values are the same.

A second reason the solution may not converge is that the model is large and is estimating too many parameters. In such cases the only option the researcher has is to reduce the size of the model. A third reason for nonconvergence is that the models are misspecified. In such cases the researcher should carefully examine the model using the underlying theory.

6.4.3 Evaluating Model Fit

The first step in interpreting the results of confirmatory factor models is to assess the overall model fit. If the model fit is adequate and acceptable to the researcher, then one can proceed with the evaluation and interpretation of the estimated model parameters. The overall model fit can be assessed statistically by the χ^2 test, and heuristically using a number of goodness-of-fit indices.

The χ^2 Test

The χ^2 statistic is used to test the following null and alternative hypotheses

$$H_o : \Sigma = \Sigma(\theta)$$
$$H_a : \Sigma \neq \Sigma(\theta)$$

where Σ is the population matrix and $\Sigma(\theta)$ is the estimated covariance matrix that would result from the vector of parameters defining the hypothesized model. To test the above hypotheses, the sample covariance matrix $\mathbf{S}$ is used as an estimate of Σ, and $\Sigma(\hat{\theta}) = \hat{\Sigma}$ is the estimate of the covariance matrix $\Sigma(\theta)$ obtained from the parameter estimates, i.e., $\hat{\Sigma}$ is the estimated covariance matrix. The null hypothesis then becomes a test of $\mathbf{S} = \hat{\Sigma}$ or $\mathbf{S} - \hat{\Sigma} = \mathbf{0}$. That is, the null hypothesis tests whether the difference between the sample and the estimated covariance matrix is a null or zero matrix. Note that in the present case failure to reject the null hypothesis is desired, as it leads to the conclusion that statistically the hypothesized model fits the data. A χ^2 value of zero results if $\mathbf{S} - \hat{\Sigma} = \mathbf{0}$. The χ^2 value of 113.02 with 9 degrees of freedom (i.e., 21 − 12) is significant at $p < .000$ thus rejecting the null hypothesis [6].[6] That is, statistically the one-factor model does not fit the data.

Heuristic Measures of Model Fit

The χ^2 statistic is sensitive to sample size. For a large sample size, even small differences in $\mathbf{S} - \hat{\Sigma}$ will be statistically significant although the differences may not be practically meaningful. Consequently, researchers typically tend to discount the χ^2 test and resort to other methods for evaluating the fit of the model to the data (Bearden, Sharma, and Teel 1982). Over thirty goodness-of-fit indices for evaluating model fit

[6]Recall that there are 21 equations (i.e., the number of nonduplicated elements of the covariance matrix) and there are 12 parameters to be estimated.

have been proposed in the literature (see Marsh, Balla, and McDonald (1988) for a review of these statistics). Most of the fit indices are designed to provide a summary measure of the residual matrix, which is the difference between the sample and the estimated covariance matrix (i.e., $\mathbf{RES} = \mathbf{S} - \hat{\boldsymbol{\Sigma}}$). Version 7 of LISREL reports three such measures: goodness-of-fit index (*GFI*); *GFI* adjusted for degrees of freedom (*AGFI*); and root mean square residual (*RMSR*).[7]

GOODNESS-OF-FIT INDEX. *GFI* is obtained using the following formula:

$$GFI = 1 - \frac{\mathrm{tr}[(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S} - \mathbf{I})^2]}{\mathrm{tr}[(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{S})^2]}, \tag{6.7}$$

and represents the amount of variances and covariances in $\mathbf{S}$ that are predicted by the model. In this sense it is analogous in interpretation to R^2 in multiple regression. Note that $GFI = 1$ when $\mathbf{S} = \hat{\boldsymbol{\Sigma}}$ (i.e., $\mathbf{RES} = \mathbf{0}$) and $GFI =< 1$ for hypothesized models that do not perfectly fit the data. However, it has been shown that *GFI*, and consequently *AGFI*, is affected by sample size and the number of indicators and that the upper bound for *GFI* may not be one.

Maiti and Mukherjee (1990) have derived the approximate sampling distribution for *GFI* under the assumption that the null hypothesis is true. The approximate expected value for *GFI* is given by

$$EGFI \approx \frac{1}{1 + (2df/pn)}, \tag{6.8}$$

where *EGFI* is the approximate expected value of *GFI*, p is the number of indicators, df are the degrees of freedom, and n is the sample size. In general, for factor models the df/p will increase as the number of indicators increases. Consequently, for a given sample size, as p increases *EGFI* decreases and vice versa. This relationship is illustrated in Panel II of Figure 6.3 for a one-factor model and a sample size of 200.

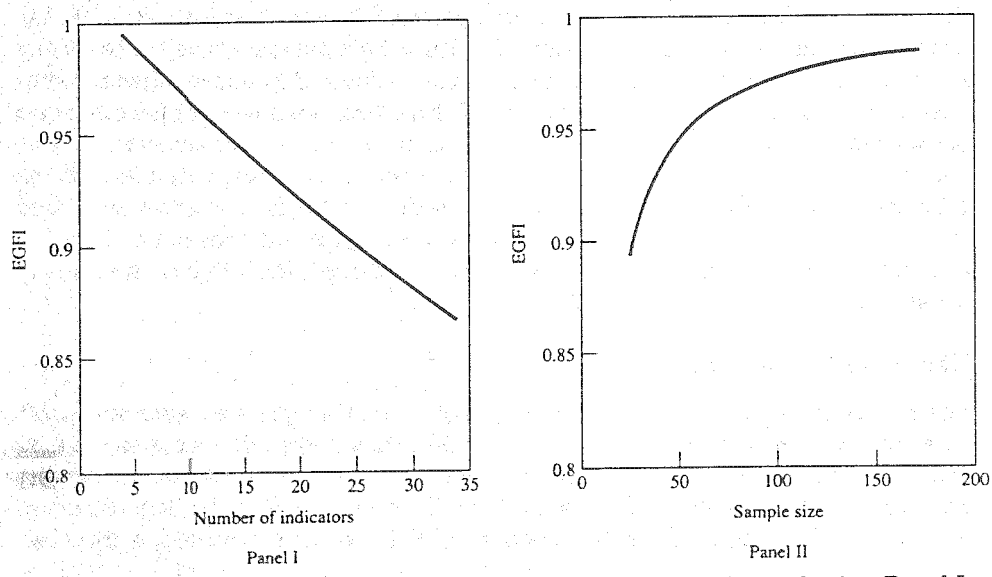


Figure 6.3 EGFI as a function of the number of indicators and sample size. Panel I: EGFI as a function of number of indicators. Panel II: EGFI as a function of sample size.

[7] Version 8 of LISREL reports the additional fit indices proposed by various researchers.

Similarly, for a given number of indicators *EGFI* increases as the sample size increases and vice versa. Panel I of Figure 6.3 illustrates this relationship for a seven-indicator model. Therefore, we suggest that rather than using *GFI*, one should use a relative goodness-of-fit index (*RGFI*), which can be computed as

$$RGFI = \frac{GFI}{EGFI}. \tag{6.9}$$

From Eq. 6.8 the *EGFI* for this example is equal to [6]

$$\begin{aligned} EGFI &= \frac{1}{1 + [(2 \times 9)/(6 \times 200)]} \\ &= .985, \end{aligned}$$

and from Eq. 6.9 the *RGFI* is equal to 0.835 (i.e., 0.822/0.985) [6]. Once again the question becomes: How high is high? One rule of thumb is that the *GFI* for good-fitting models should be greater than 0.90. The value of 0.835 for *RGFI* does not meet this criterion. Since *RGFI* is less than the suggested cutoff value of 0.90, one would conclude that the model does not fit the data. It should be noted that cutoff values are completely arbitrary. Consequently, we recommend that the goodness-of-fit indices should be used to assess the fit of a number of competing models fitted to the same data set, rather than the fit of a single model.

ADJUSTED GOODNESS-OF-FIT INDEX. The *AGFI*, analogous to adjusted R^2 in multiple regression, is essentially *GFI* that has been adjusted for degrees of freedom. *AGFI* is given as

$$AGFI = 1 - \left[\frac{p(p+1)}{2df}\right][1 - GFI]. \tag{6.10}$$

Once again, there are no guidelines regarding how high *AGFI* should be for good-fitting models, but researchers have typically used a value of 0.80 as the cutoff value. Since *AGFI* is a simple transformation of *GFI*, the expected value of *AGFI* (i.e., *EAGFI*) can be obtained by substituting *EGFI* in Eq. 6.10. This gives a value of 0.965 for *EAGFI*, resulting in a relative value of *AGFI* (i.e., *RAGFI*) equal to 0.605 (i.e., 0.584/0.965), which is less than the suggested cutoff value of 0.80, thereby indicating once again an inadequate model fit.

ROOT MEAN SQUARE RESIDUAL. The *RMSR* is given by

$$RMSR = \sqrt{\frac{\sum_{i=1}^{p}\sum_{j=1}^{i}(s_{ij} - \hat{\sigma_{ij}})^2}{p(p+1)/2}}. \tag{6.11}$$

Note that *RMSR* is the square root of the average of the square of the residuals. The larger the *RMSR*, the less is the fit between the model and the data and vice versa. Unfortunately, the residuals are scale dependent and do not have an upper bound. Therefore, it is normally recommended that the *RMSR* of a model should be interpreted relative to the *RMSR* of other competing models fitted to the same data set.

OTHER FIT INDICES. As indicated earlier, a number of other fit indices have been proposed to evaluate model fit. Marsh, Balla, and McDonald (1988) and McDonald and Marsh (1990) compared a number of fit indices, including *GFI* and *AGFI*, and concluded that the following fit indices are relatively less sensitive to sample size: (1) rescaled noncentrality parameter (*NCP*); (2) McDonald's transformation of the

noncentrality parameter (*MDN*); (3) the Tucker-Lewis index (*TLI*); and (4) the relative noncentrality index (*RNI*).[8] Each of these fit indices is discussed below.

The formulae for computing *NCP* and *MDN* are

$$NCP = \frac{\chi^2 - df}{n} \tag{6.12}$$

$$MDN = e^{-0.5 \times NCP}. \tag{6.13}$$

From these equations it is obvious that *NCP* ranges from zero to infinity; however, its transformation, *MDN*, ranges from zero to one. Good model fit is suggested by high values for *MDN*.

The *TLI* and *RNI* are relative fit indices—they are based on comparison of the fit of the hypothesized model relative to some baseline model. The typical baseline model used is one that hypothesizes no relationship between the indicators and the factor, and is normally referred to as the *null model*.[9] That is, all the factor loadings are assumed to be zero and the variances of the error terms are the only model parameters that are estimated. The null model is represented by the following equations:

$$M = \delta_1; \qquad P = \delta_2; \qquad C = \delta_3$$
$$E = \delta_4; \qquad H = \delta_5; \qquad F = \delta_6.$$

Table 6.4 gives the LISREL commands for the null model, and the resulting fit statistics are: $\chi^2 = 564.67$ with 15 *df*, $GFI = .451$, $AGFI = .231$, and $RMSR = 1.670$. The formulae for computing *TLI* and *RNI* are

$$TLI = \frac{NCP_n/df_n - NCP_h/df_h}{NCP_n/df_n} \tag{6.14}$$

$$RNI = \frac{NCP_n - NCP_h}{NCP_n} \tag{6.15}$$

where NCP_h is the NCP for the hypothesized model, NCP_n is the NCP for the null model, df_h are the degrees of freedom for the hypothesized model, and df_n are the degrees of freedom for the null model. It can be seen that *TLI* and *RNI* represent the increase in model fit relative to a baseline model, which in the present case is the null model.

Computations for the values of *NCP*, *MDN*, *TLI*, and *RNI* are shown in Table 6.5. Once again we are faced with the issue of cutoff values to be used for assessing model fit. Traditionally researchers have used cutoff values of .90. None of the goodness-of-fit indices exceed the suggested cutoff value and, as before, we conclude that the model does not fit the data.

The Residual Matrix

All the fit indices discussed in the preceding section are summary measures of the RES matrix and provide an overall measure of model fit. In many instances, especially when the model does not fit the data, further analysis of the RES matrix can provide meaningful insights regarding model fit. A brief discussion of the RES matrix follows.

The RES matrix, labeled the *fitted residuals* matrix, contains the variances and covariances that have not been explained by the model [7b]. Obviously, the larger the

[8] Version 8 of LISREL reports these and other indices.

[9] The researcher is free to use any baseline model that he or she desires as the null model. For an interesting discussion of this point see Sobel and Bohrnstedt (1985).

Table 6.4 **LISREL Commands for the Null Model**

```
LISREL
 /"NULL MODEL"
 /DATA NI=6 NO=200 MA=CM
 /LABELS FO
 /'M' 'P' 'C' 'E' 'H' 'F'
 /MODEL NX=6 NK=1 TD=SY
 /LK
 /'IQ'
 /PA LX
 /0
 /0
 /0
 /0
 /0
 /0
 /PA PHI
 /0
 /PA TD
 /1
 /0 1
 /0 0 1
 /0 0 0 1
 /0 0 0 0 1
 /0 0 0 0 0 1
 /VALUE 1.0 PHI(1,1)
 /OUTPUT TV RS MI SC TO
FINISH
```

Table 6.5 **Computations for *NCP*, *MDN*, *TLI*, and *RNI* for the One-Factor Model**

1.

$$NCP_n = \frac{564.67 - 15}{200} = 2.748.$$

2.

$$NCP_h = \frac{113.02 - 9}{200} = .520; \quad MDN = e^{-0.5\times.520} = .771.$$

$$TLI = \frac{2.748/15 - .520/9}{2.748/15} = .685; \quad RNI = \frac{2.748 - .520}{2.748} = .811.$$

residuals the worse the model fit and vice versa. It can be clearly seen that the residuals of the covariances among indicators M, P, and C are large compared to residuals of covariances among other indicators. This suggests that the model is unable to adequately explain the relationships among M, P, and C. But how large should the residuals be before one can say that the hypothesized model is not able to adequately explain the covariances among these three indicators? Unfortunately, the residuals in the **RES**

matrix are scale dependent. To overcome this problem, the **RES** matrix is standardized by dividing the residuals by their respective asymptotic standard errors. The resulting standardized residual matrix is also reported by LISREL [7c]. Standardized residuals that are greater than 1.96 (the critical Z value for $\alpha = .05$) are considered to be statistically significant and, therefore, high. Ideally, no more than 5% of standardized residuals should be greater than 1.96. From the standardized **RES** matrix it is clear that 46.67% (7 of the 15 covariance residuals) are greater than 1.96, suggesting that the hypothesized model does not fit the data. If too many of the standardized residuals are greater than 1.96, then we should take a careful look at the data or the hypothesized model.[10] We seem to have resolved the "how high is high" issue by using standardized residuals, but the issue of the sensitivity of a statistical test to sample size resurfaces. That is, for large samples even small residual covariances will be statistically significant. For this reason, many researchers tend to ignore the interpretation of the standardized residuals and simply look for residuals in the **RES** matrix that are *relatively* large and use the *RMSR* as a summary measure of the **RES** matrix.

Summary of Model Fit Assessment

Model fit was assessed using the χ^2 statistic and a number of goodness-of-fit indices. The χ^2 statistic formally tests the null and alternative hypotheses where the null hypothesis is that the hypothesized model fits the data and the alternative hypothesis is that some model other than the one hypothesized fits the data. It was seen that the χ^2 statistic indicated that the one-factor model did not fit the data. However, the χ^2 statistic is quite sensitive to sample size in that for a large sample even small differences in model fit will be statistically significant. Consequently, many researchers have proposed a number of heuristic statistics, called goodness-of-fit indices, to assess overall model fit. We discussed a number of these indices. All the indices suggested that model fit was inadequate. We also discussed an analysis of the **RES** matrix to identify reasons for lack of fit. This information, coupled with other information provided in the output, can be used to respecify the model. Model respecification is discussed in Section 6.4.5.

6.4.4 Evaluating the Parameter Estimates and the Estimated Factor Model

If the overall model fit is adequate, then the next step is to evaluate and interpret the estimated model parameters, and if the model fit is not adequate then one should attempt to determine why the model does not fit the data. In order to discuss the interpretation and evaluation of the estimated model parameters we will for the time being assume that the model fit is adequate. This is followed by a discussion of the additional diagnostic procedures available to assess reasons for lack of model fit.

Parameter Estimates

From the maximum likelihood parameter estimates the estimated factor model can be represented by the following equations [5a]:

$$\begin{aligned} M &= 1.000IQ + \delta_1; & P &= 1.134IQ + \delta_2; & C &= 1.073IQ + \delta_3; \\ E &= 1.786IQ + \delta_4; & H &= 1.770IQ + \delta_5; & F &= 1.937IQ + \delta_6, \end{aligned} \tag{6.16}$$

[10]This is similar to the analysis of residuals in multiple regression analysis for identifying possible reasons for lack of model fit.

and the variance of the latent construct is 0.836 [5b]. Note that the output gives estimates for the variances of the error terms (i.e., δ). For example, $V(\delta_1) = 3.164$ [5c].

The output also reports the standardized values of the parameter estimates [9]. Standardization is done with respect to the latent constructs and not the indicators. That is, parameter estimates are standardized such that the variances of the latent constructs are one. Consequently, for a covariance matrix input it is quite possible to have indicator loadings that are greater than one. The completely standardized solution, on the other hand, standardizes the solution such that the variances of the latent constructs *and* the indicators are one. The completely standardized solution is used to determine if there are inadmissible estimates. Inadmissible estimates result in an improper factor solution. Inadmissible estimates are: (1) factor loadings that do not lie between -1 and $+1$; (2) negative variances of the constructs and the error terms; and (3) variances of the error terms that are greater than one. It can be seen that all the factor loadings are between -1 and $+1$ [10a] and variances of the construct and the error terms are positive and less than or equal to one [10b, 10c]. Therefore, the estimated factor solution is proper or admissible.

Statistical Significance of the Parameter Estimates

The statistical significance of each estimated parameter is assessed by its t-value. As can be seen, all the parameter estimates are statistically significant at an alpha of .05 [8]. That is, the loadings of all the variables on the IQ factor are significantly greater than zero.

Are the Indicators Good Measures of the Construct?

Given that the parameter estimates are statistically significant, the next question is: To what extent are the variables good or reliable indicators of the construct they purport to measure? The output gives additional statistics for answering this question.

SQUARED MULTIPLE CORRELATIONS. The total variance of any indicator can be decomposed into two parts: the first part is that which is in common with the latent construct and the second part is that which is due to error. For example, for indicator M, out of a total variance of 4.000 for M, 3.164 [5c] is due to error and .836 (i.e., $4 - 3.164$) is in common with IQ construct. That is, the proportion of variance of M that is in common with the IQ construct it is measuring is equal to .209 (.836/4). The proportion of the variance in common with the construct is called the communality of the indicator. As discussed in Chapter 5, the higher the communality of an indicator the better or more reliable measure it is of the respective construct and vice versa. LISREL labels the communality as *squared multiple correlation*. This is because, as shown in Section A6.1 of the Appendix, the communality is the same as the square of the multiple correlation between the indicator and the construct.

The squared multiple correlation for each indicator is given in the output [5d]. It is clear that the squared multiple correlation gives the communality of the indicator as reported in exploratory factor analysis programs. Therefore, the squared multiple correlation can be used to assess how good or reliable an indicator is for measuring the construct that it purports to measure. Although there are no hard and fast rules regarding how high the communality or squared multiple correlation of an indicator should be, a good rule of thumb is that it should be at least greater than 0.5. This rule of thumb is based on the logic that an indicator should have at least 50% of its variance in common with its construct. In the present case, the communalities of the first three indicators

are not high, implying that they are not good indicators of the *IQ* construct. This may be because the indicators are poor measures or because the hypothesized model is not correct. If it is suspected that the hypothesized model is not the correct model, then one can modify or respecify the model. Model respecification is discussed in Section 6.4.5.

TOTAL COEFFICIENT OF DETERMINATION. The squared multiple correlations are used to assess the appropriateness of each indicator. Obviously, one is also interested in assessing the extent to which the indicators as a group measure the construct. For this purpose, LISREL reports the *total coefficient of determination for x variables*, which is computed using the formula

$$1 - \frac{|\Theta_\delta|}{|\mathbf{S}|},$$

where $|\Theta_\delta|$ is the determinant of the covariance matrix of the error variances and $|\mathbf{S}|$ is the determinant of the sample covariance matrix. It is obvious from this formula that the greater the communalities of the indicators, the greater the coefficient of determination and vice versa. For a one-dimensional (i.e., unidimensional) construct this measure is closely related to coefficient alpha and can be used to assess construct reliability. Once again, we are faced with the issue: How high is high? One of the commonly recommended cutoff values is 0.80; however, researchers have used values as low as 0.50. For the present model, a value of 0.895 [5e] suggests that the indicators as a group do tend to measure the *IQ* construct. However, note that *E*, *H*, and *F* are relatively better indicators than are *M*, *P*, and *C*.

6.4.5 Model Respecification

The χ^2 test and the goodness-of-fit indices suggested a poor fit for the one-factor model. The question then becomes: How can the model be modified to fit the data? LISREL provides a number of diagnostic measures that can help in identifying the reasons for poor fit. Using these diagnostic measures and the underlying theory, one can respecify or change the model. This is known as model respecification.

As indicated previously, the **RES** matrix can provide important information regarding model reformulation. An analysis of the residuals indicated that the covariances among the indicators *M*, *P*, and *C* were not being adequately explained by the model. It appears that something other than the *IQ* construct is responsible for the covariances among these three indicators. The *modification indices* provided by LISREL can also be used for model respecification. The modification index of each fixed parameter gives the approximate decrease in the χ^2 value, if that parameter is estimated. It can be seen that the modification indices of the covariances among the errors of *M*, *P*, and *C* are high, indicating that the fit of the model can be improved substantially if they are correlated [11].

The above diagnostic measures hinted that the covariances among the three indicators *M*, *P*, and *C* were not being explained by the model, suggesting further that their error terms (i.e., unique factors) should be correlated. Should this be done? It depends. The error terms should only be correlated if their correlation can be theoretically justified. That is, any model modification or respecification should be well grounded in theory. The data can only provide hints or clues as to what changes should or can be made.

In the present case, one might argue that there are not one, but two latent constructs that account for the correlation among the indicators. The first construct measures

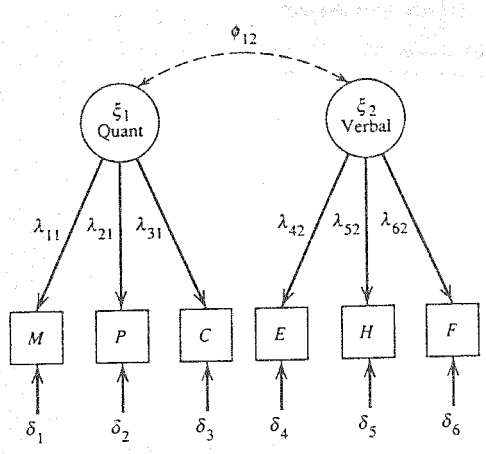


Figure 6.4 Two-factor model.

students' quantitative ability with M, P, and C as its indicators, and the second construct measures students' verbal ability with E, H, and F as its indicators. It can be further hypothesized that the two constructs are independent, i.e., they are not correlated. Figure 6.4 gives the two-factor model, which can be represented by the following equations (for the time being ignore the dotted arrow between the two constructs):

$$M = \lambda_{11}\xi_1 + \delta_1; \qquad P = \lambda_{21}\xi_1 + \delta_2; \qquad C = \lambda_{31}\xi_1 + \delta_3$$
$$E = \lambda_{42}\xi_2 + \delta_4; \qquad H = \lambda_{52}\xi_2 + \delta_5; \qquad F = \lambda_{62}\xi_2 + \delta_6.$$

Table 6.6 gives the LISREL commands[11] and Exhibit 6.2 gives the partial LISREL output. The following conclusions can be drawn from the output:

1. Statistically, the null hypothesis is rejected, indicating that the overall fit of the two-factor model is not good [2]. However, the *RGFI* of 0.937 (0.923/0.985), *RAGFI* of 0.850 (0.821/0.965), and *RNI* of 0.930 are above the recommended cutoff values, implying a good fit [2]. But values of 0.854 for *TLI* and 0.887 for *MDN* indicate a less than desirable fit. Furthermore, the *RMSR* value of 0.948 is higher than the one-factor model, indicating that some of the residuals are large and the fit might be worse than the one-factor model [2].
2. All but one of the indicators are good measures of their respective constructs as the squared multiple correlations are above 0.50 [1a].
3. The total coefficient of determination of 0.978 for the x variables indicates that the amount of variance that is in common between the two constructs and the indicators is quite high [1b]. Notice that the total coefficient of determination assesses the appropriateness of *all* the indicators as measures of *all* the constructs in the factor model. Werts, Linn, and Jöreskog (1974) recommend the use of the following formula to assess the reliability of the indicators of a given construct:

$$\frac{\left(\sum_i^p \lambda_{ij}\right)^2}{\left(\sum_i^p \lambda_{ij}\right)^2 + \sum_i^p V(\delta_i)}, \tag{6.17}$$

[11] Only the LISREL commands are given. SPSS commands are the same as that given in Table 6.2.

Table 6.6 **LISREL Commands for the Two-Factor Model**

```
LISREL
 /TITLE "TWO FACTOR ORTHOGONAL MODEL"
 /DATA NI=6 NO=200 MA=CM
 /LABELS
 /'M' 'P' 'C' 'E' 'H' 'F'
 /MODEL NX=6 NK=2 TD=SY
 /LK
 /'QUANT' 'VERBAL'
 /PA LX
 /0 0
 /1 0
 /1 0
 /0 0
 /0 1
 /0 1
 /PA PHI
 /1
 /0 1
 /PA TD
 /1
 /0 1
 /0 0 1
 /0 0 0 1
 /0 0 0 0 1
 /0 0 0 0 0 1
 /VALUE 1.0 LX(1,1) LX(4,2)
 /OUTPUT TV RS MI SC TO
FINISH
```

where λ_{ij} is the loading of the ith variable on the jth construct, $V(\delta_i)$ is the error variance for the ith variable, and p is the number of indicators of the jth construct. Only completely standardized parameter estimates should be used in the above formula. Using Eq. 6.17, the construct reliability for the *QUANT* construct is equal to [5]:

$$\frac{(.810 + .765 + .666)^2}{(.810 + .765 + .666)^2 + (.344 + .414 + .556)} = .793,$$

and for the *VERBAL* construct it is equal to [5]:

$$\frac{(.825 + .831 + .884)^2}{(.825 + .831 + .884)^2 + (.319 + .309 + .218)} = .884.$$

The reliability of both the constructs is reasonably high suggesting that the indicators of the *QUANT* and the *VERBAL* constructs are reliable indicators of their respective constructs.

4. Examination of the residuals reveals that the covariances among the variables of a construct are being perfectly explained by the respective constructs [3a, 3b]. However, the covariances between the variables M, P, and C of the *QUANT* construct and the variables E, H, and F of the *VERBAL* construct are quite high. That is, the residuals of the variables across the constructs are high. This suggests that perhaps the two constructs should be correlated. This assertion is given support by the

Exhibit 6.2

LISREL output (partial) for the two-factor model

```
(1a)0        SQUARED MULTIPLE CORRELATIONS FOR X - VARIABLES
 0                M        P        C        E        H        F
 +           --------  -------- -------- -------- -------- --------
              0.656    0.586    0.444    0.681    0.691    0.782
(1b)0        TOTAL COEFFICIENT OF DETERMINATION FOR X - VARIABLES IS  0.978

(2)0          CHI-SQUARE WITH   9 DEGREES OF FREEDOM =    57.03 (P = .000)
 0                     GOODNESS OF FIT INDEX =0.923
                 ADJUSTED GOODNESS OF FIT INDEX =0.821
                      ROOT MEAN SQUARE RESIDUAL =     0.948

(3a)0        FITTED RESIDUALS
 0                M        P        C        E        H        F
 +           --------  -------- -------- -------- -------- --------
        M     0.000
        P     0.000    0.000
        C     0.000    0.000    0.000
        E     1.280    1.520    1.440    0.000
        H     1.136    1.404    1.344    0.000    0.000
        F     1.480    1.720    1.620    0.000    0.000    0.000

(3b)0        STANDARDIZED RESIDUALS
 0                M        P        C        E        H        F
 +           --------  -------- -------- -------- -------- --------
        M     0.000
        P     0.000    0.000
        C     0.000    0.000    0.000
        E     4.514    5.361    5.078    0.000
        H     4.006    4.951    4.740    0.000    0.000
        F     5.219    6.066    5.713    0.000    0.000    0.000

(4)0         MODIFICATION INDICES FOR PHI
 0              QUANT    VERBAL
 +           --------  --------
    QUANT     0.000
   VERBAL    43.989     0.000

 0       MAXIMUM MODIFICATION INDEX IS   43.99 FOR ELEMENT ( 2, 1) OF PHI

(5) COMPLETELY STANDARDIZED SOLUTION

          LAMBDA X

                 QUANT    VERBAL
              --------  --------
        M        .810      .000
        P        .765      .000
        C        .666      .000
        E        .000      .825
        H        .000      .831
        F        .000      .884
```

(continued)

Exhibit 6.2 (continued)

```
PHI

              QUANT      VERBAL
            --------   --------
QUANT        1.000
VERBAL        .000      1.000

THETA DELTA

                M          P          C          E          H          F
            --------   --------   --------   --------   --------   --------
M             .344
P             .000       .414
C             .000       .000       .556
E             .000       .000       .000       .319
H             .000       .000       .000       .000       .309
F             .000       .000       .000       .000       .000       .218
```

high modification index of 43.989 for the fixed parameter representing the covariance between the two constructs [4]. If we argue that the verbal and quantitative abilities are not independent, but are somewhat related, then it makes sense to correlate them. That is, an oblique or correlated two-factor model may provide a better representation of the phenomenon being studied. Consequently, a two-factor correlated model given in Figure 6.4 (the dotted arrow in the figure depicts that the two constructs are related) is hypothesized.

A correlated factor model is specified by freeing up the parameter representing the correlation between the two constructs. Exhibit 6.3 gives partial LISREL output when

Exhibit 6.3

Two-factor model with correlated constructs

```
(1a)0        SQUARED MULTIPLE CORRELATIONS FOR X - VARIABLES
  0                M          P          C          E          H          F
  +           --------   --------   --------   --------   --------   --------
                 0.602      0.616      0.467      0.677      0.676      0.800
(1b)0        TOTAL COEFFICIENT OF DETERMINATION FOR X - VARIABLES IS  0.972

(2)0          CHI-SQUARE WITH    8 DEGREES OF FREEDOM =      6.05 (P = .642)
  0                     GOODNESS OF FIT INDEX =0.990
                   ADJUSTED GOODNESS OF FIT INDEX =0.974
                        ROOT MEAN SQUARE RESIDUAL =      0.111

(3a)0        FITTED RESIDUALS
  0                M          P          C          E          H          F
  +           --------   --------   --------   --------   --------   --------
          M      0.000
          P      0.045      0.000
          C      0.039     -0.106      0.000
          E     -0.169      0.054      0.163      0.000
          H     -0.312     -0.061      0.068      0.039      0.000
          F     -0.095      0.127      0.232     -0.022      0.000      0.000
```

(continued)

Exhibit 6.3 (continued)

3b STANDARDIZED RESIDUALS

	M	P	C	E	H	F
M	0.000					
P	1.338	0.000				
C	0.657	-1.886	0.000			
E	-1.135	0.371	0.935	0.000		
H	-2.095	-0.416	0.388	1.167	0.000	
F	-0.776	1.064	1.496	-1.275	-0.017	0.000

4 -T-VALUES

LAMBDA X

	QUANT	VERBAL
M	0.000	0.000
P	9.321	0.000
C	8.610	0.000
E	0.000	0.000
H	0.000	12.993
F	0.000	13.939

PHI

	QUANT	VERBAL
QUANT	5.743	
VERBAL	5.522	6.778

THETA DELTA

	M	P	C	E	H	F
M	6.213					
P	0.000	6.000				
C	0.000	0.000	7.865			
E	0.000	0.000	0.000	7.204		
H	0.000	0.000	0.000	0.000	7.216	
F	0.000	0.000	0.000	0.000	0.000	4.887

5 -COMPLETELY STANDARDIZED SOLUTION

LAMBDA X

	QUANT	VERBAL
M	0.776	0.000
P	0.785	0.000
C	0.684	0.000
E	0.000	0.823
H	0.000	0.822
F	0.000	0.894

PHI

	QUANT	VERBAL
QUANT	1.000	
VERBAL	0.568	1.000

THETA DELTA

	M	P	C	E	H	F
M	0.398					
P	0.000	0.384				
C	0.000	0.000	0.533			
E	0.000	0.000	0.000	0.323		
H	0.000	0.000	0.000	0.000	0.324	
F	0.000	0.000	0.000	0.000	0.000	0.200

Table 6.7 **Computations for *NCP*, *MDN*, *TLI*, and *RNI* for the Correlated Two-Factor Model**

1. From Table 6.5, $NCP_n = 2.748$.
2. From Exhibit 6.3 [2]

$$NCP_h = \frac{6.05 - 8}{200} = -010 = 0.000^a; \qquad MDN = e^0 = 1.000.$$

$$TLI = \frac{2.748/15 - 0.000}{2.748/15} = 1.000; \qquad RNI = \frac{2.748 - 0.000}{2.748} = 1.000.$$

[a]The expected value of NCP when the null hypothesis is true is zero. However, due to sampling errors, it is possible to get negative estimates for NCP. In such cases, the value of NCP is assumed to be zero and, consequently, values for MDN, TLI, and RNI will be one, implying an almost perfect model fit.

the covariance between the two factors is estimated. As can be seen, the χ^2 test indicates that the model fits the data quite well [2]. Table 6.7 reports the computations of the various goodness-of-fit indices. It can be seen that all the fit indices are close to one, implying an extremely good fit. Furthermore, none of the residuals are large [3a, 3b]. The results suggest that a two-factor-correlated model fits the data better than any of the previous models.

The completely standardized solution indicates that the solution is admissible [5], and all the parameter estimates are statistically significant [4]. The communalities of all the variables except C are well above .50 [1a]. The total coefficient of determination value of 0.972 is quite high [1b]. Using Eq. 6.17, the construct reliabilities for the *QUANT* and the *VERBAL* constructs are, respectively, .793 and .884, which are reasonably high. To conclude, the results suggest that the hypothesized two-factor model with correlated constructs fits the data quite well. That is, the theory postulating that students' grades are functions of two correlated constructs—verbal and quantitative ability—has empirical support.

One could argue that since the data were used to modify or respecify and test the model, the analysis is not truly confirmatory. In order to do a true confirmatory analysis the model should be developed using one sample and then tested on an independent sample. Or, one could divide the sample into two subsamples: an analysis sample and a holdout sample. The model could be developed on the analysis sample and validated on the holdout sample.

6.5 MULTIGROUP ANALYSIS

In many situations researchers are interested in determining if a hypothesized factor model is the same or different across multiple groups. For example, one might be interested in determining if the loadings, error variances, and the covariance between the two constructs of the correlated two-factor model hypothesized earlier are the same or different for males and females. Or, one might be interested in determining if that hypothesized factor model is the same for two different time periods, say 1960 and 1994. Such hypothesis testing can easily be performed using LISREL by employing multigroup analysis. Following is a discussion of multigroup analysis.

Assume that we are interested in determining if the loadings, error variances, and the covariance between the two constructs of the correlated two-factor model are the

same or different for males and females. The null and alternative hypotheses for this problem are:

$$
\begin{aligned}
H_o: \quad & \Lambda_x^{males} = \Lambda_x^{females} \\
& \Theta_\delta^{males} = \Theta_\delta^{females} \\
& \Phi^{males} = \Phi^{females} \\
H_a: \quad & \Lambda_x^{males} \neq \Lambda_x^{females} \\
& \Theta_\delta^{males} \neq \Theta_\delta^{females} \\
& \Phi^{males} \neq \Phi^{females}
\end{aligned}
$$

These hypotheses are tested by conducting two separate analyses. In the first, separate models for each sample are estimated. The total χ^2 value is equal to the sum of the χ^2 values for each model, and the total degrees of freedom are equal to the sum of the degrees of freedom of each model. This analysis is referred to as the *unconstrained analysis* as the parameter matrices of the models for the two groups are not constrained to be equal to each other. In the second analysis it is assumed that the loadings, error variances, and covariance between the two factors are the same. That is, the parameter matrices of the two samples are constrained to be equal, and the analysis is equivalent to estimating a factor model using the covariance matrix of the combined sample (i.e., the male and the female sample). This analysis is referred to as the *constrained analysis*. The hypotheses are tested by employing a χ^2 difference test. The difference in the χ^2s of the two analyses follows a χ^2 distribution with the degrees of freedom equal to the difference in the degrees of freedom of the two analyses.

Table 6.8 gives the LISREL commands for unconstrained analysis. The SPLIT option in the MATRIX DATA command is used to indicate that multiple matrices will be read. SPSS will assign to the GENDER variable a value of 1 for the first sample and a value of 2 for the second sample. In the table the first correlation matrix is for males and the second correlation matrix is for the females. Once again, it is assumed that the sample size for each group is 200 and that the standard deviations of all variables are equal to 2.

In the DATA command the NG=2 option specifies that there are two groups. The first set of LISREL commands is for the first sample, the male sample. The LISREL commands for the second sample (the female sample) follow the OUTPUT command of the first sample. The absence of any options following the DATA command for the second group indicates that the options are the same as those in the DATA command of the previous group. In the MODEL command of the second group, the PS option indicates that the pattern matrix and the starting values are the same as those of the previous sample. The constrained model is run by replacing the MODEL command in Table 6.8 with MODEL LX=IN TD=IN PHI=IN MA=CM. The IN option specifies that the estimates of the elements of the corresponding matrix should be constrained to be equal. Table 6.9 gives χ^2 values for the two analyses. The χ^2 value for the unconstrained analysis is equal to 17.36 with 16 *df*, and the χ^2 value for the constrained analysis is equal to 18.45 with 29 *df*.[12] A χ^2 difference value of 1.09 (i.e., 18.45 − 17.36) with 13 *df* is not significant at an alpha of .05 and, therefore, we cannot reject the null hypothesis. Thus, it can be concluded that the factor structures for males and females are the same.

[12]In the unconstrained analysis, the degrees of freedom for the model in each sample is equal to 8 giving a total of 16 degrees of freedom for the two samples. In the constrained analysis, the unduplicated number of elements of the covariance matrix for each sample is 21 giving a total of 42 unduplicated elements for the two samples. The number of parameters estimated for the models in the two samples is 13 giving a total of 29 *df*.

Table 6.8 **SPSS Commands for Multigroup Analysis**

```
TITLE LISREL IN SPSSX
MATRIX DATA VARIABLES=M P C E H F/CONTENTS=CORR
 STD/N=200/SPLIT=GENDER
BEGIN DATA
1.000
 .620     1.000
 .540      .510      1.000
 .320      .380       .360      1.000
 .284      .351       .336       .686     1.000
 .370      .430       .405       .730      .735      1.000
2 2 2 2 2 2
1.000
 .600     1.000
 .590      .500      1.000
 .360      .360       .370      1.000
 .274      .361       .346       .676     1.000
 .360      .435       .415       .720      .725      1.000
2 2 2 2 2 2
END DATA
MCONVERT
LISREL
 /TITLE "MULTIGROUP ANALYSIS -- MALE SAMPLE"
 /DATA NI=6 NG=2 NO=200 MA=CM
 /LABELS
 /'M' 'P' 'C' 'E' 'H' 'F'
 /MODEL NX=6 NK=2 TD=SY PH=SY
 /LK
 /'QUANT' 'VERBAL'
 /PA LX
 /0 0
 /1 0
 /1 0
 /0 0
 /0 1
 /0 1
 /PA PHI
 /1
 /1 1
 /PA TD
 /1
 /0 1
 /0 0 1
 /0 0 0 1
 /0 0 0 0 1
 /0 0 0 0 0 1
 /VALUE 1.0 LX(1,1) LX(4,2)
 /OUTPUT TV RS MI SC TO
 /TITLE "FEMALE SAMPLE"
 /DA
 /MO LX=PS TD=PS PHI=PS MA=CM
 /OUTPUT TV RS MI SC TO
```

Table 6.9 **Results of Multigroup Analysis: Testing Factor Structure for Males and Females**

Fit Statistic

Model	χ^2	*df*
Unconstrained	17.36	16
Constrained	18.45	29

Parameter Estimates

	Loadings		
Parameter	Quant	Verbal	Squared Multiple Correlations
M	.785		.616
P	.763		.581
C	.704		.495
E		.821	.673
H		.819	.671
F		.891	.793

Table 6.9 also gives the estimates of factor loadings and squared multiple correlations for the constrained analysis. As can be seen, all of the factor loadings are high and the squared multiple correlations indicate that all the measures are reliable indicators of their respective constructs. The multigroup analysis is quite powerful in that it can be used to test a variety of hypotheses. For example, one could hypothesize that the factor models for the two samples are equivalent only with respect to the covariance between the two factors. Such a hypothesis could be tested by conducting a constrained and an unconstrained analysis. In the constrained analysis only the covariance between the two factors is constrained to be equal. This is achieved by the following two commands:

```
MODEL LX=PS TD=PS PHI=PS MA=CM
EQU PHI(1,2,1) PHI(2,2,1)
```

The EQU command specifies equality of the covariance between the two factors for the two samples. The first subscript in the EQU command refers to the sample number.

6.6 ASSUMPTIONS

The maximum likelihood estimation procedure assumes that the data come from a multivariate normal distribution. Theoretical and simulation studies have shown that the violation of this assumption biases the χ^2 statistic and the standard errors of the parameter estimates (Sharma, Durvasula, and Dillon 1989); however, the parameter estimates themselves are not affected. Of the two nonnormality characteristics (i.e., kurtosis and skewness) it appears that only nonnormality due to kurtosis affects the χ^2 statistic and the standard errors. If the data do not come from a multivariate normal distribution then one can use alternative estimation methods such as generalized least squares, elliptical estimation techniques, and asymptotic distribution free methods. These estimation methods are available in version 8 of LISREL and in EQS. The simulation study

found that the performance of the elliptical methods was superior to other methods and it is recommended that this method be used when the assumption of nonnormality is violated.

6.7 AN ILLUSTRATIVE EXAMPLE

Shimp and Sharma (1987) developed a 17-item scale to measure consumers' ethnocentric tendencies (CET) related to purchasing foreign-made versus American-made products. This study also identified a shortened 10-item scale (see Table 6.10 for a list). The scale was developed using rigorous procedures and is well grounded in theory. Suppose we are interested in independently verifying the hypothesis that the 10 items given in Table 6.10 are indicators of the CET construct. That is, a 10-indicator one-factor model is hypothesized. To test our hypothesis, data were collected from a sample of 575 subjects, who were asked to indicate their degree of agreement or disagreement with each of the 10 statements using a seven-point Likert-type scale. Exhibit 6.4 gives the partial LISREL output. From the output it can be seen that:

1. The χ^2 statistic indicates that statistically the model does not fit the data [3]. However, keeping in mind the sensitivity of the χ^2 test to sample size, we use the goodness-of-fit indices to assess model fit. From Eq. 6.8 and 6.10 the *EGFI* and *EAGFI* are, respectively, 0.988 and 0.981 [3], giving a value of 0.940 (0.929/0.988) for *RGFI* and a value of 0.906 (.889/.981) for *RAGFI*. Values of 0.867 for *MDN*, 0.949 for *TLI*, and 0.960 for *RNI* suggest a good model fit.[13] The *RMSR* is 0.111 and is quite low. Thus, the goodness-of-fit indices suggest an adequate fit of the model to the data.
2. The factor solution is admissible because all the completely standardized loadings are between −1 and +1, the variances of the error terms are positive and less than one, and the variance of the CET construct is one [5]. The *t*-values indicate that all the estimated loadings and the variances of the error terms are significant at an alpha of .05 [4].

Table 6.10 **Items or Statements for the 10-item CET Scale**

Respondents stated their level of agreement or disagreement with the following statements on a seven-point Likert-type scale.

1. Only those products that are unavailable in the U.S. should be imported.
2. American products, first, last, and foremost.
3. Purchasing foreign-made products is un-American.
4. It is not right to purchase foreign products, because it puts Americans out of jobs.
5. A real American should always buy American-made products.
6. We should purchase products manufactured in America instead of letting other countries get rich off us.
7. Americans should not buy foreign products, because this hurts American business and causes unemployment.
8. It may cost me in the long-run but I prefer to support American products.
9. We should buy from foreign countries only those products that we cannot obtain within our own country.
10. American consumers who purchase products made in other countries are responsible for putting their fellow Americans out of work.

[13]The χ^2 for the null model is 4160.24 with 45 *df*.

Exhibit 6.4

LISREL output for the 10-item CETSCALE

```
(1) 1TITLE TEN ITEM CETSCALE
 0        COVARIANCE MATRIX TO BE ANALYZED
 0                V1        V2        V3        V4        V5        V6
 +           --------  --------  --------  --------  --------  --------
       V1      4.174
       V2      2.769     4.340
       V3      1.845     1.994     2.742
       V4      2.791     2.827     2.257     4.300
       V5      2.386     2.610     1.609     2.737     3.689
       V6      2.645     2.950     2.101     2.901     2.697     4.080
       V7      2.619     2.719     1.795     2.999     2.908     2.802
       V8      2.134     2.535     2.057     2.739     2.334     2.785
       V9      2.522     2.369     1.856     2.776     2.375     2.610
      V10      1.831     2.091     1.132     1.832     1.831     1.920

 0        COVARIANCE MATRIX TO BE ANALYZED
 0                V7        V8        V9       V10
 +           --------  --------  --------  --------
       V7      3.988
       V8      2.582     3.590
       V9      2.774     2.341     3.987
      V10      2.074     1.740     1.736     3.284

(2a) 0        SQUARED MULTIPLE CORRELATIONS FOR X - VARIABLES
 0                V1        V2        V3        V4        V5        V6
 +           --------  --------  --------  --------  --------  --------
              0.578     0.633     0.508     0.715     0.666     0.718
 0        SQUARED MULTIPLE CORRELATIONS FOR X - VARIABLES
 0                V7        V8        V9       V10
 +           --------  --------  --------  --------
              0.726     0.662     0.602     0.396
(2b) 0        TOTAL COEFFICIENT OF DETERMINATION FOR X - VARIABLES IS   0.947

(3) 0         CHI-SQUARE WITH  35 DEGREES OF FREEDOM =    95.45 (P = .000)
 0                    GOODNESS OF FIT INDEX =0.932
              ADJUSTED GOODNESS OF FIT INDEX =0.893
                   ROOT MEAN SQUARE RESIDUAL =     0.111

(4) -T-VALUES
 0        LAMBDA X
 0             KSI 1
 +           --------
       V1      0.000
       V2     14.009
       V3     12.317
       V4     15.073
       V5     14.441
       V6     15.109
       V7     15.214
       V8     14.381
       V9     13.598
      V10     10.693
```

(continued)

Exhibit 6.4 (continued)

```
0        PHI
0              KSI 1
+           --------
    KSI 1      7.328
0        THETA DELTA
0                V1        V2        V3        V4        V5        V6
+           --------  --------  --------  --------  --------  --------
              10.807    10.572    11.029    10.053    10.394    10.030
0        THETA DELTA
0                V7        V8        V9       V10
+           --------  --------  --------  --------
               9.959    10.421    10.713    11.277
5 -COMPLETELY STANDARDIZED SOLUTION
0        LAMBDA X
0              KSI 1
+           --------
        V1      0.760
        V2      0.796
        V3      0.713
        V4      0.846
        V5      0.816
        V6      0.847
        V7      0.852
        V8      0.813
        V9      0.776
       V10      0.629
0        PHI
0              KSI 1
+           --------
    KSI 1      1.000
0        THETA DELTA
0                V1        V2        V3        V4        V5        V6
+           --------  --------  --------  --------  --------  --------
               0.422     0.367     0.492     0.285     0.334     0.282
0        THETA DELTA
0                V7        V8        V9       V10
+           --------  --------  --------  --------
               0.274     0.338     0.398     0.604
```

3. The squared multiple correlation for all statements except statement 10 is greater than the recommended value of 0.50 [2a].
4. The total coefficient of determination of 0.947 for the total scale [2b], and the construct reliability of .942, computed using Eq. 6.17, are quite high, indicating that the 10 items combined are good indicators of the CET construct.

The preceding analysis suggests that the fit of the data to the hypothesized factor model is adequate. That is, we conclude that the CET construct is unidimensional, and the 10 items given in Table 6.10 are indeed good indicators of this construct.

6.8 SUMMARY

In this chapter we discussed the basic concepts of confirmatory factor models. Confirmatory factor analysis is different from exploratory factor analysis discussed in the previous chapter. In exploratory factor analysis, the researcher has no knowledge of the factor structure and is essentially seeking to identify the factor model that would account for the covariances among the variables. In confirmatory factor models, on the other hand, the precise structure of the model

is known and the major objective is to empirically validate the hypothesized model and estimate model parameters.

Confirmatory factor analysis can be done using a number of computer programs. These programs are available in various statistical packages such as CALIS in SAS, EQS in BMDP, and LISREL in SPSS, or as stand-alone PC programs. In this chapter we discussed the use of LISREL as it is one of the most widely used programs.

The next chapter discusses cluster analysis, a technique useful for forming groups or clusters such that the observations within each cluster are similar with respect to the clustering variables and the observations across clusters are dissimilar with respect to the clustering variables.

QUESTIONS

6.1 The common factor analytic model is given by:

$$\mathbf{x} = \Lambda_x \boldsymbol{\xi} + \boldsymbol{\delta}.$$

Although both exploratory factor analysis and confirmatory factor analysis attempt to estimate the unknown parameters of the above model, there is a fundamental difference between the two approaches. Discuss.

6.2 Explain what is meant by *identification* in the context of estimating the unknown parameters of a common factor model.

Classify the following models as under-, just-, or overidentified:

(a)

$$\begin{aligned} x_1 &= \lambda_1\xi_1 + \delta_1 \\ x_2 &= \lambda_2\xi_1 + \delta_2. \end{aligned}$$

(b)

$$\begin{aligned} x_1 &= \lambda_1\xi_1 + \delta_1 \\ x_2 &= \lambda_2\xi_1 + \delta_2 \\ x_3 &= \lambda_3\xi_1 + \delta_3 \\ x_4 &= \lambda_4\xi_1 + \delta_4 \\ x_5 &= \lambda_5\xi_1 + \delta_5. \end{aligned}$$

(c)

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \delta_1 \\ x_2 &= \lambda_{21}\xi_1 + \delta_2 \\ x_3 &= \lambda_{32}\xi_2 + \delta_3 \\ x_4 &= \lambda_{42}\xi_2 + \delta_4 \end{aligned}$$

when ξ_1 and ξ_2 are uncorrelated.

(d)

$$\begin{aligned} x_1 &= \lambda_{11}\xi_1 + \delta_1 \\ x_2 &= \lambda_{21}\xi_1 + \delta_2 \\ x_3 &= \lambda_{32}\xi_2 + \delta_3 \\ x_4 &= \lambda_{42}\xi_2 + \delta_4 \end{aligned}$$

when ξ_1 and ξ_2 are correlated.

What are the degrees of freedom associated with each of the above models?

What is paradoxical about the models in (c) and (d)? How can this paradox be explained?

6.3 Consider the following single-factor model:

$$\begin{aligned} x_1 &= \lambda_1\xi + \delta_1 \\ x_2 &= \lambda_2\xi + \delta_2 \\ x_3 &= \lambda_3\xi + \delta_3. \end{aligned}$$

If the sample covariance matrix of the indicators is given by:

$$S = \begin{pmatrix} 1.20 & 0.93 & 0.45 \\ 0.93 & 1.56 & 0.27 \\ 0.45 & 0.27 & 2.15 \end{pmatrix}$$

compute the estimates of the model parameters (λ_1, λ_2, λ_3, Var(δ_1), Var(δ_2), Var(δ_3)) *using hand calculations*. Are the parameter estimates unique?

Recompute the parameter estimates with the restriction

$$Var(\delta_1) = Var(\delta_2) = Var(\delta_3).$$

Are the new parameter estimates unique?

Use the new parameter estimates to obtain the estimated covariance matrix ($\hat{\Sigma}$). Compare $\hat{\Sigma}$ to S. Would you consider your model to provide a good fit to the data? Why?

6.4 Given the model shown in Figure Q6.1,

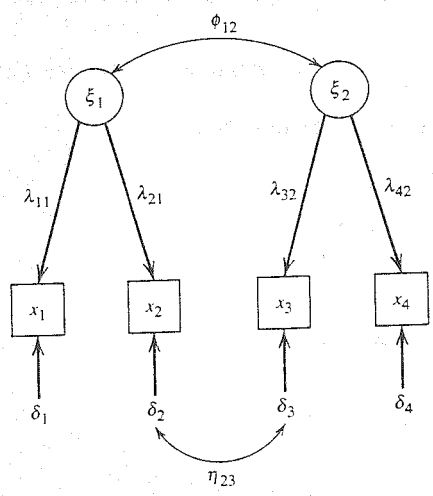


Figure Q6.1 Model.

(a) Represent the covariance matrix between the indicators as a function of the model parameters.
(b) Is the model under-, just-, or overidentified? Explain.
(c) What restriction(s) can you impose on the parameters to overidentify the model? Justify.

6.5 Table Q6.1 presents a hypothetical correlation matrix.

Table Q6.1 **Hypothetical Correlation Matrix**

Variable	1	2	3	4	5	6
1	1.00	.90	.90	.70	.70	.70
2		1.00	.90	.70	.70	.70
3			1.00	.70	.70	.70
4				1.00	.90	.90
5					1.00	.90
6						1.00

Use the above correlation matrix to estimate each of the models shown in Figures Q6.2 (a)–(d) (assume a sample size of 200). Which model would you consider to be the most acceptable? Why?

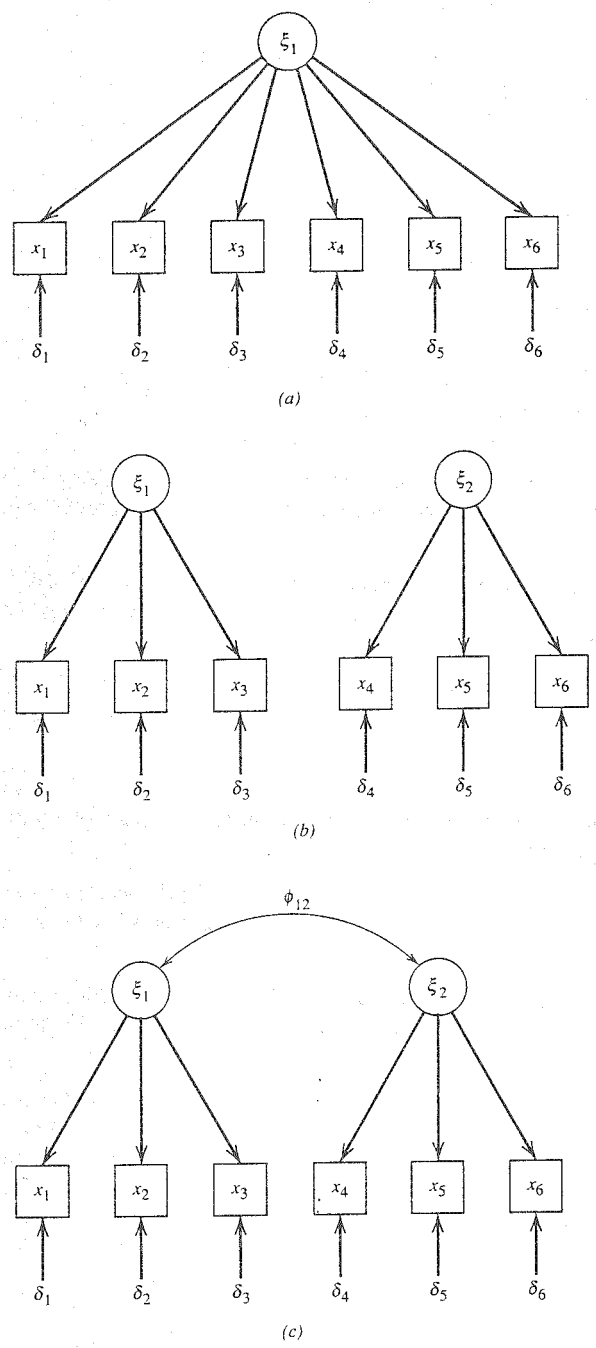
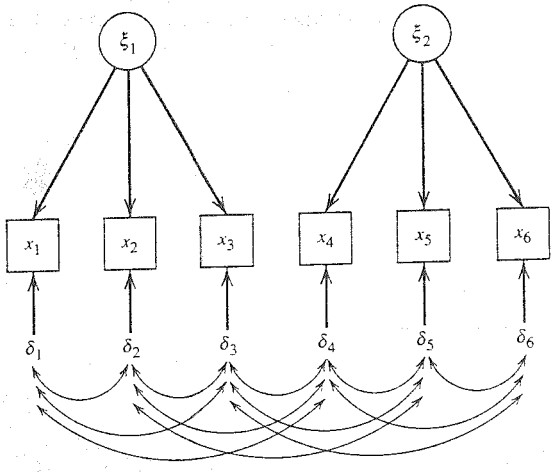


Figure Q6.2 Models. *Notes:* The loadings for the models shown in (*a*) to (*d*) have been left out to prevent cluttering. The assumption for the model shown in (*d*) is that all the error terms are correlated. It is also assumed that the covariances between the error terms are all equal. *(continues)*



(d)

Figure Q6.2 (continued)

6.6 Perform confirmatory factor analysis on the correlation data given in file PHYSATT.DAT and interpret the results. How do the results compare with the factor structure obtained using exploratory factor analysis?

6.7 Perform confirmatory factor analysis on the correlation data given in file TEST.DAT and interpret the results. How do the results compare with the factor structure obtained using exploratory factor analysis?

6.8 Perform confirmatory factor analysis on the correlation data given in BANK.DAT and interpret the results. How do the results compare with the factor structure obtained using exploratory factor analysis?

6.9 Perform a confirmatory factor analysis to determine the underlying perceptions about the energy crisis, using variables V_{19} to V_{38} of the mass transportation data given in file MASST.DAT. Interpret the results.

6.10 Perform confirmatory factor analysis on the data given in file NUT.DAT and interpret the results. How do the results compare with the factor structure obtained using exploratory factor analysis?

6.11 Perform confirmatory factor analysis on the data given in SOFTD.DAT and interpret the results. How do the results compare with the factor structure obtained using exploratory factor analysis?

6.12 Suppose a researcher has developed a seven-item unidimensional scale to measure consumer ethnocentric tendencies. The seven-item scale was administered to a random sample of 300 respondents in Korea and in the U.S. File CET.DAT gives the covariance matrices among the seven items for the two samples. Conduct a group analysis to test for equivalence of the factor structure for the two samples. What conclusions can you draw from your analysis?

Appendix

In this appendix we discuss the computational procedures for squared multiple correlations and the basic concepts of maximum likelihood estimation technique.

A6.1 SQUARED MULTIPLE CORRELATIONS

From Exhibit 6.1, the estimated factor model can be represented by the following equations:

$$M = 1.000IQ + \delta_1; \quad P = 1.134IQ + \delta_2; \quad C = 1.073IQ + \delta_3$$
$$E = 1.786IQ + \delta_4; \quad H = 1.770IQ + \delta_5; \quad F = 1.937IQ + \delta_6. \qquad \text{(A6.1)}$$

The variance of any indicator, say P, is computed as (see Eq. A5.2 in the Appendix to Chapter 5):

$$\begin{aligned} V(M) &= E(1.000IQ + \delta_1)^2 \\ &= 1.000^2E(IQ^2) + E(\delta_1^2) \\ &= 1.000^2V(IQ) + V(\delta_1) \\ &= 1.000 \times 0.836 + 3.164 \\ &= 0.836 + 3.164 \\ &= 4.000. \end{aligned}$$

That is, out of a total variance of 4.000 for P, 0.836 or 20.9% (0.836/4) is in common with the IQ construct that it is measuring, and 3.164 or 79.1% is due to error. The proportion of the variance in common with the construct is called the communality of the indicator. For indicator P this is equal to .209. As discussed in Chapter 5, the higher the communality of an indicator the better the measure it is of the respective construct and vice versa. LISREL labels the communality as *squared multiple correlation*. This is because, as shown below, communality is the same as the square of the multiple correlation between the indicator and the construct.

The covariance between any indicator, say P, and the construct IQ is given by (see Eq. A5.3 in the Appendix to Chapter 5)

$$\begin{aligned} Cov(M, IQ) &= E[(1.000IQ + \delta_1)IQ] \\ &= 1.000E(IQ^2) \\ &= 1.000(0.836) \\ &= .836, \end{aligned}$$

and the correlation between P and IQ is

$$\begin{aligned} r(M, IQ) &= \frac{.836}{\sqrt{4.000}\sqrt{.836}} \\ &= .457. \end{aligned}$$

The square of the correlation is .209 which, within rounding error, is the same as the communality.

A6.2 MAXIMUM LIKELIHOOD ESTIMATION

The basic concepts of maximum likelihood estimation technique are discussed by using two simple examples. In the first example, consider the case where a coin is tossed and the probability of obtaining a head, H, is p and the probability of obtaining a tail, T, is $1 - p$. Suppose that the coin is tossed four times with the following outcomes: H, H, H, and T. If the outcome at each trial is independent of the previous outcomes and the probability p does not change, then the joint probability of obtaining three heads and a tail is given by

$$\begin{aligned} l = P(H,H,H,T) &= p \times p \times p \times (1 - p) \\ &= p^3(1 - p). \end{aligned} \qquad \text{(A6.2)}$$

In this equation, p is the parameter of the process that is generating the data or the outcomes. Now the question is: What is the value of the parameter p that maximizes the joint probability

Table A6.1 **Value of the Likelihood Function for Various Values of p**

p	Likelihood Function (l)
0.00	0.0000
0.10	0.0009
0.20	0.0064
0.30	0.0189
0.40	0.0384
0.50	0.0625
0.60	0.0864
0.75	0.1055
0.80	0.1024
0.90	0.0729
1.00	0.0000

$P(H,H,H,T)$, i.e., the probability of obtaining three heads and a tail? The outcomes H, H, H, and T are observed data and $P(H,H,H,T)$ is referred to as the likelihood l of observing the data for a given value of p. The maximum likelihood estimate of parameter p is defined as that estimate of the parameter that results in the maximum likelihood or probability of observing the given sample data; i.e., it is the value of p for which the sample data will occur the most often.

Equation A6.2 is known as the likelihood function. The value of p can be obtained by trial and error or by using calculus, if the function is analytically tractable. The trial-and-error procedure tries different values of the parameters and selects the one that results in the highest value for l. For example, Table A6.1 gives the value of l for various estimates of p. Figure A6.1 gives a graphical representation of the results in Table A6.1. It can be seen that the maximum value of the likelihood function occurs for $p = .75$.

Since the likelihood function given by Eq. A6.2 is analytically tractable, the estimate of p can also be obtained by differentiating the likelihood function with respect to the parameter p and equating it to zero. That is,

$$\frac{dl}{dp} = 3p^2 - 4p^3 = 0$$

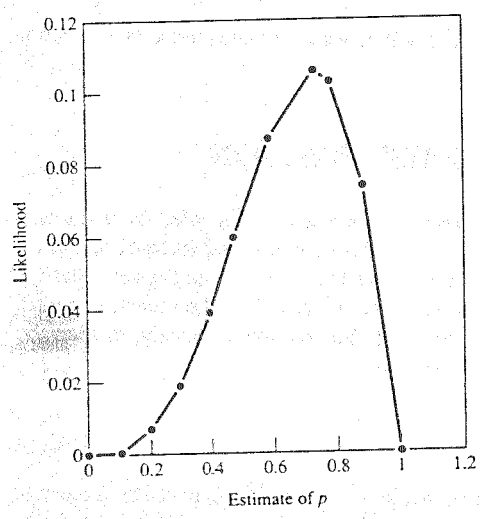


Figure A6.1 Maximum likelihood estimation procedure.

or

$$p^2(3 - 4p) = 0$$
$$p = 3/4 = .75.$$

Many times instead of maximizing l, the natural log (ln) of the maximum likelihood function is maximized (i.e., $L = \ln l$). Maximizing L does not affect the results as the ln of a variable is a monotonic function of the variable.

For the second example, consider the case of normally distributed random variable X with a mean of μ and variance σ^2. Assume that the variance of the distribution is known to be 1.0, and the following four values of x are observed (i.e., data): 3, 4, 6, and 7. What then is the maximum likelihood estimate of the mean μ? We know that the density function for a normal distribution is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-0.5(\frac{x-\mu}{\sigma})^2}$$

or

$$\ln f(x) = \ln \frac{1}{\sqrt{2\pi\sigma^2}} - 0.5\left(\frac{x-\mu}{\sigma}\right)^2.$$

Since the first term of this equation is a constant for a given value of σ, the equation can be rewritten as

$$\ln f(x) = -0.5\left(\frac{x-\mu}{\sigma}\right)^2. \tag{A6.3}$$

The likelihood function l for the data will be $f(x = 3)f(x = 4)f(x = 6)f(x = 7)$ or the ln of the likelihood function will be (note that σ is assumed to be equal to 1):

$$L = \ln[f(x = 3)f(x = 4)f(x = 6)f(x = 7)]$$
$$= \ln f(x = 3) + \ln f(x = 4) + \ln f(x = 6) + \ln f(x = 7).$$

Substituting the value of $f(x)$ from Eq. A6.3

$$L = -0.5(3 - \mu)^2 - 0.5(4 - \mu)^2 - 0.5(6 - \mu)^2 - 0.5(7 - \mu)^2. \tag{A6.4}$$

Table A6.2 and Figure A6.2 give the value of the preceding likelihood function for various estimates of μ. As can be seen, the value of 5 gives the maximum value for l and hence the maximum likelihood estimate of μ is 5.0, i.e., $\hat{\mu} = 5$.

Table A6.2 **Maximum Likelihood Estimate for the Mean of a Normal Distribution**

$\hat{\mu}$	Log of the Likelihood Function (l)
3.0	−13.0
3.5	−9.5
4.0	−7.0
4.5	−5.5
5.0	−5.0
5.5	−5.5
6.0	−7.0
6.5	−9.5
7.0	−13.0
7.5	−23.0

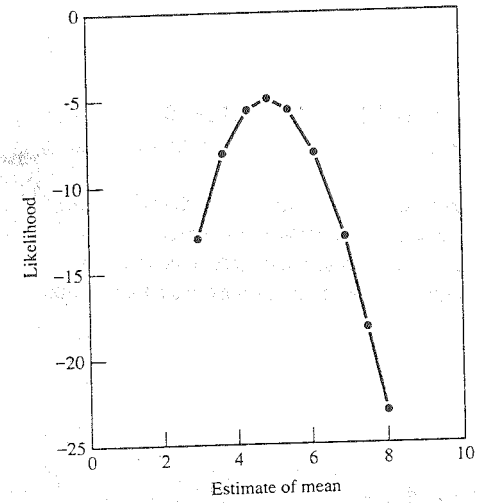


Figure A6.2 Maximum likelihood estimation for mean of a normal distribution.

Once again, since the likelihood function given by Eq. A6.4 is analytically tractable, the estimate for μ can also be obtained using calculus. Differentiating Eq. A6.4 with respect to μ and equating to zero gives

$$\frac{dL}{d\mu} = (3-\mu) + (4-\mu) + (6-\mu) + (7-\mu) = 0$$

$$3+4+6+7-4\mu = 0$$

$$\mu = \frac{3+4+6+7}{4}$$

$$= 5.$$

In general it can be easily shown that the formula for the maximum likelihood estimate of the mean is

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

It is clear from the discussion that the likelihood function must be known if maximum likelihood estimates of the parameters are desired. And, in order to obtain the likelihood function it is necessary to know the distribution from which data are generated.

The maximum likelihood estimation procedure in LISREL uses the likelihood function of the hypothesized model under the assumption that the data come from a multivariate normal distribution. However, in most cases, the resulting likelihood function is analytically intractable, and hence iterative procedures are required to identify the parameter values that would maximize the function. For further discussion regarding the derivation of the likelihood function used by LISREL and the iterative procedures used, see Bollen (1989) and Hayduk (1987).

CHAPTER 7

Cluster Analysis

Consider the following scenarios:

- The financial analyst of an investment banking firm is interested in identifying a group of firms that are prime targets for takeover.
- A marketing manager is interested in identifying similar cities that can be used for test marketing.
- The campaign manager for a political candidate is interested in identifying groups of voters who have similar views on important issues.

Each of the above scenarios is concerned with identifying groups of entities or subjects that are similar to each other with respect to certain characteristics. Cluster analysis is a useful technique for such a purpose. In this chapter we discuss the concept of cluster analysis and some of the available techniques for forming homogeneous groups or clusters.

7.1 WHAT IS CLUSTER ANALYSIS?

Cluster analysis is a technique used for combining observations into groups or clusters such that:

1. Each group or cluster is homogeneous or compact with respect to certain characteristics. That is, observations in each group are similar to each other.
2. Each group should be different from other groups with respect to the same characteristics; that is, observations of one group should be different from the observations of other groups.

The definition of similarity or homogeneity varies from analysis to analysis, and depends on the objectives of the study.

Consider a deck of playing cards. The 52 cards can be grouped using a number of different schemes. One scheme can have all red cards in one group and all black cards in another group. Or, a blackjack player might wish to group the cards into a group containing all face cards and another group containing the rest of the cards. Similarly, in the Hearts game a more meaningful grouping might be: (1) all hearts cards; (2) queen of spades; and (3) the rest of the cards. It is obvious that one can have a number of different grouping schemes, each dependent upon the purpose or objectives of the game.

Table 7.1 **Hypothetical Data**

Subject Id	Income ($ thous.)	Education (years)
S1	5	5
S2	6	6
S3	15	14
S4	16	15
S5	25	20
S6	30	19

7.2 GEOMETRICAL VIEW OF CLUSTER ANALYSIS

Geometrically, the concept of cluster analysis is very simple. Consider the hypothetical data given in Table 7.1. The table contains income and education in years for six hypothetical subjects.[1] As shown in Figure 7.1, each observation can be represented as a point in a two-dimensional space. In general, each observation can be represented as a point in a p-dimensional space, where p is the number of variables or characteristics used to describe the subjects. Now suppose we want to form three homogeneous groups. An examination of the figure suggests that subjects S1 and S2 will form one group, subjects S3 and S4 will form another group, and subjects S5 and S6 will form the third group.

As can be seen, cluster analysis groups observations such that the observations in each group are similar with respect to the clustering variables. It is also possible to cluster variables such that the variables in each group are similar with respect to the clustering observations. Geometrically, this is equivalent to representing data in an n-dimensional observation space, and identifying clusters of variables. This objective of cluster analysis appears to be similar to that of factor analysis.

Recall that in factor analysis we attempt to identify clusters of variables such that the variables in each cluster have something in common; i.e., they appear to measure the same latent factor. It is therefore possible to use factor analysis to cluster observations, and to use cluster analysis to cluster variables. The factor analysis technique used to

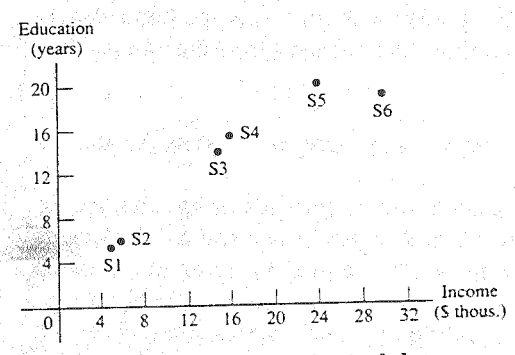


Figure 7.1 Plot of hypothetical data.

[1]The terms, *subjects* and *observations*, are used interchangeably.

cluster observations is known as *Q-factor analysis*. However, we do not recommend the use of Q-factor analysis for clustering observations as it introduces additional problems.[2] We subscribe to the philosophy that: (1) if one is interested in identifying latent factors and their indicators then one should use factor analysis as it is a technique specifically developed for this purpose; and (2) if one is interested in clustering observations then one should use cluster analysis as it is a technique specifically developed for this purpose.

The graphical procedures for identifying clusters may not be feasible when we have many observations or when we have more than three variables or characteristics. What is needed, in such a case, is an analytical technique for identifying groups or clusters of points in a given dimensional space.

7.3 OBJECTIVE OF CLUSTER ANALYSIS

The objective of cluster analysis is to group observations into clusters such that each cluster is as homogeneous as possible with respect to the clustering variables. The first step in cluster analysis is to select a measure of similarity. Next, a decision is made on the type of clustering technique to be used (e.g., a hierarchical or a nonhierarchical). Third, the type of clustering method for the selected technique is selected (e.g., centroid method in hierarchical clustering technique). Fourth, a decision regarding the number of clusters is made. Finally, the cluster solution is interpreted.

7.4 SIMILARITY MEASURES

In the geometrical approach to clustering, we visually combined subjects S1 and S2 into one group or cluster as these two subjects appeared to be close to each other in the two-dimensional space. In other words, we implicitly used the distance between the two points (i.e., the subjects) as a measure of similarity. A number of different similarity measures can be used. Therefore, one of the issues facing the researcher is the selection of an appropriate measure of similarity, an issue covered in Section 7.10 after the various clustering techniques and their methods have been discussed. For the time being let us assume that we have selected the squared euclidean distance between two points as a measure of similarity. The squared euclidean distance between subjects S1 and S2 is given by

$$\begin{aligned} D_{12}^2 &= (5-6)^2 + (5-6)^2 \\ &= 2, \end{aligned}$$

where D_{12}^2 is the squared euclidean distance between subjects S1 and S2. The more similar the subjects, the smaller the distance between them and vice versa. The formula for computing squared euclidean distances for p variables is given by

$$D_{ij}^2 = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2, \tag{7.1}$$

where D_{ij}^2 is the squared distance between subjects i and j, x_{ik} is the value of the kth variable for the ith subject, x_{jk} is the value of the kth variable for the jth subject, and p is

[2]See Stewart (1981) for a good review article on the use and misuse of factor analysis.

Table 7.2 **Similarity Matrix Containing Euclidean Distances**

	S1	S2	S3	S4	S5	S6
S1	0.00	2.00	181.00	221.00	625.00	821.00
S2	2.00	0.00	145.00	181.00	557.00	745.00
S3	181.00	145.00	0.00	2.00	136.00	250.00
S4	221.00	181.00	2.00	0.00	106.00	212.00
S5	625.00	557.00	136.00	106.00	0.00	26.00
S6	821.00	745.00	250.00	212.00	26.00	0.00

the number of variables. Table 7.2 gives the similarities, as measured by the squared euclidean distances, between the six subjects. But how do we use the similarities given in Table 7.2 for forming groups or clusters? There are two main types of analytical clustering techniques: hierarchical and nonhierarchical. Each of these techniques is discussed in the following sections using the hypothetical data presented in Table 7.1.

7.5 HIERARCHICAL CLUSTERING

From Table 7.2, subjects S1 and S2 are similar to each other, as are subjects S3 and S4, since the squared euclidean distance between each pair is the same. Either of these two pairs could be selected. The tie is broken randomly. Let us choose subjects S1 and S2 and merge them into one cluster. We now have five clusters; cluster 1 consisting of subjects S1 and S2, and subjects S3, S4, S5, and S6, each forming the remaining four clusters.

The next step is to develop another similarity matrix representing the distances between the five clusters. Since cluster 1 consists of two subjects, we must use some rule for determining the distance or similarity between clusters consisting of more than one subject. A number of different rules or methods have been suggested for computing distances between two clusters. In fact, the various hierarchical clustering algorithms or methods differ mainly with respect to how the distances between the two clusters are computed. Some of the popular methods are:

1. Centroid method.
2. Nearest-neighbor or single-linkage method.
3. Farthest-neighbor or complete-linkage method.
4. Average-linkage method.
5. Ward's method.

We use the centroid method to complete the discussion of the hierarchical clustering algorithm. This is followed by a discussion of the other methods.

7.5.1 Centroid Method

In the centroid method each group is replaced by an *Average Subject* which is the centroid of that group. For example, the first cluster, formed by combining subjects S1 and S2, is represented by the centroid of subjects S1 and S2. That is, cluster 1 has an average education of 5.5 years [i.e., $(5 + 6) \div 2$] and an average income of 5.5 thousand

dollars [i.e., (5 + 6) ÷ 2]. Table 7.3 gives the data for the five new clusters that have been formed. The similarity between the clusters is obtained by using Eq. 7.1 to compute the squared euclidean distance. The table also gives the similarity matrix (squared euclidean distances) among the five clusters.

As can be seen, subjects S3 and S4 have the smallest distance and, therefore, are most similar. Consequently, we can group these two subjects into a new group or cluster. Once again, this cluster will be represented by the centroid of the subjects in this group. Table 7.4 gives the data and similarity matrix containing squared euclidean distance for the four clusters.

Table 7.3 **Centroid Method: Five Clusters**

Data for Five Clusters

Cluster	Cluster Members	Income ($ thous.)	Education (years)
1	S1&S2	5.5	5.5
2	S3	15.0	14.0
3	S4	16.0	15.0
4	S5	25.0	20.0
5	S6	30.0	19.0

Similarity Matrix

	S1&S2	S3	S4	S5	S6
S1&S2	0.00	162.50	200.50	590.50	782.50
S3	162.50	0.00	2.00	135.96	250.00
S4	200.50	2.00	0.00	106.00	212.00
S5	590.50	135.96	106.00	0.00	26.00
S6	782.50	250.00	212.00	26.00	0.00

Table 7.4 **Centroid Method: Four Clusters**

Data for Four Clusters

Cluster	Cluster Members	Income ($ thous.)	Education (years)
1	S1&S2	5.5	5.5
2	S3&S4	15.5	14.5
3	S5	25.0	20.0
4	S6	30.0	19.0

Similarity Matrix

	S1&S2	S3&S4	S5	S6
S1&S2	0.00	181.00	590.50	782.50
S3&S4	181.00	0.00	120.50	230.50
S5	590.50	120.50	0.00	26.00
S6	782.50	230.50	26.00	0.00

Table 7.5 **Centroid Method: Three Clusters**

Data for Three Clusters

Cluster	Cluster Members	Income ($ thous.)	Education (years)
1	S1&S2	5.5	5.5
2	S3&S4	15.5	14.5
3	S5&S6	27.5	19.5

Similarity Matrix

	S1&S2	S3&S4	S5&S6
S1&S2	0.00	181.00	680.00
S3&S4	181.00	0.00	169.00
S5&S6	680.00	169.00	0.00

Subjects S5 and S6 have the smallest distance and therefore are combined to form the third cluster or group, which once again will be represented by the centroid of the subjects in this group. Table 7.5 gives the data for the three clusters and the similarity matrix containing the squared euclidean distances among the clusters.

As can be seen from the matrix in Table 7.5, clusters comprised of subjects S3 and S4 and S5 and S6 have the smallest distance. Therefore, these two clusters are combined to form a new cluster comprised of subjects S3, S4, S5, and S6. The other cluster consists of subjects S1 and S2. Obviously the next step is to group all the subjects into one cluster.

Thus, the hierarchical clustering algorithm forms clusters in a hierarchical fashion. That is, the number of clusters at each stage is one less than the previous one. If there are n observations then at Step 1, Step 2, ..., Step $n-1$ of the hierarchical process the number of clusters, respectively, will be $n-1, n-2, \ldots, 1$. In the case of the centroid method, each cluster is represented by the centroid of that cluster for computing distances between clusters.

Frequently the various steps or stages of the hierarchical clustering process are represented graphically in what is called a dendrogram or tree. Figure 7.2 gives the

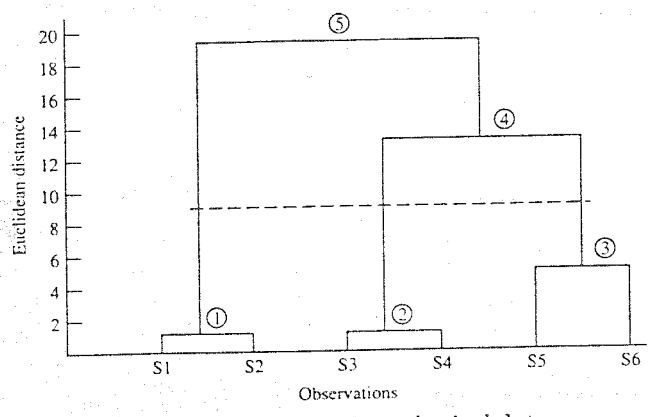


Figure 7.2 Dendrogram for hypothetical data.

dendrogram for the hypothetical data. The circled numbers represent the various steps or stages of the hierarchical process. The observations (i.e., subjects) are listed on the horizontal axis and the vertical axis represents the euclidean distance between the centroids of the clusters. For example, in Step 4 clusters formed in Steps 2 and 3 are merged or combined to form one cluster. The squared euclidean distance between the two merged clusters is 169 or the euclidean distance is 13. In order to determine the cluster composition for a given number of clusters the dendrogram can be *cut* at the appropriate place. A number of different criteria can be used for determining the best number of clusters. These are discussed in Section 7.6.1. For example, the cut shown by the dotted line in Figure 7.2 gives the composition of a three-cluster solution. The three-cluster solution consists of cluster 1 containing subjects S1 and S2, cluster 2 containing subjects S3 and S4, and cluster 3 containing subjects S5 and S6. The dendrogram gives a visual representation of the clustering process; however, it may not be very useful for a large number of subjects as it could become too cumbersome to interpret.

As mentioned previously, there are other hierarchical methods. The first step (i.e., the formation of the first cluster) is the same for all the methods, but after the first step the various methods differ with respect to the procedure used to compute the distances between clusters. The following section discusses other available hierarchical methods.

7.5.2 Single-Linkage or the Nearest-Neighbor Method

Consider the similarity matrix given in Table 7.2. In the centroid method, the distance between clusters was obtained by computing the squared euclidean distance between the centroids of the respective clusters. In the single-linkage method, the distance between two clusters is represented by the minimum of the distance between all possible pairs of subjects in the two clusters. For example, the distance between cluster 1 (consisting of subjects S1 and S2) and subject S3 is the minimum of the following distances:

$$D^2_{13} = 181 \quad \text{and} \quad D^2_{23} = 145.$$

Similarly, the distance between cluster 1 and subject S4 is the minimum of the following distances:

$$D^2_{14} = 221 \quad \text{and} \quad D^2_{24} = 181.$$

This procedure results in the following similarity matrix of squared euclidean distances:

	S1&S2	S3	S4	S5	S6
S1&S2	0.00	145.00	181.00	557.00	745.00
S3	145.00	0.00	2.00	136.00	250.00
S4	181.00	2.00	0.00	106.00	212.00
S5	557.00	136.00	106.00	0.00	26.00
S6	745.00	250.00	212.00	26.00	0.00

The next step is to merge subjects S3 and S4 to form a new cluster and develop a new similarity matrix. The squared euclidean distance between cluster 1 (consisting of subjects S1 and S2) and cluster 2 (consisting of subjects S3 and S4) is the minimum of the following distances: $D^2_{13}, D^2_{14}, D^2_{23}$, and D^2_{24}. In general, if cluster k contains n_k subjects and cluster l contains n_l subjects then the distance between the two clusters is the minimum of the distance between $n_k \times n_l$ pairs of distances. The next cluster is formed using the resulting similarity matrix and the procedure is repeated until all the subjects are merged into one cluster.

7.5.3 Complete-Linkage or Farthest-Neighbor Method

The complete-linkage method is the exact opposite of the nearest-neighbor method. The distance between two clusters is defined as the *maximum* of the distances between all possible pairs of observations in the two clusters. Once again consider the similarity matrix given in Table 7.2. The first cluster is formed by merging subjects S1 and S2. The distance between cluster 1 and subject S3 is the *maximum* of the following distances:

$$D_{13}^2 = 181 \quad \text{and} \quad D_{23}^2 = 145,$$

and the distance between cluster 1 and subject S5 is maximum of the following distances:

$$D_{15}^2 = 625.00 \quad \text{and} \quad D_{25}^2 = 557.00.$$

Following the above rule the similarity matrix after the first step (i.e., the five-cluster solution) is

	S1&S2	S3	S4	S5	S6
S1&S2	0.00	181.00	221.00	625.00	821.00
S3	181.00	0.00	2.00	136.00	250.00
S4	221.00	2.00	0.00	106.00	212.00
S5	625.00	136.00	106.00	0.00	26.00
S6	821.00	250.00	212.00	26.00	0.00

From this similarity matrix, it can be seen that the next cluster will consist of subjects S3 and S4. The squared euclidean distance between cluster 1 (consisting of subjects S1 and S2) and cluster 2 (consisting of subjects S3 and S4) will be the *maximum* of the following distances: $D_{13}^2, D_{14}^2, D_{23}^2$, and D_{24}^2. In general, if cluster k contains n_k subjects and cluster l contains n_l subjects then the distance between the two clusters is the maximum of the distances between $n_k \times n_l$ pairs of distances. The next cluster is formed using the resulting similarity matrix and the procedure is repeated until all the observations are merged into one cluster.

7.5.4 Average-Linkage Method

In the average-linkage method the distance between two clusters is obtained by taking the average distance between all pairs of subjects in the two clusters. For example, from the similarity matrix given in Table 7.2 the first cluster is formed by merging subjects S1 and S2. The distance between cluster 1 and subject S3 is the average of

$$D_{13}^2 \text{ and } D_{23}^2,$$

which is equal to $(181 + 145) \div 2 = 163$. The resulting similarity matrix, after the first cluster has been formed, is given by:

	S1&S2	S3	S4	S5	S6
S1&S2	0.00	163.00	201.00	591.00	783.00
S3	163.00	0.00	2.00	136.00	250.00
S4	201.00	2.00	0.00	106.00	212.00
S5	591.00	136.00	106.00	0.00	26.00
S6	783.00	250.00	212.00	26.00	0.00

Once again, the second cluster is formed by combining subjects S3 and S4. And the distance between the second and the first cluster is given by the average of the

following distances: $D^2_{13}, D^2_{14}, D^2_{23}$, and D^2_{24}. In general, the distance between cluster k and cluster l is given by the average of the $n_k \times n_l$ squared euclidean distances, where n_k and n_l are the number of subjects in clusters k and l, respectively.

7.5.5 Ward's Method

The Ward's method does not compute distances between clusters. Rather, it forms clusters by maximizing within-clusters homogeneity. The within-group (i.e., within-cluster) sum of squares is used as the measure of homogeneity. That is, the Ward's method tries to minimize the total within-group or within-cluster sums of squares. Clusters are formed at each step such that the resulting cluster solution has the fewest within-cluster sums of squares. The within-cluster sums of squares that is minimized is also known as the error sums of squares (ESS). Once again, consider the hypothetical data given in Table 7.1. Initially each observation is a cluster and therefore the ESS is zero. The next step is to form five clusters, one cluster of size two and the other four clusters of size one (i.e., each subject being a cluster). For example, we can have one cluster consisting of subjects S1 and S2 and the other four clusters consisting of subjects S3, S4, S5, and S6, respectively. The ESS for the cluster with two observations (i.e., S1 and S2) is

Table 7.6 **Ward's Method**

	Members in Cluster					
Cluster Solution	**1**	**2**	**3**	**4**	**5**	**ESS**
(a) All Possible Five-Cluster Solutions						
1	S1,S2	S3	S4	S5	S6	1.0
2	S1,S3	S2	S4	S5	S6	90.5
3	S1,S4	S2	S3	S5	S6	110.5
4	S1,S5	S2	S3	S4	S6	312.5
5	S1,S6	S2	S3	S4	S5	410.5
6	S2,S3	S1	S4	S5	S6	72.5
7	S2,S4	S1	S3	S5	S6	90.5
8	S2,S5	S1	S3	S4	S6	278.5
9	S2,S6	S1	S3	S4	S5	372.5
10	S3,S4	S1	S2	S5	S6	1.0
11	S3,S5	S1	S2	S4	S6	68.0
12	S3,S6	S1	S2	S4	S5	125.0
13	S4,S5	S1	S2	S3	S6	53.0
14	S4,S6	S1	S2	S3	S5	106.0
15	S5,S6	S1	S2	S3	S4	13.0
(b) All Possible Four-Cluster Solutions						
1	S1,S2,S3	S4	S5	S6		109.333
2	S1,S2,S4	S3	S5	S6		134.667
3	S1,S2,S5	S3	S4	S6		394.667
4	S1,S2,S6	S3	S4	S5		522.667
5	S1,S2	S3,S4	S5	S6		2.000
6	S1,S2	S3,S5	S4	S6		69.000
7	S1,S2	S3,S6	S4	S5		126.000
8	S1,S2	S4,S5	S3	S6		54.000
9	S1,S2	S4,S6	S3	S5		107.000
10	S1,S2	S5,S6	S3	S4		14.000

$(5-5.5)^2+(6-5.5)^2+(5-5.5)^2+(6-5.5)^2 = 1.0$ and the ESS for the remaining four clusters is zero as each cluster consists of only one observation. Therefore, the total ESS for the cluster solution is 1.0.

Table 7.6 gives all the fifteen possible five-cluster solutions along with their ESS.[3] Based on the criterion of minimizing ESS, cluster solution 1 or 10 can be selected. The tie is broken randomly. Let us select the first cluster solution; that is, merge subjects S1 and S2.

The next step is to form four clusters. There are ten possible four-cluster solutions (i.e., $(5\times4)/2$). Table 7.6 also gives the four-cluster solutions along with their ESS. For example, the ESS for the cluster consisting of subjects S1, S2, and S3 is $(5-8.67)^2+(6-8.67)^2+(15-8.67)^2+(5-8.33)^2+(6-8.33)^2+(14-8.33)^2 = 109.33$, and the ESS for the first four-cluster solution will be 109.33 as the ESS for the remaining three clusters is zero. Cluster solution 5 is the one that minimizes the ESS. This procedure is repeated for all the remaining steps.

7.6 HIERARCHICAL CLUSTERING USING SAS

In this section we will discuss the resulting output from the hierarchical clustering procedure, PROC CLUSTER, in SAS. Table 7.7 gives the SAS commands for clustering the hypothetical data in Table 7.1. The SIMPLE option requests simple or descriptive statistics for the data. NOEIGEN instructs that the eigenvalues and eigenvectors of the covariance matrix among the variables should not be reported. This information is not needed for interpreting the cluster solution. The METHOD option specifies that the centroid method be used for clustering the observations. RMSSTD and RSQUARE request certain statistics used to evaluate the cluster solution. NONORM indicates that the euclidean distances should not be normalized. Normalizing essentially divides the euclidean distance between the two observations or clusters by the average of the euclidean distances between all pairs of observations. Consequently, normalizing of the euclidean distances does not affect the cluster solution and hence is not really required.

Table 7.7 **SAS Commands**

```
TITLE CLUSTER ANALYSIS FOR DATA IN TABLE 7.1;
DATA TABLE1;
 INPUT SID $ 1-2 INCOME 4-5 EDUC 7-8;
CARDS;
insert data here
PROC CLUSTER SIMPLE NOEIGEN METHOD=CENTROID RMSSTD RSQUARE
 NONORM OUT=TREE;
  ID SID;
  VAR INCOME EDUC;
PROC TREE DATA=TREE OUT=CLUS3 NCLUSTERS=3;
 ID SID;
 COPY INCOME EDUC;
PROC SORT; BY CLUSTER;
PROC PRINT; BY CLUSTER;
   VAR SID INCOME EDUC;
TITLE2 '3-CLUSTER SOLUTION ';
```

[3]In general, there will be $n(n-1)/2$ possible cluster solutions.

The PROC TREE procedure uses the output from PROC CLUSTER to develop a list of cluster members for a given cluster solution. For illustration purposes assume that we are interested in obtaining cluster membership for a three-cluster solution.

7.6.1 Interpreting the SAS Output

The resulting output is given in Exhibit 7.1. To facilitate discussion the SAS output is labeled with circled numbers that correspond to the bracketed numbers in the text.

Descriptive Statistics

Basic statistics such as the mean, standard deviation, skewness, kurtosis, and the coefficient of bimodality are reported [1]. These statistics are normally used to give some indication regarding the distribution of the variables; however, knowledge of the

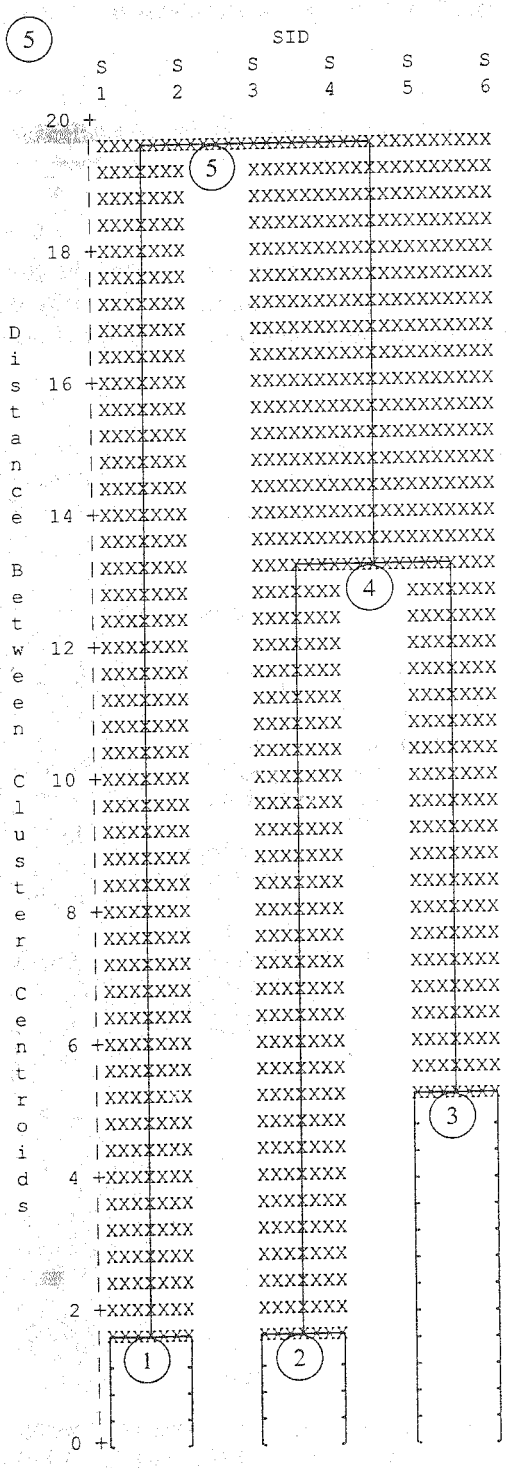
Exhibit 7.1

SAS output for cluster analysis on data in Table 7.1

```
Centroid Hierarchical Cluster Analysis

(1) Simple Statistics

                Mean      Std Dev    Skewness    Kurtosis  Bimodality

INCOME       16.1667       9.9883      0.2684     -1.4015      0.2211
EDUC         13.1667       6.3692     -0.4510     -1.8108      0.2711

(2) Root-Mean-Square Total-Sample Standard Deviation = 8.376555

      (3a)                (3b)                (3c)      (3d)      (3e)
      Number                                  Frequency RMS STD
Step     of                                   of New    of New    Semipartial
Number Clusters    Clusters Joined            Cluster   Cluster   R-Squared

  1      5          S1       S2                2       0.707107  0.001425
  2      4          S3       S4                2       0.707107  0.001425
  3      3          S5       S6                2       2.549510  0.018527
  4      2          CL4      CL3               4       5.522681  0.240855
  5      1          CL5      CL2               6       8.376555  0.737767

(3f)       (3g)
          Centroid
R-Squared Distance

0.998575   1.4142
0.997150   1.4142
0.978622   5.0990
0.737767  13.0000
0.000000  19.7041

(4) CLUSTER=1                CLUSTER=2                    CLUSTER=3

OBS  SID  INCOME  EDUC     OBS  SID  INCOME  EDUC     OBS  SID  INCOME  EDUC
 1   S1     5      5        3   S3    15     14        5   S5    25     20
 2   S2     6      6        4   S4    16     15        6   S6    30     19
```

(continued)

Exhibit 7.1 (continued)



distribution of the data is not very useful as cluster analysis does not make any distributional assumptions.

The *root-mean-square total-sample standard deviation* (RMSSTD) is simply a measure of the standard deviation of all the variables [2]. The RMSSTD is given by

$$\begin{aligned} \text{RMSSTD} &= \sqrt{\frac{(n-1)\sum_{j=1}^{p}\hat{s}_j^2}{p(n-1)}} \\ &= \sqrt{\frac{\sum_{j=1}^{p}\hat{s}_j^2}{p}}, \end{aligned} \tag{7.2}$$

which is equal to

$$\begin{aligned} \text{RMSSTD} &= \sqrt{\frac{9.9883^2 + 6.3692^2}{2}} \\ &= 8.377, \end{aligned}$$

and is the same as that reported in the output [2]. The smaller the value, the more homogeneous the observations are with respect to the variables and vice versa. Since root-mean-square is scale dependent, it should only be used to compare the homogeneity of data sets whose variables are measured using similar scales.

Cluster Solution

The Step Number column is not a part of the output generated by SAS and has been added to facilitate discussion of the output. At each step a cluster or group is formed either by joining two observations, by joining two previously formed clusters, or by joining an observation and a previously formed cluster.

The Number of Clusters column gives the total number of clusters, including the one formed in the current step [3a]. The cluster formed at any given step is labeled as CLj, where j is the total number of clusters at the given step. There will be $n - 1$ clusters in the first step, $n - 2$ clusters in the second step, $n - 3$ clusters in the third step, and so on. Therefore, the cluster formed in Step 1 is represented as CL($n - 1$), the cluster formed in Step 2 as CL($n - 2$), and so on.

For any given step, the Clusters Joined column gives the clusters or the observations that are joined to form the cluster at the given step [3b]. A cluster consisting of a single observation or subject is denoted by the observation identification number, whereas a cluster consisting of two or more observations is represented by the cluster identification number.

The Frequency of New Cluster column gives the size of the cluster formed at any given step [3c]. The information provided in the columns discussed above can be used to determine the number of clusters at any given step, the size of the cluster formed, and its composition. For example, in Step 1 there are a total of 5 clusters; the size of the CL5 cluster formed at this step is 2, and it consists of subjects S1 and S2. CL4 is the cluster formed at Step 2 and it consists of subjects S3 and S4. CL3 is formed at Step 3 and consists of subjects S5 and S6. In Step 4, the total number of clusters is 2. The CL2 cluster is formed at Step 4 by merging clusters CL4 and CL3; therefore, cluster CL2 consists of subjects S3, S4, S5, and S6.

Evaluating the Cluster Solution and Determining the Number of Clusters

Given the cluster solution, the next obvious steps are to evaluate the solution and determine the number of clusters present in the data. A number of statistics are available

for evaluating the cluster solution at any given step and to determine the number of clusters. The most widely used statistics are:

1. Root-mean-square standard deviation (RMSSTD) of the new cluster.
2. Semipartial R-squared (SPR).
3. R-squared (RS).
4. Distance between two clusters.

These statistics all provide information about the cluster solution at any given step, the new cluster formed at this step, and the consequences of forming the new cluster. A conceptual understanding and use of all these statistics can be gained by computing them for the hypothetical data. Information in Table 7.8 will be used in computing the various statistics. The table reports the within-group sum of squares and the corresponding degrees of freedom (see Section 3.1.6 of Chapter 3). For example, in Step 4 the newly formed cluster, *CL2*, consists of observations S3, S4, S5, and S6. The within-group sum of squares for the four observations in *CL2* for income and education, respectively, are 157.000 and 26.000, and the corresponding degrees of freedom are 3 for each variable. The within-group sum of squares pooled across all the clustering variables (i.e., income and education) is 183.000 and the pooled degrees of freedom are 6.

RMSSTD OF THE CLUSTER [3d]. RMSSTD is the pooled standard deviation of all the variables forming the cluster. For example, the cluster in Step 4 is formed by merging clusters CL2 and CL4 and consists of subjects S3, S4, S5, and S6. The variance of income is given by the *SS* for this variable divided by its degrees of freedom. That is, the variance is equal to 52.333 (157/3). Similarly, the variance of education is given by 8.667 (26/3). The pooled variance of all the variables used for clustering is obtained by dividing the pooled *SS* by the pooled degrees of freedom. That is,

$$\begin{aligned}\text{Pooled variance} &= \frac{\text{Pooled } SS \text{ for all the variables}}{\text{Pooled degrees of freedom for all the variables}}\\ &= \frac{157 + 26}{3 + 3}\\ &= 30.500,\end{aligned}$$

or the pooled standard deviation is equal to 5.523 ($\sqrt{30.500}$) and is known as the root mean square standard deviation (RMSSTD) of the cluster formed at a given step. As can be seen, RMSSTD is simply the pooled standard deviation of all variables for observations comprising the cluster formed at a given step. Since the objective of cluster analysis is to form homogeneous groups, the RMSSTD of a cluster should be as small as possible. Greater values of RMSSTD suggest that the new cluster may not be homogeneous and vice versa. However, it should be noted that there are no guidelines to decide what is "small" and what is "large."

R-SQUARED (RS) [3f]. RS is the ratio of SS_b to SS_t. As discussed in Chapter 3, SS_b is a measure of the extent to which groups are different from each other. Since $SS_t = SS_b + SS_w$, the greater the SS_b the smaller the SS_w and vice versa. Consequently, for a given data set the greater the differences between groups the more homogeneous each group is and vice versa. Therefore, RS measures the extent to which groups or clusters are different from each other. Alternatively, one can say it also measures the extent to which the groups are homogeneous. The value of RS ranges from 0 to 1, with 0 indicating no differences among groups or clusters and 1 indicating maximum

Table 7.8 **Within-Group Sum of Squares and Degrees of Freedom for Clusters Formed in Steps 1, 2, 3, 4, and 5**

		Within-Group Sum of Squares			Degrees of Freedom		
Step Number	Cluster	Income	Education	Pooled	Income	Education	Pooled
1	CL5	0.500	0.500	1.000	1	1	2
2	CL4	0.500	0.500	1.000	1	1	2
3	CL3	12.500	0.500	13.000	1	1	2
4	CL2	157.000	26.000	183.000	3	3	6
5	CL1	498.333	202.833	701.166	5	5	10

differences among groups. Following is an example of the computation of RS for the clusters at Step 4.

At Step 4 we have two clusters or groups: CL2, the cluster formed in Step 4 consisting of subjects S3, S4, S5, and S6, and CL5 consisting of two subjects (i.e., S1 and S2) formed in Step 1 [3b]. The SS_w for income for CL4 is equal to 157 and for CL2 it is equal to 0.50, giving a pooled SS_w of 157.50. Similarly, the pooled SS_w for education for CL4 and CL2 is 26.50 (i.e., 26 + 0.50). Therefore, the total pooled SS_w across all the clustering variables is 184 (i.e., 157.50 + 26.50). Since the total pooled sum of squares is 701.166, the pooled SS_b is 517.166 (701.166 − 184), giving an RS of .738 (517.166 ÷ 701.166).

SEMIPARTIAL R-SQUARED (SPR) [3e]. As discussed previously, the new cluster formed at any given step is obtained by merging two clusters formed in previous steps. The difference between the pooled SS_w of the new cluster, and the sum of pooled SS_w's of clusters joined to obtain the new cluster is called *loss of homogeneity*. If the loss of homogeneity is zero, then the new cluster is obtained by merging two perfectly homogeneous clusters. On the other hand, if loss of homogeneity is large then the new cluster is obtained by merging two heterogeneous clusters.

For example, cluster CL2, formed at Step 4, is obtained by joining CL4 and CL3 [3b]. The loss of homogeneity due to joining CL4 and CL3 is given by the pooled SS_w of CL2 formed at Step 4 minus the sum of the pooled SS_w's of CL3 and CL4 formed at Steps 3 and 2, respectively. Usually this quantity is divided by the pooled SS for the total sample. The resulting ratio is referred to as SPR. Thus, SPR is the loss of homogeneity due to combining two groups or clusters to form a new group or cluster. A smaller value would imply that we are merging two homogeneous groups and vice versa. Therefore, for a good cluster solution SPR should be low. SPR for the two-cluster solution in Step 4 is given by

$$\frac{(183) - (1.00 + 13.00)}{701.166} = 0.241.$$

DISTANCE BETWEEN CLUSTERS [3g]. The output reports the distance between the two clusters that are merged at a given step. In the centroid method it is simply the euclidean distance between the centroids of the two clusters that are to be joined or merged and it is termed the centroid distance (CD); for single linkage it is the minimum euclidean distance (MIND) between all possible pairs of points or subjects; for complete linkage it is the maximum euclidean distance (MAXD) between all pairs of subjects; and for Ward's method it is the between-group sum of squares for the two clusters (i.e., SS_b). The CD for the two-cluster solution obtained by merging clusters CL4 and CL3 is (see Table 7.5)

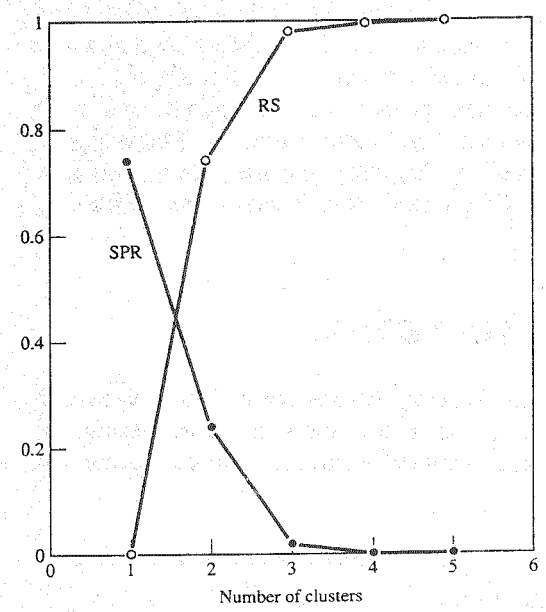
$$\sqrt{(27.5 - 15.5)^2 + (19.5 - 14.5)^2} = 13.00.$$

CD, obviously, should be small to merge the two clusters. A large value for CD would indicate that two dissimilar groups are being merged. If the NONORM option had not been specified then the preceding CD would be divided by the average of the euclidean distances between all observations. This is the only effect the NONORM option has.

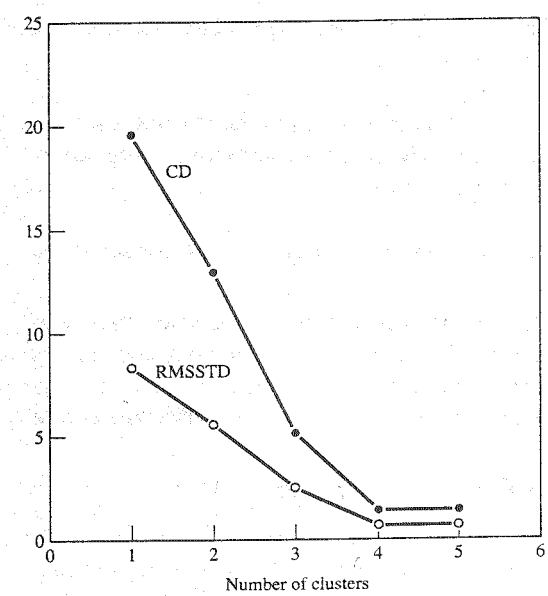
Table 7.9 gives a summary of the statistics previously discussed for evaluating the cluster solution. These statistics can also be used for determining the number of clusters in the data set. Essentially, one looks for a big jump in the value of a given statistic. One could plot the statistics and look for an elbow. For example, Figure 7.3 gives plots of SPR, RS, RMSSTD, and CD. It is clear that there is a "big" change in the values when

Table 7.9 **Summary of the Statistics for Evaluating Cluster Solution**

Statistic	Concept Measured	Comments
RMSSTD	Homogeneity of new cluster	Value should be small
SPR	Homogeneity of merged clusters	Value should be small
RS	Heterogeneity of clusters	Value should be high
CD	Homogeneity of merged clusters	Value should be small



(a)



(b)

Figure 7.3 Plots of (*a*) SPR and RS and (*b*) RMSSTD and CD.

going from a three-cluster to a two-cluster solution. Consequently, it appears that there are three clusters in the data set. Furthermore, the three clusters are well separated as suggested by RS [3f], and the clusters are homogeneous as evidenced by the low value of RMSSTD [3d], SPR [3e], and CD [3g].

It should be noted that the sampling distributions of all these statistics are not known and, therefore, these statistics are basically heuristics. It is recommended that the researcher consider all of these heuristics and the objective of the study in order to assess the cluster solution and determine the number of clusters. SAS also gives the dendrogram of the cluster solution [5]. The additional hand-drawn lines and the step numbers are included in the dendrogram to show its similarity to that depicted in Figure 7.2.

The output from the PROC TREE procedure gives the cluster membership for a given cluster solution [4]. For a given cluster solution, cluster members of each cluster, along with the values of the variables used for clustering, are listed in this section. For example, members of the first cluster for a three-cluster solution are subjects S1 and S2.

7.7 NONHIERARCHICAL CLUSTERING

In nonhierarchical clustering, the data are divided into k partitions or groups with each partition representing a cluster. Therefore, as opposed to hierarchical clustering, the number of clusters must be known *a priori*. Nonhierarchical clustering techniques basically follow these steps:

1. Select k initial cluster centroids or seeds, where k is the number of clusters desired.
2. Assign each observation to the cluster to which it is the closest.
3. Reassign or reallocate each observation to one of the k clusters according to a predetermined stopping rule.
4. Stop if there is no reallocation of data points or if the reassignment satisfies the criteria set by the stopping rule. Otherwise go to Step 2.

Most of the nonhierarchical algorithms differ with respect to: (1) the method used for obtaining initial cluster centroids or seeds; and (2) the rule used for reassigning observations. Some of the methods used to obtain initial seeds are

1. Select the first k observations with nonmissing data as centroids or seeds for the initial clusters.
2. Select the first nonmissing observation as the seed for the first cluster. The seed for the second cluster is selected such that its distance from the previous seed is greater than a certain selected distance. The third seed is selected such that its distance from previously selected seeds is greater than the selected distance, and so on.
3. Randomly select k nonmissing observations as cluster centers or seeds.
4. Refine the selected seeds using certain rules such that they are as far apart as possible. Some of these rules are discussed in Sections 7.7.2 and 7.8.
5. Use a heuristic that identifies cluster centers such that they are as far apart as possible.
6. Use seeds supplied by the researcher.

Once the seeds are identified, initial clusters are formed by assigning each of the remaining $n - k$ observations to the seed to which the observation is the closest.

Nonhierarchical algorithms also differ with respect to the procedure used for reassigning subjects to the k clusters. Some of the reassignment rules are

1. Compute the centroid of each cluster and reassign subjects to the cluster whose centroid is the nearest. The centroids are not updated while assigning each observation to the k clusters; they are recomputed after the assignment for all the observations have been made. If the change in the cluster centroids is greater than a selected convergence criterion then another pass at reassignment is made and cluster centroids are recomputed. The reassignment process is continued until the change in the centroids is less than the selected convergence criterion.
2. Compute the centroid of each cluster and reassign subjects to the cluster whose centroid is the nearest. For the assignment of each observation, recompute the centroid of the cluster *to* which the observation is assigned and the cluster *from* which the observation is assigned. Once again, reassignment is continued until the change in cluster centroids is less than the selected convergence criterion.
3. Reassign the observations such that some statistical criterion is minimized. These methods are commonly referred to as hill-climbing methods. Some of the objective functions or the statistical criteria that can be minimized are
   (a) trace of the within-group SSCP matrix (i.e., minimize ESS).
   (b) determinant of the within-group SSCP matrix.
   (c) trace of $\mathbf{W}^{-1}\mathbf{B}$, where $\mathbf{W}$ and $\mathbf{B}$ are, respectively, the within-group and between-group SSCP matrices.
   (d) largest eigenvalue of the $\mathbf{W}^{-1}\mathbf{B}$ matrix.

As can be seen, a variety of clustering algorithms can be developed depending on the combination of the initial partitioning and the reassignment rule employed. Three popular types of nonhierarchical algorithms will be discussed and illustrated using the hypothetical data given in Table 7.1. For illustration purposes we will assume that three clusters are desired and that a convergence criterion of .02 has been specified.

7.7.1 Algorithm I

This algorithm selects the first k observations as cluster centers. For the present example, the first three observations are selected as seeds or centroids for the clusters. Table 7.10 gives the initial cluster centroids, the squared euclidean distance of each observation from the centroid of each cluster, and the assignment of each observation. The next step is to compute the centroid of each cluster, given in Table 7.11, and the change in cluster centroids, also reported in the table. For example, the change in cluster centroid of cluster 3 with respect to income is 6.5 (21.5 − 15). Because the change in cluster seeds is greater than the convergence criterion of .02, a reallocation of observations is done in the next iteration.

Observations are reassigned by computing the distance of each observation from the centroid. Table 7.12 gives the recomputed distance, previous assignment and reassignment of each observation, and the cluster centroids. As can be seen, none of the observations are reassigned and the change in cluster centroids is zero. Consequently, no more reassignments are made and the final three-cluster solution consists of one cluster having four observations and the remaining two clusters having one observation each.

Table 7.10 **Initial Cluster Centroids, Distance from Cluster Centroids, and Initial Assignment of Observations**

Initial Cluster Centroids

	Cluster		
Variable	1	2	3
Income	5	6	15
Education	5	6	14

Distance from Cluster Centroids and Initial Assignment of Observations

	Distance from Cluster Centroid			
Observation	1	2	3	Assigned to Cluster
S1	0	2	181	1
S2	2	0	145	2
S3	181	145	0	3
S4	221	181	2	3
S5	625	557	136	3
S6	821	745	250	3

Table 7.11 **Centroid of the Three Clusters and Change in Cluster Centroids**

	Clusters		
Variable	1	2	3
Cluster Centroids			
Income	5	6	21.5
Education	5	6	17.0

	Clusters		
Variable	1	2	3
Change in Cluster Centroids			
Income	0	0	6.5
Education	0	0	3.0

Table 7.12 **Distance from Centroids and First Reassignment of Observations to Clusters**

	Distance from Cluster			Cluster Assignment	
Observation	1	2	3	Previous	Reassignment
S1	0	2	416.25	1	1
S2	2	0	361.25	2	2
S3	181	145	51.25	3	3
S4	221	181	34.25	3	3
S5	625	557	21.25	3	3
S6	821	990	76.25	3	3

7.7.2 Algorithm II

This algorithm differs from Algorithm I with respect to how the initial seeds are modified. The first three observations are selected as cluster seeds. Then each of the remaining observations is evaluated to determine if it can replace any of the previously selected seeds according to the following rule: The seed that is a candidate for replacement is from the two seeds (i.e., pair of seeds) that are closest to each other. An observation qualifies to replace one of the two identified seeds if the distance between the seeds is less than the distance between the observation and the nearest seed. If the observation qualifies, then the seed that is replaced is the one closest to the observation. This rule, and its variants, are used in the nonhierarchical clustering procedure in SAS.

For example, in the previous algorithm observations S1, S2, and S3 were selected as seeds for the three clusters. Table 7.2 gives the squared euclidean distances among the observations. The smallest distance between the seeds is for seeds S1 and S2, and this distance is equal to 2. Observation S4 does not qualify as a replacement seed because the distance between S1 and S2 is not less than the distance between S4 and the nearest seed (i.e., distance between S4 and seed S3). However, observation S5 qualifies as the replacement seed because the distance between seeds S1 and S2 is less than the distance between S5 and the nearest seed (i.e., S5 and S3). Seed S2 is replaced by S5 because the distance between S5 and S2 is smaller than the distance between S5 and S1. The three seeds now are S1, S3, and S5 and the two closest seeds are S3 and S5 with a distance of 136.00 (see Table 7.2). Observation S6 does not qualify for replacement as the distance between the S3 and S5 is not less than the distance between S6 and the nearest seed (i.e., S6 and S5). Therefore, the resulting seeds are S1, S3, and S5.

Table 7.13 gives the assignment of each observation to the three clusters and also the reassignment. As can be seen, none of the observations are reassigned, resulting in no change in the cluster centroids. Consequently, no more reassignments are done and the resulting three-cluster solution is that given in Table 7.13. However, the cluster solution of this step is different than the cluster solution of Algorithm I. Consequently, as has also been shown in simulation studies, nonhierarchical clustering techniques are quite sensitive to the selection of the initial seeds. Algorithms I and II are commonly referred to as *K-means* clustering.

7.7.3 Algorithm III

As mentioned previously, the nonhierarchical clustering programs differ with respect to initial partitioning and the reassignment rule. Here we describe an alternative heuristic for selecting the initial seeds and a reassignment rule that explicitly minimizes the ESS (i.e., trace of the within-group SSCP matrix).

Let $Sum(i)$ be the sum of the values of the variables for each observation and k be the desired number of clusters. The initial allocation of observation i to cluster C_i is given by the integer part of the following equation:

$$C_i = \frac{(Sum(i) - Min)(k - 0.0001)}{Max - Min} + 1 \tag{7.3}$$

where C_i is the cluster to which observation i should be assigned, Max and Min are, respectively, the maximum and minimum of $Sum(i)$, and k is the number of clusters desired. Table 7.14 gives the $Sum(i)$, C_i, and the initial allocation of the data points and the centroid of the three clusters.

Table 7.13 **Initial Assignment, Cluster Centroids, and Reassignment**

Initial Assignment

Observation	Distance from Cluster Centroid 1	2	3	Assigned to Cluster
S1	0.00	181.00	625.00	1
S2	2.00	145.00	557.00	1
S3	181.00	0.00	136.00	2
S4	221.00	2.00	106.00	2
S5	625.00	136.00	0.00	3
S6	821.00	250.00	26.00	3

Cluster Centroids

Variable	Clusters 1	2	3
Income	5.5	15.5	27.5
Education	5.5	14.5	19.5

Reassignment

Observation	Clusters 1	2	3	Cluster Assignment Previous	Reassignment
S1	0.50	200.50	716.50	1	1
S2	0.50	162.50	644.50	1	1
S3	162.50	0.50	186.50	2	2
S4	200.50	0.50	152.50	2	2
S5	590.50	120.50	6.50	3	3
S6	600.50	230.50	6.50	3	3

Table 7.14 **Initial Assignment**

Subject	Income ($ thous.)	Education (years)	*Sum(i)*	C_i	Assigned to Cluster
S1	5	5	10	1	1
S2	6	6	12	1	1
S3	15	14	29	2	2
S4	16	15	31	2	2
S5	25	20	45	3	3
S6	30	19	49	3	3

Centroid of Three Clusters

Variable	Clusters 1	2	3
Income	5.5	15.5	27.5
Education	5.5	14.5	19.5

Table 7.15 **Change in ESS Due to Reassignment**

Observation	Cluster	Change in ESS if Assigned to Cluster 3	2	1	Reassignment
S1	1	1074.50	300.50	-	1
S2	1	966.50	243.75	-	1
S3	2	279.50	-	243.50	2
S4	2	228.50	-	300.75	2
S5	3	-	177.50	882.50	3
S6	3	-	585.50	1170.50	3

Next, the observations are reassigned such that the statistical criterion, ESS, is minimized. For example, the change in ESS if $S1$ belonging to cluster 1 is reassigned to cluster 3 will be

$$\text{Change in ESS} = \frac{3}{2}[(5-27.5)^2 + (5-19.5)^2] - \frac{1}{2}[(5-5.5)^2 + (5-5.5)^2]$$
$$= 1074.750 - 0.250 = 1074.500.$$

In this equation, the quantity $(5-27.5)^2 + (5-19.5)^2$ gives the increase in the sum of squares of the cluster *to* which the observation is assigned (i.e., cluster 3), and the quantity $(5-5.5)^2 + (5-5.5)^2$ gives the decrease in the sum of squares of the cluster *from* which the observation is assigned (i.e., cluster 1). The weight for each term is the ratio of the number of observations after and before the reassignment. A negative ESS for the quantity in the preceding equation indicates that the total ESS will decrease if the observation is reassigned to the respective cluster. This change in ESS is computed for reassignment of the observation to each of the other clusters, and the observation is reassigned to the cluster that results in the greatest decrease in ESS. This procedure is repeated for all the observations. Table 7.15 gives the change in ESS for each observation and the reassignment. As can be seen, the reassignment does not result in a reduction of ESS. Therefore, the initial cluster solution is the final one.

7.8 NONHIERARCHICAL CLUSTERING USING SAS

The data in Table 7.1 are used to discuss the output resulting from FASTCLUS, a nonhierarchical clustering procedure in SAS. Table 7.16 gives the SAS commands. Following is a discussion of the various options for the FASTCLUS procedure. The RADIUS

Table 7.16 **SAS Commands for Nonhierarchical Clustering**

```
OPTIONS NOCENTER;
TITLE NONHIERARCHICAL CLUSTERING OF DATA IN TABLE 7.1;
DATA TABLE1;
 INPUT SID $ 1-2 INCOME 4-5 EDUC 7-8;
CARDS;
insert data here
PROC FASTCLUS RADIUS=0 REPLACE=FULL MAXCLUSTERS=3 MAXITER=20
 LIST DISTANCE;
```

Table 7.17 **Observations Selected as Seeds for Various Combinations of Radius and Replace Options**

	Radius		
Replace	0	10	20
None	S1, S2 and S3	S1, S3 and S5	S1 and S5
Part	S1, S5 and S6	S1, S3 and S5	S1 and S5
Full	S1, S4 and S6	S1, S3 and S6	S1 and S5

and the REPLACE options control the selection of initial cluster seeds and the rules used to replace them. The RADIUS option specifies the minimum euclidean distance between an observation in consideration for potential seed and the existing seeds. If the observation does not meet this criterion then it is not selected as the seed. Caution needs to be exercised in specifying the minimum distance because too large a distance may result in the number of seeds being less than the number of desired clusters, or it may result in outliers being selected as seeds. For example, if one were to specify a RADIUS of 20 for the data in Table 7.1 then only two observations qualify as seeds, resulting in a two-cluster solution even though a three-cluster solution is desired.

The REPLACE option controls seed replacement after the initial selection. One can specify partial, full, or no replacement. If REPLACE = NONE then the seeds are not replaced. The replacement procedure described in Algorithm II can be obtained by specifying REPLACE = PART. The REPLACE = FULL option uses two criteria or rules for replacing seeds. The first criterion is the same as that discussed in Algorithm II but if this criterion is not satisfied then a second criterion is used to determine if the observation qualifies for replacing a current seed. We do not discuss this criterion here; however, the interested reader is referred to the SAS manual for a discussion of this criterion. To illustrate the effect of the radius and replace options, Table 7.17 gives the selection of seeds for some of the various combinations of radius and replace options. It is clear that different combinations result in different sets of seeds. Note that for the RADIUS = 20 option only two seeds are selected, as there are only two observations with distance greater than 20. Consequently, there will only be two clusters. We suggest using a radius of zero with full replacement option (which is the default option) as this gives seeds that are reasonably far apart and also guards against the selection of outliers as seeds.

The MAXCLUSTERS option specifies the number of clusters desired. The maximum number of iterations or reallocations can be specified by the MAXITER option. The iterations (i.e., reallocation of observations among clusters) are continued until the change in the cluster centroids of two successive iterations is less than the convergence value specified by the researcher. Default values for MAXITER and CONVERGE, respectively, are 20 and 0.02. Exhibit 7.2 gives the SAS output.

7.8.1 Interpreting the SAS Output

Initial Cluster Seeds and Reassignment

This part of the output gives the selected options [1]. Next, the initial cluster seeds are reported along with the minimum distance between the seeds [2]. Note that the

Exhibit 7.2

Nonhierarchical clustering on data in Table 7.1

```
FASTCLUS Procedure

(1) Replace=FULL  Radius=0  Maxclusters=3  Maxiter=20  Converge=0.02

(2) Initial Seeds

Cluster          INCOME          EDUC
-----------------------------------------
   1             5.0000          5.0000
   2            30.0000         19.0000
   3            16.0000         15.0000


Minimum Distance Between Seeds = 14.56022

(3) Iteration   Change in Cluster Seeds
                     1           2           3
-----------------------------------------------
     1        0.707107     2.54951    0.707107
     2               0           0           0


(4)                    Statistics for Variables

Variable       Total STD    Within STD    R-Squared     RSQ/(1-RSQ)
-------------------------------------------------------------------
INCOME          9.988327      2.121320     0.972937       35.950617
EDUC            6.369197      0.707107     0.992605      134.222222
(4a) OVER-ALL   8.376555      1.581139     0.978622       45.777778

                      Pseudo F Statistic =     68.67
Approximate Expected Over-All R-Squared =      .
             Cubic Clustering Criterion =      .
WARNING: The two above values are invalid for correlated variables.

(5) Cluster Means

Cluster          INCOME          EDUC
-----------------------------------------
   1             5.5000          5.5000
   2            27.5000         19.5000
   3            15.5000         14.5000
```

initial seeds pertain to observations S1, S6, and S4. The minimum distance is used to assess how far apart the seeds are relative to some other set of initial seeds. For example, the user can try different combinations of the replace and radius options to assess the selection of initial seeds such that the minimum distance between the seeds is the largest. In the iteration history, the first iteration corresponds to the initial allocation of observations into clusters [3]. The second iteration corresponds to a reallocation of observations. Because the change in cluster seeds at the second iteration is less than the convergence criterion, the cluster solution at the second iteration is the final cluster solution.[4] In some data sets it is quite possible that the cluster solution may not converge in the desired number of iterations. That is, each iteration results in reallocation of observations. In this case the user may have to increase the maximum number of iterations or the convergence criterion.

Evaluation of the Cluster Solution

The cluster solution is evaluated using the same statistics discussed earlier. The *overall* RS of 0.978 is quite large suggesting that the clusters are quite homogeneous and well separated [4a]. The *within-std* reported in the output is the same as the RMSSTD discussed earlier except that it is the RMSSTD pooled across all the clusters [4a]. Since the RMSSTD is dependent upon the measurement scale we recommend that for a given cluster solution it should be interpreted only relative to the total RMSSTD. In the present case a value of 0.189 (1.581 ÷ 8.377) is quite low suggesting that the resulting clusters are quite homogeneous [4a]. As discussed earlier, in hierarchical clustering techniques RS and RMSSTD are used to determine the number of clusters present in the data set. If one is not sure about the number of clusters in nonhierarchical clustering technique then one can rerun the analysis to obtain a solution for a different number of clusters and use RS and RMSSTD to determine the number of clusters. Table 7.18 gives the RS and RMSSTD for 2-, 3-, 4-, and 5-cluster solutions. These statistics suggest a 3-cluster solution.

Occasionally the researcher is also interested in determining how good the cluster solution is with respect to each clustering variable. This can be done by examining the RS and RMSSTD for each variable reported by the output [4]. RS values of .993 and .973, respectively, for education and income suggest that the clusters are well separated with respect to these two variables; however, the separation with respect to education is slightly more than with respect to income. Similarly, relative values of .111 (.707 ÷ 6.369) and .212 (2.121 ÷ 9.988), respectively, for education and income suggest that the cluster solution is homogeneous with respect to these variables and once again the clusters are more homogeneous with respect to education than with respect to income.

Table 7.18 **RS and RMSSTD for 2-, 3-, 4-, and 5-Cluster Solutions**

Number of Clusters	RS	RMSSTD
2	0.681	5.292
3	0.979	1.581
4	0.997	0.707
5	0.999	0.707

[4]Note that a zero change in the centroid of the cluster seeds for the second iteration implies that the reallocation did not result in any reassignment of observations.

Interpreting the Cluster Solution

The cluster solution can be labeled or profiled using the centroids of each cluster. For example, using the cluster means [5], cluster 1 consists of subjects who have low education and low income and therefore this cluster can be labeled as low-education, low-income cluster. Similarly, cluster 2 can be labeled as high-education, high-income cluster, and cluster 3 as medium-education, medium-income cluster.

7.9 WHICH CLUSTERING METHOD IS BEST?

Which of the two types of clustering techniques (i.e., hierarchical and nonhierarchical) should one use? Then, given that the researcher selects one of these clustering techniques, which particular method or algorithm for a given clustering technique (e.g., centroid or nearest neighbor for hierarchical method) should one select? Obviously, the decision depends on the objective of the study and the properties of the various clustering algorithms. Punj and Stewart (1983) have provided comprehensive summaries of the various clustering algorithms and the empirical studies which have compared those algorithms.[5] These summaries are reproduced in Exhibit 7.3. In the following sections we briefly discuss some of the main properties of hierarchical and nonhierarchical clustering algorithms, which could help us in selecting from the various clustering algorithms.

7.9.1 Hierarchical Methods

Hierarchical clustering methods do not require *a priori* knowledge of the number of clusters or the starting partition. This is a definite advantage over nonhierarchical methods. However, hierarchical methods have the disadvantage that once an observation is assigned to a cluster it cannot be reassigned to another cluster. Therefore, hierarchical methods are sometimes used in an exploratory sense and the resulting solution is submitted to a nonhierarchical method to further refine the cluster solution. That is, hierarchical and nonhierarchical methods could be viewed as complementary clustering methods rather than as competing methods.

Since the various hierarchical methods differ with respect to the procedure used for computing the intercluster distance, the obvious question is: Which hierarchical method should one use? Unfortunately, there is no clear-cut answer. It depends on the data, the amount of noise or outliers present in the data, and the nature of groups present in the data (of which we are unaware). However, based on results of simulation studies and applications of these techniques, the following points can be made about the various techniques:

1. Hierarchical methods are susceptible to a chaining or linking effect. That is, observations are sometimes assigned to existing clusters rather than being grouped in new clusters. This is more of a problem if chaining starts early in the clustering process. In general, the nearest neighbor is more susceptible to this problem than the complete-linkage technique. However, chaining sometimes becomes an

[5]See Milligan (1980, 1981, and 1985) for a summary of simulation studies that have compared various clustering algorithms.

Exhibit 7.3

Empirical comparisons of the performance of clustering algorithms

Reference	Methods Examined	Data Sets Employed	Coverage[a]	Criteria	Summary of Results
Cunningham and Ogilvie (1972)	Single, complete, average linkage with euclidean distances and Ward's minimum variance technique	Normal mixtures	Complete	Measures of "stress" to compare input similarity/dissimilarity matrix with similarity relationship among entities portrayed by the clustering method	Average linkage outperformed other methods
Kuiper and Fisher (1975)	Single, complete, average, centroid, median linkage, all using euclidean distances and Ward's minimum variance technique	Bivariate normal mixtures	Complete	Rand's statistic (Rand 1971)	Ward's technique consistently outperformed other methods
Blashfield (1976)	Single, complete, average linkage, all using euclidean distance and Ward's minimum variance technique	Multinormal mixtures	Complete	Kappa (Cohen 1960)	Ward's technique demonstrated highest median accuracy
Mojena (1977)	Simple average, weighted average, median, centroid, complete linkage, all using Euclidean distances and Ward's minimum variance technique	Multivariate gamma distribution mixtures	Complete	Rand's statistic	Ward's method outperformed other methods

Blashfield (1977)	Eight iterative partitioning methods: Anderberg and CLUSTAN K-means methods, each with cluster statistics updated after each reassignment and only after a complete pass through the data; CLUS and MIKCA (both hill-climbing algorithms), each with optimization of tr **W** and **W**	Multinormal mixtures	Complete	Kappa	For 15 of the 20 data sets examined, a hill-climbing technique which optimized **W** performed best, i.e., MIKCA or CLUS. In two other cases a hill-climbing method which optimized tr **W** performed best, CLUS
Milligan and Isaac (1978)	Single, complete average linkage, and Ward's minimum variance technique, all using Euclidean distances	Data sets differing in degree of error perturbation	Complete	Rand's statistic and kappa	Average linkage and Ward's technique superior to single and complete linkage
Mezzich (1978)	Single, complete linkage, and K-means, each with city-block and Euclidean distances and correlation coefficient, ISO-DATA, Friedman and Rubin method, Q factor analysis, multidimensional scaling with city-block and Euclidean metrics and correlation coefficients, NORMAP/NORMIX, average linkage with correlation coefficient	Psychiatric ratings	Complete	Replicability; agreement with "expert" judges; goodness of fit between raw input dissimilarity matrix and matrix of 0's and 1's indicating entities clustered together	K-means procedure with Euclidean distances performed best followed by K-means procedure with the city-block metric; average linkage also performed well as did complete linkage with a correlation coefficient & city-block metric & ISO-DATA; the type of metric used (r, city-block, or Euclidean distance) had little impact on results

(continued)

Exhibit 7.3 (continued)

Reference	Methods Examined	Data Sets Employed	Coverage[a]	Criteria	Summary of Results
Edelbrock (1979)	Single, complete, average, and centroid, each with correlation coefficients, Euclidean distances, and Ward's minimum variance technique	Multivariate normal mixtures, standardized & unstandardized	70, 80, 90, 95, 100%	Kappa	Ward's method and simple average were most accurate; performance of all algorithms deteriorated as coverage increased but this was less pronounced when the data were standardized or correlation coefficients were used. The latter finding is suggested to result from the decreased extremity of outliers associated with standardization or use of the correlation coefficient
Edelbrock and McLaughlin (1980)	Single, complete, average, each with correlation coefficients, Euclidean distances, one-way and two-way intraclass correlations, and Ward's minimum variance technique	Multivariate normal mixtures & multivariate gamma mixtures	40, 50, 60, 70, 80, 90, 95, 100%	Kappa and Rand's statistic	Ward's method and the average method using one-way intraclass correlations were most accurate; performance of all algorithms deteriorated as coverage increased
Blashfield and Morey (1980)	Ward's minimum variance technique, group average linkage, *Q* factor analysis, Lorr's nonhierarchical procedure, all using Pearson product moment correlations as the similarity measure	Multivariate normal mixtures	Varying levels	Kappa	Group average method best at higher levels of coverage; at lower levels of coverage Ward's method and group average performed similarly

Milligan (1980)	Single, complete, group average, weighted average, centroid & median linkage, Ward's minimum variance technique, minimum average sum of squares, minimum total sum of squares, beta-flexible (Lance & Williams 1970a, b), average link in the new cluster, MacQueen's method, Jancey's method, *K*-means with random starting point, *K*-means with derived starting point, all with Euclidean distances, Cattell's (1949) r_p, and Pearson r	Multivariate normal mixtures, standardized and varying in the number of underlying clusters and the pattern of distribution of points of the clusters. Data sets ranged from error free to two levels of error perturbations of the distance measures, from containing no outliers to two levels of outlier conditions, and from no variables unrelated to the clusters to one or two randomly assigned dimensions unrelated to the underlying clusters	Complete	Rand's statistic; the point biserial correlation between the raw input dissimilarity matrix and a matrix of 0's and 1's indicating entities clustering together	*K*-means procedure with a derived point generally performed better than other methods across all conditions. 1. Distance measure selection did not appear critical; methods generally robust across distance measures. 2. Presence of random dimensions produced decrements in cluster recovery. 3. Single linkage method strongly affected by error-perturbations; other hierarchical methods moderately so; nonhierarchical methods only slightly affected by perturbations. 4. Complete linkage and Ward's method exhibited noticeable decrements in performance in the outlier conditions; single, group average, & centroid methods only slightly affected by presence of outliers; nonhierarchical methods generally unaffected by presence of outliers. 5. Group average method best among hierarchical methods used to derive starting point for *K*-means procedure. 6. Nonhierarchical methods using random starting points performed poorly across all conditions

(continued)

Exhibit 7.3 (continued)

Reference	Methods Examined	Data Sets Employed	Coverage[a]	Criteria	Summary of Results
Bayne, Beauchamp, Begovich, and Kane (1980)	Single, complete, centroid, simple average, weighted average, median linkage and Ward's minimum variance technique, and two new hierarchical methods, the variance and rank score methods; four hierarchical methods: Wolfe's NORMIX, *K*-means, two variants of the Friedman-Rubin procedure (trace **W** & |**W**|). Euclidean distances served as similarity measure	Six parameterizations of two bivariate normal populations	Complete	Rand's statistic	*K*-means, trace **W**, and |**W**| provided the best recovery of cluster structure. NORMIX performed most poorly. Among hierarchical methods, Ward's technique, complete linkage, variance & rank score methods performed best. Variants of average linkage method also performed well but not as well as other methods. Single linkage performed poorly

[a]The percentage of observations included in the cluster solution. With complete coverage, clustering continues until all observations have been assigned to a cluster. Ninety percent coverage could imply that the most extreme 10 percent of the observations were not included in any cluster.

Source: Punj, Girish, and David W. Stewart (1983), "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research*, 20 (May), 134–148.

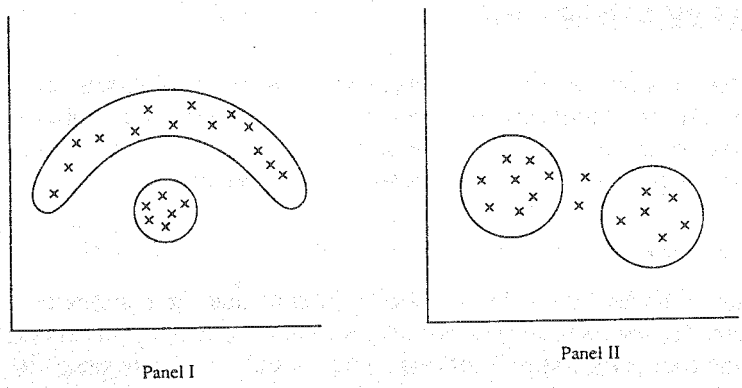


Figure 7.4 Hypothetical cluster configurations. Panel I, Nonhomogeneous clusters. Panel II, Effect of outliers on single-linkage clustering.

advantage for identifying nonhomogeneous clusters such as those depicted in Figure 7.4, Panel I.

2. Compared to the single-linkage method, the complete-linkage method (farthest-neighbor method) is less affected by the presence of noise or outliers in the data. For example, the single-linkage method will tend to join the two different clusters shown in Panel II of Figure 7.4, whereas the complete-linkage method will not.
3. The farthest-neighbor technique typically identifies compact clusters in which the observations are very similar to each other.
4. The Ward's method tends to find clusters that are compact and nearly of equal size and shape.

In order to apply the above rules to identify the best methods, knowledge of the spatial dispersion of data is required, and this is normally not known. Therefore, it is recommended that one use the various methods, compare the results for consistency, and use the method that results in an interpretable solution.

7.9.2 Nonhierarchical Methods

As discussed previously, nonhierarchical clustering techniques require knowledge about the number of clusters. Consequently, the cluster centers or the initial partition has to be identified before the technique can proceed to cluster observations. The nonhierarchical clustering algorithms, in general, are very sensitive to the initial partition. It should be further noted that since a number of starting partitions can be used, the final solution could result in local optimization of the objective function. As was evident from the previous examples, the two initial partitioning algorithms gave different cluster solutions. Results of simulation studies have shown that the K-mean algorithm and other nonhierarchical clustering algorithms perform poorly when random initial partitions are used. However, their performance is much superior when the results from hierarchical methods are used to form the initial partition. Therefore, it is recommended that for nonhierarchical clustering methods one should use an *a priori* initial partition or cluster solution. In other words, hierarchical and nonhierarchical techniques should be viewed as complementary clustering techniques rather than as competing techniques. The Appendix gives the SAS commands for first doing a hierarchical clustering, and then refining the solution using a nonhierarchical clustering technique.

7.10 SIMILARITY MEASURES

All clustering algorithms require some type of a measure to assess the similarity of a pair of observations or clusters. Similarity measures can be classified into the following three types: (1) distance measures, (2) association coefficients, and (3) correlation coefficients. In the following section we discuss these similarity measures.

7.10.1 Distance Measures

Section 3.2 of Chapter 3 discusses various distance measures and their properties. These distance measures are reviewed here briefly in the context of their use in cluster analysis. In general, the euclidean distance between points i and j in p dimensions is given by

$$D_{ij} = \left(\sum_{k=1}^{p} (X_{ik} - X_{jk})^2 \right)^{1/2},$$

where D_{ij} is the distance between observations i and j, and p is the number of variables. Euclidean distance is a special case of a more general metric called the Minkowski metric and is given by

$$D_{ij} = \left(\sum_{k=1}^{p} (|X_{ik} - X_{jk}|)^n \right)^{1/n}, \tag{7.4}$$

where D_{ij} is the Minkowski distance between observations i and j, p is the number of variables, and $n = 1, 2, \ldots, \infty$. As can be seen from Eq. 7.4, a value of 2 for n gives the euclidean distance and a value of $n = 1$ results in what is called a city-block or Manhattan distance. For example, in Figure 7.5 the city-block distance between points i and j is given by $a + b$. As the name implies, the city-block distance is the path one would normally take in a city to get from point i to point j. In general, the city-block distance is given by

$$D_{ij} = \sum_{k=1}^{p} |X_{ik} - X_{jk}|.$$

Other values of n in Eq. 7.4 result in other types of distances; however, they are not used commonly. Distance measures of similarity are based on the concept of a metric whose properties were discussed in Section 3.2 of Chapter 3. As mentioned previously, euclidean distance is the most widely used measure of similarity; however, it is not scale invariant. That is, distances between observations could change with a change in scale.

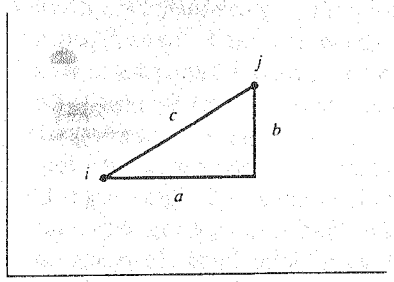


Figure 7.5 City-block distance.

Consider the data given in Table 7.1. Let us assume that income is measured in dollars instead of thousands of dollars. Squared euclidean distance between observations 1 and 2 is then given by

$$\begin{aligned} D_{12}^2 &= (5000 - 6000)^2 + (5 - 6)^2 \\ &= 100000 + 1 = 100001. \end{aligned}$$

As can be seen, the income variable dominates the computation of the distance. Thus, the scale used to measure observations can have a substantial effect on distances. Clearly, it is important to have variables that are measured on a comparable scale. However, if they cannot be measured on comparable scales then one can use the statistical distance, which has the advantageous property of scale invariance; that is, as will be seen below, the distance measure is not affected by a change in scale. The two measures of statistical distance discussed in Chapter 3 are the euclidean distance for standardized data (i.e., statistical distance) and the Mahalanobis distance. Each of these is discussed below.

Euclidean Distance for Standardized Data

Let us compute the squared euclidean distance between observations S1 and S2 for the hypothesized data given in Table 7.1 after it has been standardized. That is,

$$\begin{aligned} S\,D_{12}^2 &= \left[\left(\frac{5 - 16.167}{9.988}\right) + \left(\frac{6 - 16.167}{9.988}\right)\right]^2 + \left[\left(\frac{5 - 13.167}{6.369}\right) + \left(\frac{6 - 13.167}{6.369}\right)\right]^2 \\ &= \left(\frac{5 - 6}{9.988}\right)^2 + \left(\frac{5 - 6}{6.369}\right)^2 \\ &= 0.010 + 0.025 = 0.035. \end{aligned}$$

Note that, compared to unstandardized data the squared euclidean distance for standardized data (i.e., the statistical distance) is weighted by $1/\hat{s}_i^2$ where $\hat{s}_i$ is the standard deviation of variable i. In other words, a variable with a large variance is given a smaller weight than a variable with a smaller variance. That is, the greater the variance the less the weight and vice versa. The critical question, then, is: Why should variance be a factor for determining the importance assigned to a given variable for determining the euclidean distance? If there is a strong rationale for doing so, one can use standardized data. If not, the data should not be standardized. The important point to remember is that standardization can affect the cluster solution.

A useful property of the euclidean distance for standardized data is that it is scale invariant. For example, suppose that we change the scale for income to dollars instead of thousands of dollars. A change in scale also changes the standard deviation of the income variable to 9.988×1000. The euclidean distance between observations 1 and 2 for standardized data, when the scale of the income variable is changed, is

$$\begin{aligned} S\,D_{12}^2 &= \left(\frac{5000 - 6000}{9.988 \times 1000}\right)^2 + \left(\frac{5 - 6}{6.369}\right)^2 \\ &= \left(\frac{5 - 6}{9.988}\right)^2 + \left(\frac{5 - 6}{6.369}\right)^2 \\ &= 0.010 + 0.025 \\ &= 0.035, \end{aligned}$$

which is the same as before.

Mahalanobis Distance

The second measure of statistical distance is the Mahalanobis distance. It is designed to take into account the correlation among the variables and is also scale invariant. For uncorrelated variables the Mahalanobis distance reduces to the euclidean distance for unstandardized data. That is, euclidean distance for standardized data is a special case of Mahalanobis distance. For a two-variable case, Mahalanobis distance between observations i and j is given by Eq. 3.9 of Chapter 3 and for more than two variables by Eq. 3.10 of Chapter 3. The clustering routines in **SAS** and **SPSS** do not have the option of using the Mahalanobis distance. However, clustering routines in the **BIOMED** (1990) package do have the option of using the Mahalanobis distance. Once again, it can be shown that the Mahalanobis distance is also scale invariant.

Association Coefficients

This type of measure is used to represent similarity for binary variables. For binary data one can use such measures as polychoric correlation or simple matching coefficients or its variations to represent the similarity between observations. Consider the following 2×2 table for two binary variables:

	1	0
1	a	b
0	c	d

where a, b, c, and d are frequencies of occurrences. The similarity between the two variables is given by

$$\frac{a+d}{a+b+c+d}.$$

There are other variations of the above measure such as the Jackards coefficient. For a detailed discussion of these and other measures see Sneath and Sokal (1973) and Hartigan (1975). Association coefficients, however, do not satisfy some of the properties of a true metric discussed in Chapter 3.

Correlation Coefficients

One can also use the Pearson product moment correlation coefficient as a measure of similarity. Strictly speaking, correlation coefficients and association coefficients are dissimilarity measures; that is, a high value represents similarity and vice versa. Correlation coefficients can easily be converted into similarity measures by subtracting them from one; however, they do not satisfy some of the properties of a true metric.

It should be noted that these are not the only measures one can use to cluster observations. One can use any measure of similarity between objects that is meaningful to the researcher. For example, in locating bank branches a meaningful measure of similarity between two potential locations might be the driving time, and not the distances in miles or kilometers. Or in the case of image-related research, perceptual distances or similarities might be more meaningful than euclidean distances. In conclusion, one should choose that measure of similarity which is consistent with the objective of the study. Suffice it to say that different measures of distance could result in different cluster configurations.

7.11 RELIABILITY AND EXTERNAL VALIDITY OF A CLUSTER SOLUTION

Cluster analysis is a heuristic technique; therefore a clustering solution or grouping will result even when there may not be any natural groups or clusters in the data. Thus, establishing the reliability and external validity of a cluster solution is all the more important.

7.11.1 Reliability

Reliability can be established by a cross-validation procedure suggested by McIntyre and Blashfield (1980). The data set is first split into two halves. Cluster analysis is done on the first half of the sample and the cluster centroids are identified. Observations in the second half of the sample are assigned to the cluster centroid that has the smallest euclidean distance. Degree of agreement between the assignment of the observations and a separate cluster analysis of the second sample is an indicator of reliability. The procedure can be repeated by performing cluster analysis on the second sample and by assigning observations in the first sample and computing the degree of agreement between the assignment and cluster analysis on the first half.

7.11.2 External Validity

External validity is obtained by comparing the results from cluster analysis with an external criterion. For example, suppose we cluster firms based on certain financial ratios and thereby obtain two clusters: firms that are financially healthy and firms that are not financially healthy. External validity can then be established by correlating the results of cluster analysis with classification obtained by independent evaluators (e.g., auditors, financial analysts, stockbrokers, industry analysts).

7.12 AN ILLUSTRATIVE EXAMPLE

In this section we illustrate the use of cluster analysis using the food nutrient data given in Table 7.19. First, we cluster the observations using the single-linkage, complete-linkage, centroid, and the Ward's methods. Multiple methods are used to determine if different methods produce similar cluster solutions. This is followed by a nonhierarchical clustering technique. The "best" solution(s) obtained from hierarchical procedure will be used as the starting or initial solutions.

7.12.1 Hierarchical Clustering Results

Exhibit 7.4 contains partial outputs for the centroid, single-linkage, complete-linkage, and Ward's methods. The dendrograms are not included as they are quite cumbersome to interpret. Figure 7.6 gives plots for the (a) RS and (b) RMSSTD criteria, which can be used for assessing the cluster solution and for determining the number of clusters. Recall that we are looking for a "big" change or an elbow in the plot of a given criterion against number of clusters. The obvious elbows in the plots have been labeled; however, visual identification of the elbows is a subjective process and could vary across researchers.

Table 7.19 **Food Nutrient Data**

Food Item	Calories	Protein	Fat	Calcium	Iron
Braised beef	340	20	28	9	2.6
Hamburger	245	21	17	9	2.7
Roast beef	420	15	39	7	2.0
Beef steak	375	19	32	9	2.6
Canned beef	180	22	10	17	3.7
Broiled chicken	115	20	3	8	1.4
Canned chicken	170	25	7	12	1.5
Beef heart	160	26	5	14	5.9
Roast lamb leg	265	20	20	9	2.6
Roast lamb shoulder	300	18	25	9	2.3
Smoked ham	340	20	28	9	2.5
Roast pork	340	19	29	9	2.5
Simmered pork	355	19	30	9	2.4
Beef tongue	205	18	14	7	2.5
Veal cutlet	185	23	9	9	2.7
Baked bluefish	135	22	4	25	0.6
Raw clams	70	11	1	82	6.0
Canned clams	45	7	1	74	5.4
Canned crabmeat	90	14	2	38	0.8
Fried haddock	135	16	5	15	0.5
Broiled mackerel	200	19	13	5	1.0
Canned mackerel	155	16	9	157	1.8
Fried perch	195	16	11	14	1.3
Canned salmon	120	17	5	159	0.7
Canned sardines	180	22	9	367	2.5
Canned tuna	170	25	7	7	1.2
Canned shrimp	110	23	1	98	2.6

Source: The Yearbook of Agriculture 1959 (The U.S. Department of Agriculture, Washington D.C.), p. 244.

It is evident from the plots that most of the criteria for the complete-linkage, centroid, and Ward's method suggest that there are four clusters, although there is some evidence that there might be only three clusters. In addition, all the criteria indicate a reasonably good cluster solution. The plots for the single-linkage method are interesting. They suggest a seven-cluster or a four-cluster solution. A final decision on the number of clusters that should be retained can be made by further examining the membership of the clusters formed by the four methods. Table 7.20 gives the cluster membership of each food item for each of the four methods. Except for the single-linkage method, the methods produce an almost similar cluster solution. The four-cluster solutions for the Ward's and complete-linkage methods are the same and differ only slightly from those of the centroid method. Cluster 4 for the complete-linkage, Ward's, and centroid methods and cluster 7 for the single-linkage method contain a single observation (i.e., canned sardines). Clusters 5, 6, and 7 of the seven-cluster solution resulting from the single-linkage method also consist of one member each, and cluster 4 consists of only two members. The membership of the remaining clusters (clusters 1, 2, and 3) is very similar to that of the other three methods. It appears that there are four clusters; however, the four-cluster solution obtained from the single-linkage method is quite different from the other three methods. Table 7.21 gives the centroids or the cluster centers for each clustering method.

Exhibit 7.4

Hierarchical cluster analysis for food data

SINGLE LINKAGE CLUSTER ANALYSIS

SIMPLE STATISTICS

	MEAN	STD DEV	SKEWNESS	KURTOSIS	BIMODALITY
CALORIES	207.407	101.208	0.542	-0.675	0.478
PROTEIN	19.000	4.252	-0.824	1.327	0.357
FAT	13.481	11.257	0.790	-0.624	0.589
CALCIUM	43.963	78.034	3.159	11.345	0.746
IRON	2.381	1.461	1.230	1.469	0.518

ROOT-MEAN-SQUARE TOTAL-SAMPLE STANDARD DEVIATION = 57.4096

NUMBER OF CLUSTERS	CLUSTERS JOINED		FREQUENCY OF NEW CLUSTER	RMS STD OF NEW CLUSTER	SEMIPARTIAL R-SQUARED	R-SQUARED	MINIMUM DISTANCE
10	CANNED MACKEREL	CANNED SALMON	2	11.16786	0.001455	0.973438	35.3159
9	CL14	ROAST LAMB SHOUL	3	12.59929	0.003226	0.970211	35.4131
8	CL11	CANNED CRABMEAT	12	16.80697	0.014701	0.955510	39.5267
7	CL15	CL9	8	20.48901	0.028341	0.927169	40.1627
6	CL7	CL8	20	40.04817	0.285060	0.642109	40.2746
5	CL12	CANNED SHRIMP	3	16.10565	0.005231	0.636878	44.8504
4	CL6	ROAST BEEF	21	43.49500	0.085924	0.550954	45.7642
3	CL4	CL5	24	48.72189	0.189548	0.361406	48.7139
2	CL3	CL10	26	50.53988	0.106595	0.254811	62.2624
1	CL2	CANNED SARDINES	27	57.40958	0.254811	0.000000	211.5691

(continued)

Exhibit 7.4 (continued)

COMPLETE LINKAGE CLUSTER ANALYSIS

NUMBER OF CLUSTERS	CLUSTERS JOINED		FREQUENCY OF NEW CLUSTER	RMS STD OF NEW CLUSTER	SEMIPARTIAL R-SQUARED	R-SQUARED	MAXIMUM DISTANCE
10	CL15	CANNED CRABMEAT	4	11.32324	0.003476	0.985594	50.6665
9	CL17	ROAST LAMB SHOUL	3	12.59929	0.003226	0.982367	55.6611
8	CL14	CANNED SHRIMP	3	16.10565	0.005231	0.977136	71.1677
7	CL13	ROAST BEEF	6	14.34190	0.009755	0.967381	80.9343
6	CL10	CL8	7	22.14096	0.023782	0.943599	108.1758
5	CL9	CL11	11	20.22234	0.039103	0.904496	141.7814
4	CL6	CL12	9	30.07489	0.048662	0.855835	154.4447
3	CL7	CL5	17	38.73570	0.220433	0.635402	262.5666
2	CL4	CANNED SARDINES	10	51.36181	0.192623	0.442779	364.8934
1	CL3	CL2	27	57.40958	0.442779	0.000000	433.7617

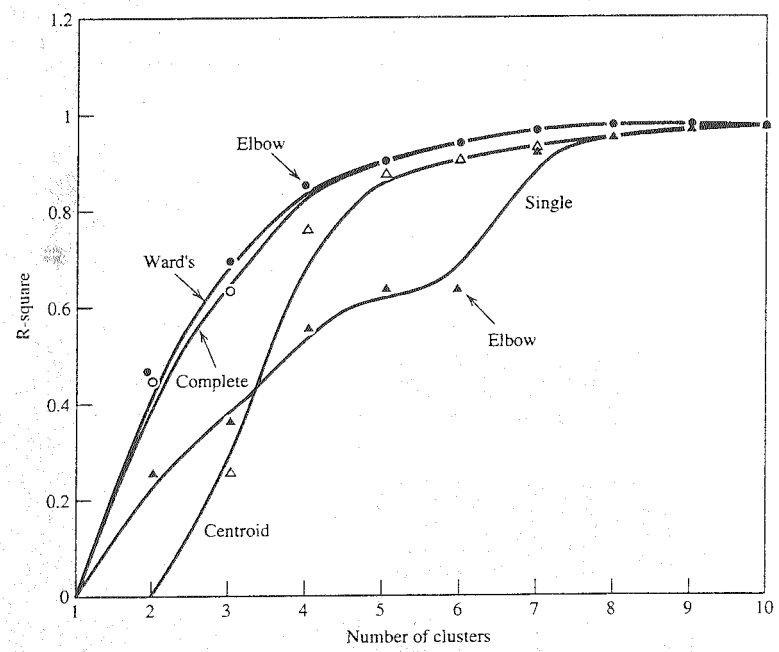
CENTROID HIERARCHICAL CLUSTER ANALYSIS

NUMBER OF CLUSTERS	CLUSTERS JOINED		FREQUENCY OF NEW CLUSTER	RMS STD OF NEW CLUSTER	SEMIPARTIAL R-SQUARED	R-SQUARED	CENTROID DISTANCE
10	CL15	CANNED CRABMEAT	4	11.32324	0.003476	0.985594	44.5633
9	CL16	ROAST LAMB SHOUL	3	12.59929	0.003226	0.982367	45.5370
8	CL14	CANNED SHRIMP	3	16.10565	0.005231	0.977136	57.9815

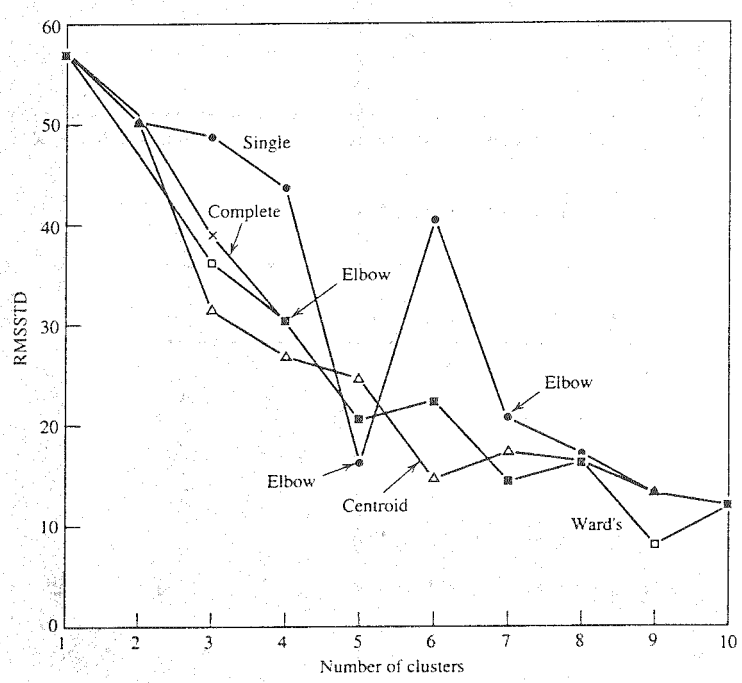
7	CL13	CL10	12	16.80697	0.026857	0.950279	65.6901
6	CL12	ROAST BEEF	6	14.34190	0.009755	0.940524	70.8222
5	CL6	CL9	9	24.36751	0.039727	0.900797	92.2533
4	CL8	CL11	5	26.85628	0.026158	0.874639	96.6423
3	CL7	CL4	17	31.36108	0.113709	0.760930	117.4906
2	CL5	CL3	26	50.53988	0.506119	0.254811	191.9655
1	CL2	CANNED SARDINES	27	57.40958	0.254811	0.000000	336.7134

WARD'S MINIMUM VARIANCE CLUSTER ANALYSIS

NUMBER OF CLUSTERS	CLUSTERS JOINED		FREQUENCY OF NEW CLUSTER	RMS STD OF NEW CLUSTER	SEMIPARTIAL R-SQUARED	R-SQUARED	BETWEEN-CLUSTER SUM OF SQUARES
10	CL14	CANNED CRABMEAT	4	11.32324	0.003476	0.985908	1489.42
9	CL16	CL20	8	7.75641	0.003541	0.982367	1517.12
8	CL15	CANNED SHRIMP	3	16.10565	0.005231	0.977136	2241.24
7	CL12	ROAST BEEF	6	14.34190	0.009755	0.967381	4179.83
6	CL10	CL8	7	22.14096	0.023782	0.943599	10189.5
5	CL11	CL9	11	20.22234	0.039103	0.904496	16754.1
4	CL6	CL13	9	30.07489	0.048662	0.855835	20849.7
3	CL5	CL4	20	36.22080	0.158726	0.697109	68007.8
2	CL3	CANNED SARDINES	21	47.72546	0.240715	0.456394	103137
1	CL7	CL2	27	57.40958	0.456394	0.000000	195548



(a)



(b)

Figure 7.6 Cluster analysis plots. (*a*) R-square. (*b*) RMSSTD.

Table 7.20 **Cluster Membership for the Four-Cluster Solution**[a]

Food Item	Hierarchical Clustering Methods			
	Complete Linkage	Ward's	Centroid	Single Linkage
Braised beef	1	1	1	1(1)
Hamburger	2	2	1	1(1)
Roast beef	1	1	1	1(6)
Beef steak	1	1	1	1(1)
Canned beef	2	2	2	1(2)
Broiled chicken	3	3	2	1(2)
Canned chicken	2	2	2	1(2)
Beef heart	2	2	2	1(2)
Roast lamb leg	2	2	1	1(1)
Roast lamb shoulder	2	2	1	1(1)
Smoked ham	1	1	1	1(1)
Roast pork	1	1	1	1(1)
Simmered pork	1	1	1	1(1)
Beef tongue	2	2	2	1(2)
Veal cutlet	2	2	2	1(2)
Baked bluefish	3	3	2	1(2)
Raw clams	3	3	3	2(3)
Canned clams	3	3	3	2(3)
Canned crabmeat	3	3	2	1(2)
Fried haddock	3	3	2	1(2)
Broiled mackerel	2	2	2	1(2)
Canned mackerel	3	3	3	3(4)
Fried perch	2	2	2	1(2)
Canned salmon	3	3	3	3(4)
Canned sardines	4	4	4	4(7)
Canned tuna	2	2	2	1(2)
Canned shrimp	3	3	3	2(5)

[a]Numbers in parentheses for the single-linkage method are cluster membership for a seven-cluster solution.

Table 7.21 **Cluster Centers for Hierarchical Clustering of Food Nutrient Data**

Cluster Number	Size	Calories	Proteins	Fat	Calcium	Iron
(a) Cluster Centers for the Complete-Linkage and Ward's Method						
1	6	361.667	18.667	31.000	8.667	2.433
2	11	206.818	21.182	12.546	10.182	2.491
3	9	108.333	16.222	3.444	72.889	2.200
4	1	180.000	22.000	9.000	367.000	2.500
(b) Cluster Centers for the Centroid Method						
1	9	331.111	19.000	27.556	8.778	2.467
2	12	161.667	20.500	7.500	14.250	1.925
3	5	100.000	14.800	3.400	114.000	3.000
4	1	180.000	22.000	9.000	367.000	2.500
(c) Cluster Centers for the Single-Linkage Method						
1	21	234.286	19.857	16.095	11.905	2.157
2	3	75.000	13.667	1.000	84.667	4.667
3	2	137.500	16.500	7.000	158.000	1.250
4	1	180.000	22.000	9.000	367.000	2.500

7.12.2 Nonhierarchical Clustering Results

As mentioned previously, it is recommended that hierarchical cluster analysis be followed by nonhierarchical clustering. That is, nonhierarchical clustering is used to refine the clustering solution obtained from the hierarchical method. For illustration purposes, the FASTCLUS procedure in SAS is used. Because the various hierarchical methods resulted in different solutions, each of these solutions will be refined by nonhierarchical clustering. The cluster means given in Table 7.21 were used as the initial or starting seeds. Note that the fourth cluster of each solution consists of only one observation, canned sardines, which is clearly an outlier because of its high calcium content. Consequently, this food item was deleted from further analysis and therefore we have essentially three clusters. The final nonhierarchical solutions differed only slightly when cluster centroids from various hierarchical methods were used as the initial seeds; these differences are discussed later. Table 7.22 gives the SAS commands. Exhibit 7.5 gives the resulting output when the cluster means from the centroid method of the hierarchical clustering algorithm were used as the initial or starting seeds. As before, the output is labeled to facilitate the discussion.

Initial Solution and Final Solution

The initial cluster centers or seeds are printed [1]. Note that the cluster seeds are the same as those reported in Table 7.21. The iteration history for reassignment is also printed [2]. A total of three iterations or reassignments were required for the cluster solution to converge. In the final cluster solution, clusters 1, 2, and 3, respectively, consist of 8, 12, and 6 members [3a]. A list of the members comprising each cluster is given at the end of the output [6].

Evaluating the Cluster Solution

For a good cluster solution, each cluster should be as homogeneous as possible and the various clusters should be as heterogeneous as possible. The three clusters appear to be well separated as the distance between the centroids of the clusters is quite large. For example, the nearest cluster to cluster 1 is cluster 2 [3d], and the distance between

Table 7.22 **Commands for FASTCLUS Procedure**

```
OPTIONS NOCENTER;
TITLE FASTCLUS ON FOOD NUTRIENT DATA GIVEN IN TABLE 7.19;
DATA INITIAL;
INPUT CALORIES PROTEIN FAT CALCIUM IRON;
CARDS;
 insert initial cluster seeds
DATA FOOD;
INPUT CODE $ 1-2 NAME $ 1-25 CALORIES 27-29 PROTEIN 31-32 FAT 34-35
   CALCIUM 37-39 IRON 41-43;
CARDS;
insert data
PROC FASTCLUS SEED=INITIAL MAXCLUSTERS=3 OUT=NEW MAXITER=20;
   VAR CALORIES FAT PROTEIN CALCIUM IRON;
PROC SORT;BY CLUSTER;
PROC PRINT;BY CLUSTER;
   VAR NAME CLUSTER DISTANCE CALORIES PROTEIN FAT CALCIUM IRON;
```

Exhibit 7.5

Nonhierarchical analysis for food-nutrient data

(1) INITIAL SEEDS

CLUSTER	CALORIES	PROTEIN	FAT	CALCIUM	IRON
1	331.111	19.000	27.556	8.778	2.467
2	161.667	20.500	7.500	14.250	1.925
3	100.000	14.800	3.400	114.000	3.000

(2) MINIMUM DISTANCE BETWEEN SEEDS = 117.4876

ITERATION	CHANGE IN CLUSTER SEEDS 1	2	3
1	10.8475	6.46446	0.3
2	0	6.85281	12.7855
3	0	0	0

CLUSTER SUMMARY

CLUSTER NUMBER	(3a) FREQUENCY	(3b) RMS STD DEVIATION	(3c) MAXIMUM DISTANCE FROM SEED TO OBSERVATION	(3d) NEAREST CLUSTER	(3e) CENTROID DISTANCE
1	8	20.8936	78.8882	2	168.5
2	12	16.3651	70.9576	3	117.9
3	6	27.8059	79.6672	2	117.9

STATISTICS FOR VARIABLES

VARIABLE	(4a) TOTAL STD	(4b) WITHIN STD	(4c) R-SQUARED	RSQ/(1-RSQ)
CALORIES	103.06085	39.89286	0.86216	6.25453
PROTEIN	4.29257	3.58590	0.35798	0.55758
FAT	11.44357	4.52989	0.85584	5.93681
CALCIUM	44.70188	22.76009	0.76150	3.19291
IRON	1.49005	1.51663	0.04688	0.04919
OVER-ALL	50.53988	20.71299	0.84547	5.47135

(4d) PSEUDO F STATISTIC = 62.92
APPROXIMATE EXPECTED OVER-ALL R-SQUARED = 0.78678
CUBIC CLUSTERING CRITERION = 2.186
WARNING: THE TWO ABOVE VALUES ARE INVALID FOR CORRELATED VARIABLES

(5) CLUSTER MEANS

CLUSTER	CALORIES	PROTEIN	FAT	CALCIUM	IRON
1	341.875	18.750	28.875	8.750	2.437
2	174.583	21.083	8.750	11.833	2.083
3	98.333	14.667	3.167	101.333	2.883

(continued)

Exhibit 7.5 (continued)

6 CLUSTER=1

OBS	NAME	CLUSTER	DISTANCE	CALORIES	PROTEIN	FAT	CALCIUM	IRON
1	BRAISED BEEF	1	2.4357	340	20	28	9	2.6
2	ROAST BEEF	1	78.8882	420	15	39	7	2.0
3	BEEF STEAK	1	33.2744	375	19	32	9	2.6
4	ROAST LAMB LEG	1	77.3963	265	20	20	9	2.6
5	ROAST LAMB SHOULDER	1	42.0616	300	18	25	9	2.3
6	SMOKED HAM	1	2.4311	340	20	28	9	2.5
7	PORK ROAST	1	1.9132	340	19	29	9	2.5
8	PORK SIMMERED	1	13.1779	355	19	30	9	2.4

CLUSTER=2

OBS	NAME	CLUSTER	DISTANCE	CALORIES	PROTEIN	FAT	CALCIUM	IRON
9	HAMBURGER	2	70.9576	245	21	17	9	2.7
10	CANNED BEEF	2	7.8135	180	22	10	17	3.7
11	BROILED CHICKEN	2	59.9964	115	20	3	8	1.4
12	CANNED CHICKEN	2	6.3070	170	25	7	12	1.5
13	BEEF HEART	2	16.4369	160	26	5	14	5.9
14	BEEF TONGUE	2	31.3971	205	18	14	7	2.5
15	VEAL CUTLET	2	10.9841	185	23	9	9	2.7
16	BAKED BLUEFISH	2	42.0215	135	22	4	25	0.6
17	FRIED HADDOCK	2	40.2403	135	16	5	15	0.5
18	BROILED MACKEREL	2	26.7634	200	19	13	5	1.0
19	FRIED PERCH	2	21.2850	195	16	11	14	1.3
20	CANNED TUNA	2	7.9719	170	25	7	7	1.2

CLUSTER=3

OBS	NAME	CLUSTER	DISTANCE	CALORIES	PROTEIN	FAT	CALCIUM	IRON
21	RAW CLAMS	3	34.7046	70	11	1	82	6.0
22	CANNED CLAMS	3	60.5092	45	7	1	74	5.4
23	CANNED CRABMEAT	3	63.9273	90	14	2	38	0.8
24	CANNED MACKEREL	3	79.6672	155	16	9	157	1.8
25	CANNED SALMON	3	61.7127	120	17	5	159	0.7
26	CANNED SHRIMP	3	14.8809	110	23	1	98	2.6

the centroids of these two clusters is 168.5 [3e]. A high overall value of 0.845 for RS further confirms this conclusion [4c]. As discussed previously, a high value for RS indicates that the clusters are well separated and consequently the clusters are quite homogeneous. The RMSSTD of the clusters suggests that, relatively, cluster 2 is more homogeneous than the other two clusters [3b]. Overall, it appears that the cluster solution is reasonable.

The cluster solution can also be evaluated with respect to each clustering variable. As discussed earlier, the RMSSTD can be compared across variables only if the measurement scales are the same. If the measurement scales are not the same, then for each variable one should obtain the ratio of the respective within-group RMSSTD [4b] to the total RMSSTD [4a], and compare this ratio across the variables. For example, the ratio

for calories is equal to .387 (39.893/103.061). The ratios for protein, fat, calcium, and iron, respectively, are .835, .396, .509, and 1.018, suggesting that the clusters are more homogeneous with respect to calories, fat, and calcium than protein and iron.

The reported RS for each variable can be used to assess differences among the clusters with respect to each variable [4c]. RS values of 0.358 and 0.047, respectively, for protein and iron suggest that the clusters are not different from each other with respect to these two variables. Previously it was seen that the clusters also were not homogeneous with respect to these two variables. This suggests that protein and iron may not be appropriate for forming clusters. One may want to repeat the analysis after deleting these two variables. The expected overall RS and the cubic clustering criterion are not interpreted because these statistics are not very meaningful for highly correlated variables [4d].

As mentioned previously there were only slight differences in the solutions obtained when the centroids from the single-linkage, complete-linkage, and Ward's methods were used as initial seeds or starting points. Specifically, there was no difference in the nonhierarchical solution when the centroids for the single-linkage and the centroid methods were used as starting points. When the centroids from the Ward's and complete-linkage methods were used, there was only one difference in the resulting nonhierarchical solution: roast lamb leg was a member of cluster 2 instead of cluster 1 (see [6] of Exhibit 7.5). Thus, it can be seen that although some of the hierarchical clustering methods gave different results, the nonhierarchical clustering method gave very similar results when the result from each of the hierarchical methods was used as the starting solution. This suggests the importance or need for further refining the cluster solution from a hierarchical method.

Interpreting the Clusters

The next question is: What do the clusters represent? That is, can we give a name to the clusters? Answering this requires knowledge of the subject area, in this case, nutrition. From the cluster means [5], it can be clearly seen that the clusters differ mainly with respect to calories, fat, and calcium. Consequently, these three nutrients are used to obtain the following description of the three clusters:

- Cluster one is high in calories and fat, and low in calcium. One could call this a "high-fat food group."
- The second cluster is low in calories and fat, and is also low in calcium. It can be labeled as "medium-fat food group."
- Cluster three is very low in calories and fat, and high in calcium, and could be labeled as "low-fat, high-calcium food group."

External Validation of the Cluster Solution and Additional Analysis

The final part of the output gives the cluster membership [6]. The clustering solution was externally validated by showing the clustering solution to a registered dietitian who indicated that the food item clusters were as expected.

Note that one can repeat the above analysis in a number of ways. First, fat and calories are related and, therefore, they are bound to be correlated.[6] Table 7.23 gives the

[6]This was pointed out by the registered dietitian to whom these results were shown for an independent evaluation.

Table 7.23 **Correlation Matrix**

	Calories	Protein	Fat	Calcium	Iron
Calories	1.000				
Protein	.179	1.000			
Fat	.987	.025	1.000	−.308	−.056
Calcium	−.320	−.085	−.308	1.000	.043
Iron	−.096	−.175	−.056	.043	1.000

correlation matrix among the variables. The correlation between calories and fat is high. The following two approaches can be taken when the variables are correlated:

1. The data can be subjected to a principal components analysis. Then either the principal components scores or a representative variable from each component can be taken and these representative variables used for performing cluster analysis. In our example, a principal components analysis of the food nutrient data resulted in a total of four principal components. After the components were varimax rotated, calories and fat, as expected, loaded high on the first component, and protein, calcium, and iron loaded, respectively, on the second, third, and fourth components. A reanalysis of the data was done by choosing calories as the representative variable from the first component. There was no change in the resulting cluster solution.
2. If the variables are correlated among themselves, one can use the Mahalanobis distance. The clustering algorithms in SAS and SPSS do not have the option of using Mahalanobis distance.[7]

Second, one can argue that the data should be standardized since the variables have a different measurement scale. However, standardization of the data implicitly gives different weights to variables. Since no theoretical reason exists for assigning different weights to the variables, the data were not standardized. Finally, it could be argued that the recommended daily allowance of the various nutrients provided by the food items should be used as the clustering variables. One can repeat the analysis by using these new variables. Obviously, the food items might cluster differently when these new variables are used.

7.13 SUMMARY

In this chapter we used hypothetical data to provide a geometrical and analytical view of the basic concepts of cluster analysis. The objective of cluster analysis is to form groups such that each group is as homogeneous as possible with respect to characteristics of interest and the groups are as different as possible. In hierarchical cluster analysis, clusters are formed hierarchically such that the number of clusters at each step is $n - 1, n - 2, n - 3$, and so on. A number of different algorithms for hierarchical clustering were discussed. These algorithms differed mainly with respect to how distances between two clusters are computed.

In nonhierarchical clustering, observations are assigned to clusters to which they are the closest. Consequently, one has to know *a priori* the number of clusters present in the data set. Nonhierarchical clustering techniques also present the user with a number of different algorithms that differ mainly with respect to how the initial cluster centers are obtained and how the observations are reallocated among clusters. Simulation studies that have compared hierarchical and

[7]The Mahalanobis distance option is available in the BIOMED package.

nonhierarchical clustering algorithms have concluded that the two techniques should be viewed as complementary techniques. That is, solutions obtained from hierarchical techniques should be further refined by the nonhierarchical techniques. This approach gave the best cluster solution.

Since all clustering techniques use some sort of similarity measure, a brief discussion of the different similarity measures used in clustering was provided. Finally, an example was used to illustrate the use of cluster analysis to group food items based on their nutrient content.

This chapter concludes the coverage of interdependent multivariate techniques. The next several chapters cover the following dependent techniques: two-group and multiple-group discriminant analysis, logistic regression, multivariate analysis of variance, canonical correlation, and structural equation models.

QUESTIONS

7.1 Can cluster analysis be considered to be a data reduction technique? In what ways is the data reduction obtained by cluster analysis different from that obtained using principal components analysis? In what way is it different from exploratory factor analysis?

Use a spreadsheet or calculator to solve Questions 7.2 and 7.3.

7.2 Table Q7.1 presents price (P) and quality rating (Q) data for six brands of beer.

Table Q7.1

Brand	Price	Quality
Bud Ice Draft	7.89	10
Milwaukee's Best	4.79	4
Coor's Light	7.65	9
Miller Genuine Draft	6.39	7
Schlitz	4.50	3
Michelob	6.25	6

Notes:

1. Quality ratings were given on a 10-point scale with 1 = worst quality and 10 = best quality.
2. Prices are in dollars per 12 pack of 12 oz. cans.
3. The above data are purely hypothetical and not based on any survey.

(a) Plot the data in two-dimensional space and perform a visual clustering of the brands from the plot.
(b) Compute the similarity matrix for the six brands.
(c) Use the centroid method and the complete-linkage method to perform a hierarchical clustering of the brands. Compare your solutions to that obtained in part (a).

7.3 Six retail chain stores were evaluated by a consumer panel on 10 service quality attributes. Based on their evaluations, the following similarity matrix containing squared euclidean distances was computed:

Store #	1	2	3	4	5	6
1	0.00					
2	3.65	0.00				
3	46.81	13.29	0.00			
4	24.62	51.88	17.21	0.00		
5	39.87	26.75	8.79	16.27	0.00	
6	82.48	40.05	18.91	6.20	65.22	0.00

Use (a) single-linkage and (b) average-linkage methods to find suitable groupings of the six stores based on their service quality ratings.

7.4 Table Q7.2 presents information on three nutrients for six fish types. Use a suitable hierarchical method to cluster the fish types. Identify the number of clusters and interpret them.

Table Q7.2

Fish Type	Energy	Fat	Calcium
Mackerel	5	9	20
Perch	6	11	2
Salmon	4	5	20
Sardines	6	9	46
Tuna	5	7	1
Shrimp	3	1	12

Source: The Yearbook of Agriculture 1959 (The U.S. Department of Agriculture, Washington, D.C.), p. 244.

7.5 Refer to file MASST.DAT (for a description of the data refer to file MASST.DOC). Variables V_7 to V_{18} pertain to Question 7 from the survey (see file MASST.DOC).

Use cluster analysis to segment the respondents based on their "latent" demand for mass transportation at various gasoline prices. Describe the characteristics of each segment.

Hint: Use a suitable hierarchical method to determine the number of clusters and then use nonhierarchical clustering to refine the cluster solution.

7.6 File FIN.DAT gives the financial performance information for 25 companies from three industries: textile companies, pharmaceutical companies, and supermarket companies. The column labeled "type," indicates the industry to which the company belongs. Perform cluster analysis on the financial performance data to find suitable groupings of the companies. Describe the characteristics of these groups. Comment on the agreement/disagreement between the groups obtained by you and the industry membership of the companies.

7.7 Refer to Q5.13 (Chapter 5). Use the factor scores obtained in that question to segment respondents based on their attitudes and opinions toward nutrition. Describe the characteristics of respondents belonging to each segment.

Note: You will need to know the interpretation of the factors obtained in Q5.13 to be able to describe segment characteristics.

7.8 Refer to the food price data in file FOODP.DAT. Perform cluster analysis on the principal components scores to group the cities. How are the cities belonging to the same group similar, and how are they different from those belonging to other groups?

7.9 File SCORE.DAT gives the percentage points scored by 30 students in four courses: math, physics, English, and French. Use cluster analysis to group the students and describe the aptitudes of each group.

7.10 Discuss the advantages and disadvantages of hierarchical versus nonhierarchical clustering methods. Under what circumstances is one method more appropriate than another?

Appendix

Hierarchical and nonhierarchical clustering can be viewed as complementary techniques. That is, a hierarchical clustering method can be used to identify the number of clusters and cluster seeds, then the resulting clustering solution can be refined using a nonhierarchical clustering technique. This appendix briefly discusses the SAS commands that can be used to achieve these objectives.

The data in Table 7.1 are used for illustration purposes. It will be assumed that the data set was analyzed using various hierarchical procedures to determine which hierarchical algorithm gave the best cluster solution and the number of clusters in the data set. Table A7.1 gives the SAS commands.

The data are first subjected to hierarchical clustering and an SAS output data set TREE is obtained. In the PROC TREE procedure the NCLUSTERS option specifies that a three-cluster solution is desired. The OUT = CLUS3 option requests a new data set, CLUS3, which contains the variable CLUSTER whose value gives the membership of each observation. For example, if the third observation is in cluster 2 then the value of CLUSTER for the third observation will

Table A7.1 **Using a Nonhierarchical Clustering Technique to Refine a Hierarchical Cluster Solution**

```
OPTIONS NOCENTER;
TITLE1 HIERARCHICAL ANALYSIS FOR DATA IN TABLE 7.1;
DATA TABLE1;
 INPUT SID $ 1-2 INCOME 4-5 EDUC 7-8;
CARDS;
insert data here

*Commands for hierarchical clustering;
PROC CLUSTER NOPRINT METHOD=CENTROID
 NONORM OUT=TREE;
  ID SID;
  VAR INCOME EDUC;

*Commands for creating CLUS3 data set;
PROC TREE DATA=TREE OUT=CLUS3 NCLUSTERS=3 NOPRINT;
 ID SID;
 COPY INCOME EDUC;
PROC SORT; BY CLUSTER;
TITLE2 '3-CLUSTER SOLUTION ';

*Commands for obtaining cluster means or centroids;
PROC MEANS NOPRINT;
 BY CLUSTER;
 OUTPUT OUT=INITIAL MEAN=INCOME EDUC;
 VAR INCOME EDUC;

*Commands for non-hierarchical clustering;
PROC FASTCLUS DATA=TABLE1 SEED=INITIAL LIST DISTANCE
 MAXCLUSTERS=3 MAXITER=30;
VAR INCOME EDUC;
TITLE3 'NONHIERARCHICAL CLUSTERING';
```

be 2. The COPY command requests that the CLUS3 data set should also contain the values of INCOME and EDUC variables. The PROC MEANS command uses the CLUS3 data set to compute the means of each variable for each cluster, and this information is contained in the SAS data set INITIAL. Finally, the PROC FASTCLUS, which is the nonhierarchical clustering procedure in SAS, is employed to obtain a nonhierarchical cluster solution for the data in Table 7.1. The DATA option in the FASTCLUS procedure specifies that the data set for clustering is in the SAS data set TABLE1, and the SEED option specifies that the initial cluster seeds are in the SAS data set INITIAL.

CHAPTER 8

Two-Group Discriminant Analysis

Consider the following examples:

- The IRS is interested in identifying variables or factors that significantly differentiate between audited tax returns that resulted in underpayment of taxes and those that did not. The IRS also wants to know if it is possible to use the identified factors to form a composite index that will parsimoniously represent the differences between the two groups of tax returns. Finally, can the computed index be used to predict which future tax returns should be audited?
- A medical researcher is interested in determining factors that significantly differentiate between patients who have had a heart attack and those who have not yet had a heart attack. The medical researcher then wants to use the identified factors to predict whether a patient is likely to have a heart attack in the future.
- A criminologist is interested in determining differences between on-parole prisoners who have and who have not violated their parole, then using this information for making future parole decisions.
- The marketing manager of a consumer packaged goods firm is interested in identifying salient attributes that successfully differentiate between purchasers and non-purchasers of brands, and employing this information to predict purchase intentions of potential customers.

Each of the above examples attempts to meet the following three objectives:

1. Identify the variables that discriminate "best" between the two groups.
2. Use the identified variables or factors to develop an equation or function for computing a new variable or index that will parsimoniously represent the differences between the two groups.
3. Use the identified variables or the computed index to develop a rule to classify future observations into one of the two groups.

Discriminant analysis is one of the available techniques for achieving the preceding objectives. This chapter discusses the case of two groups. The next chapter presents the case of more than two groups.

8.1 GEOMETRIC VIEW OF DISCRIMINANT ANALYSIS

The data given in Table 8.1 are used for the discussion of the geometric approach to discriminant analysis. The table gives financial ratios for a sample of 24 firms, the 12

Table 8.1 **Financial Data for Most-Admired and Least-Admired Firms**

Group 1: Most-Admired				Group 2: Least-Admired			
Firm Number	*EBITASS*	*ROTC*	Z	Firm Number	*EBITASS*	*ROTC*	Z
1	0.158	0.182	0.240	13	−0.012	−0.031	−0.030
2	0.210	0.206	0.294	14	0.036	0.053	0.063
3	0.207	0.188	0.279	15	0.038	0.036	0.052
4	0.280	0.236	0.365	16	−0.063	−0.074	−0.097
5	0.197	0.193	0.276	17	−0.054	−0.119	−0.122
6	0.227	0.173	0.283	18	0.000	−0.005	−0.004
7	0.148	0.196	0.243	19	0.005	0.039	0.031
8	0.254	0.212	0.329	20	0.091	0.122	0.151
9	0.079	0.147	0.160	21	−0.036	−0.072	−0.076
10	0.149	0.128	0.196	22	0.045	0.064	0.077
11	0.200	0.150	0.247	23	−0.026	−0.024	−0.035
12	0.187	0.191	0.267	24	0.016	0.026	0.030

Note: Z computed using $w_1 = .707$ and $w_2 = .707$.

most-admired firms and the 12 least-admired firms. The financial ratios are: *EBITASS*, earnings before interest and taxes to total assets, and *ROTC*, return on total capital.

8.1.1 Identifying the "Best" Set of Variables

Figure 8.1 gives a plot of the data, which can be used to visually assess the extent to which the two ratios discriminate between the two groups. The projections of the points onto the two axes, representing *EBITASS* and *ROTC*, give the values for the respective ratios. It is clear that the two groups of firms are well separated with respect to each ratio. In other words, each ratio does discriminate between the two groups of firms. Examining differences between groups with respect to a single variable is referred to as a *univariate analysis*. That is, does each variable (ratio) discriminate between the two groups? The univariate term is used to emphasize the fact that the differences between the two groups are assessed for each variable independent of the remaining variables. It is also clear that the two groups are well separated in the two-dimensional space, which implies that both the ratios *combined* or *jointly* provide a good separation of the two groups of firms. Examining differences with respect to two or more variables simultaneously is referred to as *multivariate analysis*. It is clear that the multivariate term is used to emphasize that the differences in the means of the two groups are assessed simultaneously for all the variables.

In the above example, based on a visual analysis, both of the variables do seem to discriminate between the two groups. This may not always be the case. Consider, for instance, the case where data on four financial ratios, X_1, X_2, X_3, and X_4, are available for the two groups of firms. Figure 8.2 portrays the distribution of each financial ratio. From the figure it is apparent that there is a greater difference between most-admired and least-admired firms with respect to ratios X_1 and X_2 than with respect to ratios X_3 and X_4. That is, ratios X_1 and X_2 are the variables that provide the "best" discrimination between the two groups. *Identifying a set of variables that "best" discriminates between the two groups is the first objective of discriminant analysis.* Variables providing the best discrimination are called *discriminator variables*.

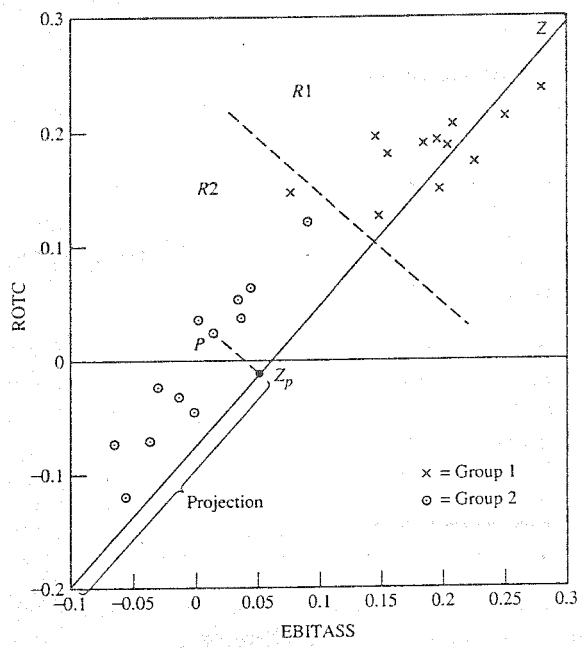


Figure 8.1 Plot of data in Table 8.1 and new axis.

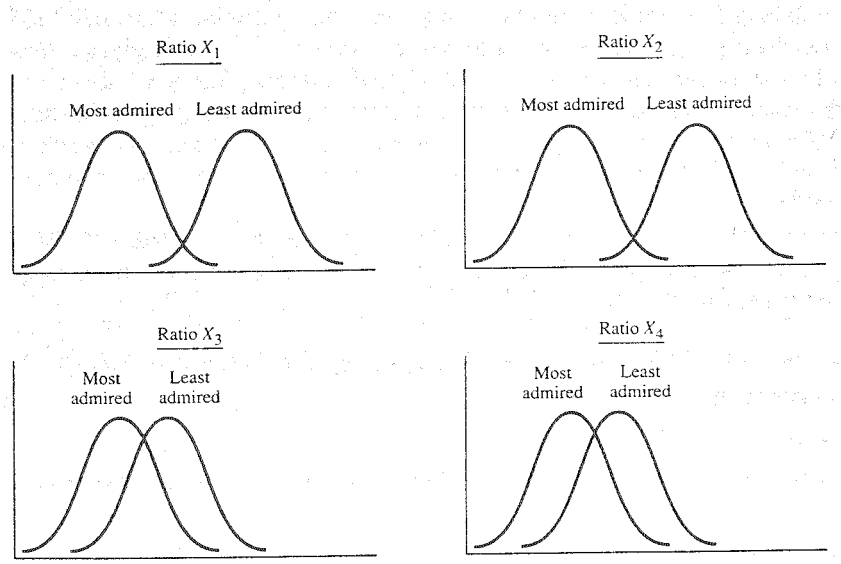


Figure 8.2 Distributions of financial ratios.

8.1.2 Identifying a New Axis

In Figure 8.1, consider a new axis, Z, in the two-dimensional space which makes an angle of, say, 45° with the *EBITASS* axis. The projection of any point, say P, on Z will be given by:

$$Z_p = w_1 EBITASS + w_2 ROTC,$$

Table 8.2 **Summary Statistics for Various Linear Combinations**

	Weights		Sum of Squares			λ
θ	w_1	w_2	SS_t	SS_w	SS_b	(SS_b/SS_w)
0	1.000	0.000	0.265	0.053	0.212	4.000
10	0.985	0.174	0.351	0.069	0.282	4.087
20	0.940	0.342	0.426	0.083	0.343	4.133
21	0.934	0.358	0.432	0.084	0.348	4.143
30	0.866	0.500	0.481	0.094	0.387	4.117
40	0.766	0.643	0.510	0.101	0.409	4.050
50	0.643	0.766	0.509	0.102	0.407	3.999
60	0.500	0.866	0.479	0.098	0.381	3.888
70	0.342	0.940	0.422	0.089	0.333	3.742
80	0.174	0.985	0.347	0.077	0.270	3.506
90	0.000	1.000	0.261	0.062	0.199	3.210

where Z_p is the projection of point or firm P on the Z axis. According to Eq. 2.24 of Chapter 2, $w_1 = \cos 45^\circ = .707$ and $w_2 = \sin 45^\circ = .707$. Therefore,

$$Z_p = .707 \times EBITASS + .707 \times ROTC.$$

This equation clearly represents a linear combination of the financial ratios, *EBITASS* and *ROTC*, for firm P. That is, the projection of points onto the Z axis gives a new variable Z, which is a linear combination of the original variables. Table 8.1 also gives the values of this new variable. The total sum of squares (SS_t), the between-group sum of squares (SS_b), and the within-group sum of squares (SS_w) for Z are, respectively, 0.513, 0.411, and 0.102. The ratio, λ, of the between-group to the within-group sum of squares is 4.029 (0.411/0.102).

Table 8.2 gives SS_t, SS_w, SS_b, and λ for various angles between Z and *EBITASS*. Figure 8.3 gives a plot of λ and θ, the angle between Z and *EBITASS*. From the table and the figure we see that:

1. When $\theta = 0^\circ$ or $\theta = 90^\circ$, only the corresponding variable, *EBITASS* or *ROTC*, is used for forming the new variable.
2. The value of λ changes as θ changes.
3. There is *one and only one* angle (i.e., $\theta = 21^\circ$) that results in a maximum value for λ.[1]
4. The following linear combination results in the maximum value for the λ:

$$\begin{aligned} Z &= \cos 21^\circ \times EBITASS + \sin 21^\circ \times ROTC \\ &= 0.934 \times EBITASS + 0.358 \times ROTC. \end{aligned} \tag{8.1}$$

The new axis, Z, is chosen such that the new variable Z gives a maximum value for λ. As shown below, the maximum value of λ implies that the new variable Z provides the maximum separation between the two groups.

[1]The other angle that gives a maximum value is 201° ($180^\circ + 21^\circ$) and the resulting axis is merely a reflection of the Z axis.

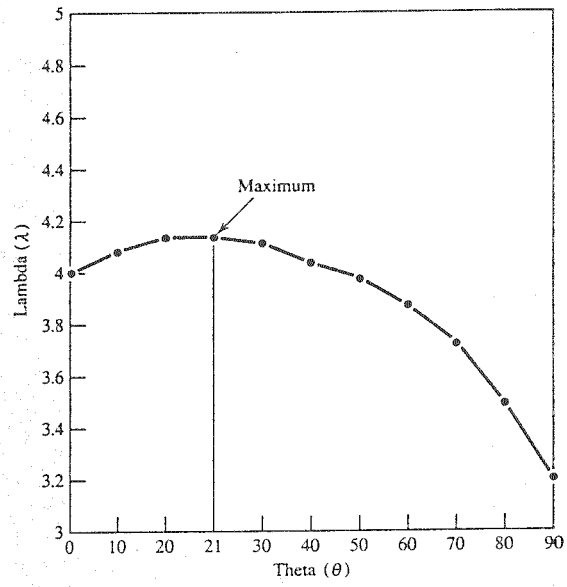


Figure 8.3 Plot of lambda versus theta.

For Z to provide maximum separation, the following two conditions must be satisfied.

1. The means of Z for the two groups should be as far apart as possible (which is equivalent to having a maximum value for the between-group sum of squares).
2. Values of Z for each group should be as homogeneous as possible (which is equivalent to having a minimum value for the within-group sum of squares).

Satisfying just one of these conditions will not result in a maximum separation. For example, compare the distribution of two possible new variables, Z, shown in Panels I and II of Figure 8.4. The difference in the means of the two groups for the new variable shown in Panel I is greater than that of the new variable shown in Panel II. However, the new variable in Panel II provides a better separation or discrimination than that in Panel I because the two groups of firms in Panel II, compared to Panel I, are more homogeneous with respect to the new variable. A measure of group homogeneity is provided by the within-group sum of squares, and a good measure of the difference in the means of the two groups is given by the between-group sum of squares. Therefore, it is obvious that for maximum separation or discrimination the new axis, Z, should be selected such that the ratio of SS_b to SS_w for the new variable is maximum. *The second objective of discriminant analysis is to identify a new axis, Z, such that the new variable Z, given by the projection of observations onto this new axis, provides the maximum separation or discrimination between the two groups.*

Note the similarity and the difference between discriminant analysis and principal components analysis. In both cases, a new axis is identified and a new variable is formed that is a linear combination of the original variables. That is, the new variable is given by the projection of the points onto this new axis. The difference is with respect to the criterion used to identify the new axis. In principal components analysis, a new axis is identified such that the projection of the points onto the new axis accounts for *maximum variance* in the data, which is equivalent to maximizing SS_t, because there

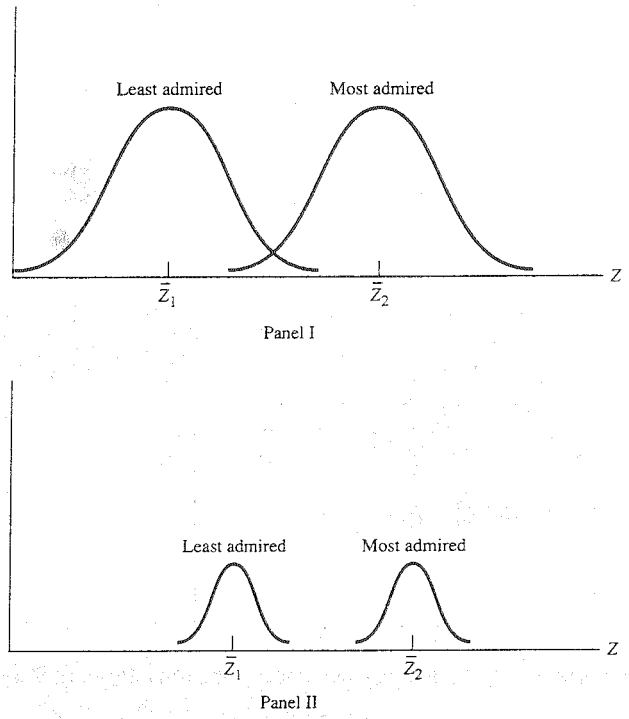


Figure 8.4 Examples of linear combinations.

is no criterion variable for dividing the sample into groups. In discriminant analysis, on the other hand, the objective is not to account for maximum variance in the data (i.e., maximize SS_t), but to maximize the between-group to within-group sum of squares ratio (i.e., SS_b/SS_w) that results in the best discrimination between the groups. The new axis, or the linear combination, that is identified is called the *linear discriminant function,* henceforth referred to as the *discriminant function.* The projection of a point onto the discriminant function (i.e., the value of the new variable) is called the *discriminant score.* For the data set given in Table 8.1 the discriminant function is given by Eq. 8.1 and the discriminant scores are given in Table 8.3.

8.1.3 Classification

The third objective of discriminant analysis is to classify future observations into one of the two groups. Actually, classification can be considered as an independent procedure unrelated to discriminant analysis. However, most textbooks and computer programs treat it as a part of the discriminant analysis procedure. We will discuss both approaches—classification as a separate procedure and as a part of discriminant analysis. It should be noted that, under certain conditions, both procedures give identical classification results. Section 8.3.3 provides further discussion of the two classification approaches.

Classification as a Part of Discriminant Analysis

Classification of future observations is done by using the discriminant scores. Figure 8.5 gives a one-dimensional plot of the discriminant scores, commonly referred to as a plot

Table 8.3 **Discriminant Score and Classification for Most-Admired and Least-Admired Firms (w_1 = .934 and w_2 = .358)**

Group 1			Group 2		
Firm Number	Discriminant Score	Classification	Firm Number	Discriminant Score	Classification
1	0.213	1	13	−0.022	2
2	0.270	1	14	0.053	2
3	0.261	1	15	0.048	2
4	0.346	1	16	−0.085	2
5	0.253	1	17	−0.093	2
6	0.274	1	18	−0.002	2
7	0.208	1	19	0.019	2
8	0.313	1	20	0.129	1
9	0.126	1	21	−0.059	2
10	0.185	1	22	0.065	2
11	0.241	1	23	−0.033	2
12	0.243	1	24	0.024	2
Average	0.244			0.00367	

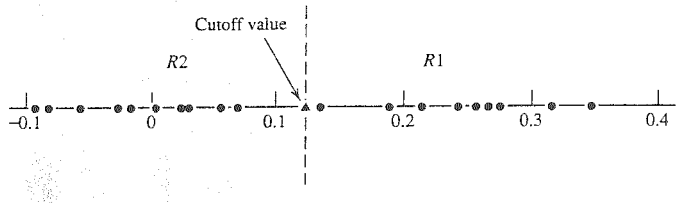


Figure 8.5 Plot of discriminant scores.

of the observations in the discriminant space. Classification of observations is done as follows. First, the discriminant space is divided into two mutually exclusive and collectively exhaustive regions, $R1$ and $R2$. Now since there is only one discriminant score, the plot shown in Figure 8.5 is a one-dimensional plot. Consequently, a point will divide the space into two regions. The value of the discriminant score that divides the one-dimensional space into the two regions is called the *cutoff* value. Next, the discriminant score of a given firm is plotted in the discriminant space and is classified as most-admired if the computed discriminant score for the firm falls in region $R1$ and least-admired if it falls in region $R2$. In other words, a given firm is classified as least-admired if the discriminant score of the observation is less than the cutoff value, and most-admired if the discriminant score is greater than the cutoff value. Once again, the estimated cutoff value is one that minimizes a given criterion (e.g., minimizes misclassification errors or misclassification costs).

Classification as an Independent Procedure

Classification essentially reduces to first partitioning a given p-dimensional variable space into two *mutually exclusive and collectively exhaustive* regions. Next, any given observation is plotted or located in the p-dimensional space and the observation is assigned to the group in whose region it falls. For example, in Figure 8.1 the dotted line divides the two-dimensional space into regions $R1$ and $R2$. Observations falling in region $R1$ are classified as most-admired firms, and those firms falling in region $R2$ are classified as least-admired firms. It is clear that the classification problem reduces to developing an algorithm or a rule for dividing the given variable space into mutually exclusive and collectively exhaustive regions. Obviously, one would like to divide the space such that a certain criterion, say the number of incorrect classifications or misclassification costs, is minimized. The various algorithms or classification rules differ mainly with respect to the minimization criterion used to divide the p-dimensional space into the appropriate number of regions. Section A8.2 of the Appendix discusses in detail some of the commonly used criteria for dividing the variable space into classification regions and the ensuing classification rules.

8.2 ANALYTICAL APPROACH TO DISCRIMINANT ANALYSIS

The data given in Table 8.1 are used for discussing the analytical approach to discriminant analysis.

8.2.1 Selecting the Discriminator Variables

Table 8.4 gives means and standard deviations for the two groups. The differences in the means of the two groups can be assessed by using an independent sample t-test. The

Table 8.4 **Means, Standard Deviations, and t-values for Most- and Least-Admired Firms**

	Group 1		Group 2		
Variable	Mean	Std. Dev.	Mean	Std. Dev.	t-value
EBITASS	.191	.053	.003	.045	9.367
ROTC	.184	.030	.001	.069	8.337

t-values for testing equality of the means of the two groups are 9.367 for *EBITASS* and 8.337 for *ROTC*. The t-test suggests that the two groups are significantly different with respect to both of the financial ratios at a significance level of .05. That is, both financial ratios do discriminate between the two groups and consequently will be used to form the discriminant function. This conclusion is based on a univariate approach. That is, a separate independent t-test is done for each financial ratio. However, a preferred approach is to perform a multivariate test in which both financial ratios are tested *simultaneously* or *jointly*. A discussion of the multivariate test is provided in Section 8.3.2.

8.2.2 Discriminant Function and Classification

Let the linear combination or the discriminant function that forms the new variable (or the discriminant score) be

$$Z = w_1 \times EBITASS + w_2 \times ROTC \tag{8.2}$$

where Z is the discriminant function.[2] Analytically, the objective of discriminant analysis is to identify the weights, w_1 and w_2, of the above discriminant function such that

$$\lambda = \frac{\text{between-group sum of squares}}{\text{within-group sum of squares}} \tag{8.3}$$

is maximized. The discriminant function, given by Eq. 8.2, is obtained by maximizing Eq. 8.3 and is referred to as *Fisher's linear discriminant function*. This is clearly an optimization problem and its technical details are provided in the Appendix.

Normally the cutoff value selected for classification purposes is the one that minimizes the number of incorrect classifications or misclassification costs. Details pertaining to the formulae used for obtaining cutoff value are given in the Appendix.

8.3 DISCRIMINANT ANALYSIS USING SPSS

The data given in Table 8.1 are used to discuss the output generated by the discriminant analysis procedure in SPSS. Table 8.5 gives the SPSS commands. Following is a brief description of the procedure commands for discriminant analysis. The GROUPS subcommand specifies the variable defining group membership. The ANALYSIS

[2]Note that Eq. 8.2 can be viewed as a general linear model of the form $\mathbf{Z} = \boldsymbol{\beta}\mathbf{X}$ where $\mathbf{Z}$ and $\mathbf{X}$ are vectors of dependent and independent variables, respectively, and $\boldsymbol{\beta}$ is a vector of coefficients. In the present case, Z is the dependent variable, *EBITASS* and *ROTC* are the independent variables, and w_1 and w_2 are the coefficients.

Table 8.5 **SPSS Commands for Discriminant Analysis of Data in Table 8.1**

```
DATA LIST FREE
 /EBITASS ROTC EXCELL
BEGIN DATA
insert data here
END DATA
DISCRIMINANT GROUPS=EXCELL(1,2)
 /VARIABLES=EBITASS ROTC
 /ANALYSIS=EBITASS ROTC
 /METHOD=DIRECT
 /STATISTICS=ALL
 /PLOT=ALL
```

subcommand gives a potential list of variables to be used for forming the discriminant function. The METHOD subcommand specifies the method to be used for selecting variables to form the discriminant function. The DIRECT method is used when all the variables specified in the ANALYSIS subcommand are used to formulate the discriminant function. However, many times the researcher is not sure which of the potential discriminating variables should be used to form the discriminant function. In such cases, a list of potential variables is provided in the ANALYSIS subcommand and the program selects the "best" set of variables using a given statistical criterion. The selection of the best set of variables by the program is referred to as *stepwise discriminant analysis*.[3] A number of statistical criteria are available for conducting a stepwise discriminant analysis. These criteria will be discussed in Section 8.6. The STATISTICS and the PLOT subcommands are used for obtaining the relevant statistics and plots for interpretation purposes. The ALL option indicates that the program should compute and print all the possible statistics and plots.

Exhibit 8.1 gives the partial output and is labeled for discussion purposes. The following discussion is keyed to the circled numbers in the output. For presentation clarity, most values reported in the text have been rounded to three significant digits; any differences between computations reported in the text and the output are due to rounding errors.

8.3.1 Evaluating the Significance of Discriminating Variables

The first step is to assess the significance of the discriminating variables. Do the selected discriminating variables significantly differentiate between the two groups? It appears that the means of each variable are different for the two groups [1]. A discussion of the formal statistical test for testing the difference between means of the two groups follows.

The null and the alternative hypotheses for each discriminating variable are:

$$H_o : \mu_1 = \mu_2$$

$$H_a : \mu_1 \neq \mu_2$$

where μ_1 and μ_2 are the population means, respectively, for groups 1 and 2. In Section 8.2.1, the above hypotheses were tested using an independent sample t-test. Alternatively, one can use the Wilks' Λ test statistic. Wilks' Λ is computed using the following

[3]The concept is similar to that used in stepwise multiple regression.

Exhibit 8.1

Discriminant analysis for most-admired and least-admired firms

(1) Group means

EXCELL	EBITASS	ROTC
1	.19133	.18350
2	.00333	.00125
Total	.09733	.09238

(2) Pooled within-groups covariance matrix with 22 degrees of freedom

	EBITASS	ROTC
EBITASS	2.4261515E-03	
ROTC	2.0336818E-03	2.8041477E-03

(3) Pooled within-groups correlation matrix

	EBITASS	ROTC
EBITASS	1.00000	
ROTC	.77969	1.00000

(4) Wilks' Lambda (U-statistic) and univariate F-ratio with 1 and 22 degrees of freedom

Variable	Wilks' Lambda	F	Significance
EBITASS	.20108	87.4076	.0000
ROTC	.23638	71.0699	.0000

(5) Total covariance matrix with 23 degrees of freedom

	EBITASS	ROTC
EBITASS	.0115	
ROTC	.0109	.0113

(6) Minimum tolerance level..... .00100

Canonical Discriminant Functions

Maximum number of functions..... 1
Minimum cumulative percent of variance... 100.00
Maximum significance of Wilks' Lambda.... 1.0000

Prior probability for each group is .50000

(7) Classification function coefficients (Fisher's linear discriminant functions)

EXCELL =	1	2
EBITASS	61.2374430	2.5511703
ROTC	21.0268971	-1.4044441
(Constant)	-8.4807470	-.6965214

(continued)

Exhibit 8.1 (continued)

8

Canonical Discriminant Functions

Fcn	Eigenvalue	Pct of Variance	Cum Pct	Canonical Corr	After Fcn	Wilks' Lambda	Chi-square	df	Sig
				:	0	.195162	34.312	2	.0000
1*	4.1239	100.00	100.00	.8971 :					

* Marks the 1 canonical discriminant functions remaining in the analysis.

9 Standardized canonical discriminant function coefficients

	Func 1
EBITASS	.74337
ROTC	.30547

10 Structure matrix:

Pooled within-groups correlations between discriminating variables and canonical discriminant functions
(Variables ordered by size of correlation within function)

	Func 1
EBITASS	.98154
ROTC	.88506

11 Unstandardized canonical discriminant function coefficients

	Func 1
EBITASS	15.0919163
ROTC	5.7685027
(Constant)	-2.0018120

12 Canonical discriminant functions evaluated at group means (group centroids)

Group	Func 1
1	1.94429
2	-1.94429

13 Test of Equality of Group Covariance Matrices Using Box's M

The ranks and natural logarithms of determinants printed are those of the group covariance matrices.

Group Label	Rank	Log Determinant
1	2	-13.516047
2	2	-14.107651
Pooled within-groups covariance matrix	2	-12.834397

Box's M	Approximate F	Degrees of freedom	Significance
21.50395	6.46365	3, 87120.0	.0002

14

Case Number	Mis Val	Sel	Actual Group	Highest Group	Probability P(D/G)	P(G/D)	2nd Highest Group	P(G/D)	Discrim Scores
1			1	1	.6088	.9962	2	.0038	1.4326
2			1	1	.6807	.9999	2	.0001	2.3558

(continued)

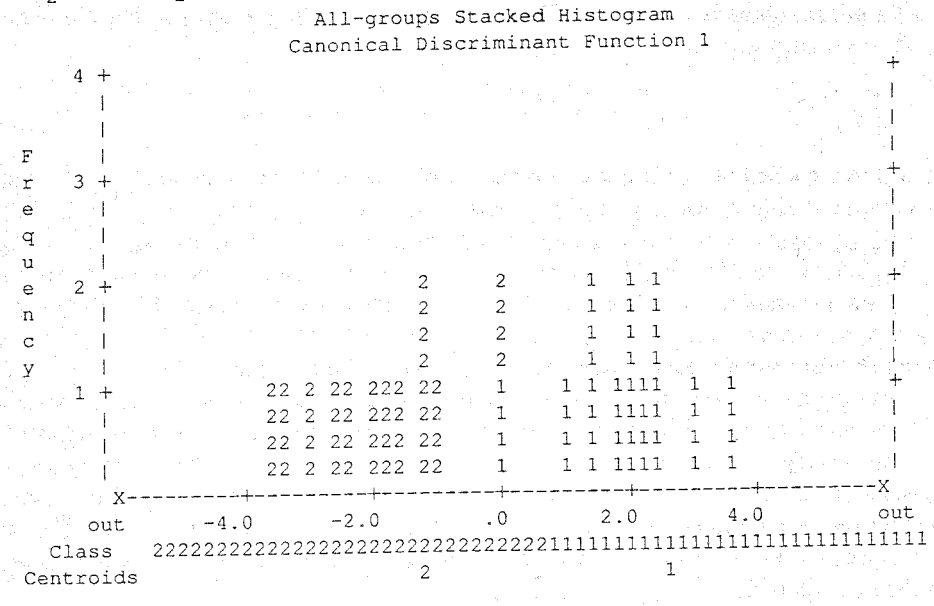
Exhibit 8.1 (continued)

3	1		1	.7930	.9998	2	.0002	2.2067
4	1		1	.1008	1.0000	2	.0000	3.5853
.	.		.	.	.	.	.	.
.	.		.	.	.	.	.	.
.	.		.	.	.	.	.	.
20	2	**	1	.0616	.5727	2	.4273	.0753
21	2		2	.3096	1.0000	1	.0000	-2.9605
22	2		2	.3218	.9761	1	.0239	-.9535
23	2		2	.5563	.9999	1	.0001	-2.5326
24	2		2	.7384	.9981	1	.0019	-1.6104

Symbols used in plots

15

Symbol	Group	Label
1	1	
2	2	



16 Classification results -

Actual Group		No. of Cases	Predicted Group Membership 1	2
Group	1	12	12 100.0%	0 .0%
Group	2	12	1 8.3%	11 91.7%

Percent of "grouped" cases correctly classified: 95.83%

formula:

$$\Lambda = \frac{SS_w}{SS_t}. \tag{8.4}$$

SS_w is obtained from the $\mathbf{SSCP}_w$ matrix, which in turn can be computed by multiplying the $\mathbf{S}_w$ [2] matrix by its pooled degrees of freedom.[4] The

$$\mathbf{SSCP}_w = \begin{pmatrix} 0.0534 & 0.0447 \\ 0.0447 & 0.0617 \end{pmatrix}.$$

$\mathbf{SS}_t$ is obtained from the $\mathbf{SSCP}_t$ matrix, which in turn is computed by multiplying $\mathbf{S}_t$ [5] by the total degrees of freedom.[5] Therefore, $\mathbf{SSCP}_t$ is equal to

$$\mathbf{SSCP}_t = \begin{pmatrix} 0.265 & 0.250 \\ 0.250 & 0.261 \end{pmatrix}.$$

Using Eq. 8.4, Wilks' Λ's for *EBITASS* and *ROTC* are, respectively, equal to .202 (i.e., .0534 ÷ .265) and .236 (i.e., 0.0617 ÷ 0.261), which, within rounding error, are the same as reported in the output [4]. Note that the smaller the value for Λ the greater the probability that the null hypothesis will be rejected and vice versa. To assess the statistical significance of the Wilks' Λ, it can be converted into an F-ratio using the following transformation:

$$F = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n_1 + n_2 - p - 1}{p}\right) \tag{8.5}$$

where p (which is 1 in this case) is the number of variable(s) for which the statistic is computed. Given that the null hypothesis is true, the F-ratio follows an F-distribution, with p and $n_1 + n_2 - p - 1$ degrees of freedom. The corresponding F-ratios, using Eq. 8.5, are 86.911 and 71.220, respectively, which again, within rounding errors, are the same as reported in the output [4]. Based on the critical F-value, the null hypotheses for both variables can be rejected at a significance level of .05. That is, the two groups are different with respect to *EBITASS* and *ROTC*. Once again, because the two groups were compared *separately* for each variable, the statistical significance tests are referred to as *univariate tests*. The univariate Wilks' Λ test for the difference in means of two groups is identical to the t-test discussed in Section 8.2.1. In fact, for two groups $F = t^2$. Note that the t^2 values obtained from Table 8.4, within rounding errors, are the same as the F-values computed above (i.e., $9.367^2 = 87.741$ and $8.337^2 = 69.506$). The reason for employing the Wilks' Λ test statistic instead of the t-test will become apparent in the next section.

8.3.2 The Discriminant Function

Options for Computing the Discriminant Function

The program prints the various parameters or options that have been selected for computing the discriminant function and for classifying observations in the sample [6]. A discussion of these options follows.

Since discriminant analysis involves the inversion of within-group matrices (see the Appendix), the accuracy of the computations is severely affected if the matrices are

[4]Recall that the pooled degrees of freedom is equal to $n_1 + n_2 - 2$ where n_1 and n_2 are, respectively, the number of observations for groups 1 and 2.

[5]Recall that the total degrees of freedom is equal to $n_1 + n_2 - 1$.

near singular (i.e., some of the discriminator variables are highly correlated or are linear combinations of other variables). The *tolerance level* provides a control for the desired amount of computational accuracy or the degree of multicollinearity that one is willing to tolerate. The tolerance of any variable is equal to $1 - R^2$, where R^2 is the squared multiple correlation between this variable and other variables in the discriminant function. The higher the multiple correlation between a given variable and the variables in the discriminant function, the lower the tolerance and vice versa. That is, tolerance is a measure of the amount of multicollinearity among the discriminator variables. If the tolerance of a given variable is less than the specified value, then the variable is not included in the discriminant function. Tolerance, therefore, is used to specify the degree to which the variables can be correlated and still be used to form the discriminant function. A default value of .001 is used by the SPSS program; however, one can specify any desired level for the tolerance (see the SPSS manual for the necessary commands).

The maximum number of discriminant functions that can be computed is the minimum of $G - 1$ or p, where G is the number of groups and p is the number of variables. Since the number of groups is equal to 2, only one discriminant function is possible. The *minimum cumulative percent of variance* is discussed in the next chapter, because the amount of variance extracted is only meaningful when there are more than two groups. The *maximum level of significance for Wilks'* Λ is only meaningful for stepwise discriminant analysis, and is discussed in Section 8.6.

The prior probability of each group is the probability of any random observation belonging to that group (i.e., Group 1 or Group 2). That is, it is the probability that a given firm is the most- or least-admired firm if no other information about the firm is available. In the present case, it is assumed that the priors are equal; that is, the prior probabilities of a given firm being most- or least-admired are each equal to 0.50.

Estimate of the Discriminant Function

The unstandardized estimate of the discriminant function is [11]:

$$Z = -2.00181 + 15.0919 \times EBITASS + 5.769 \times ROTC. \tag{8.6}$$

This is referred to as the unstandardized discriminant function because unstandardized (i.e., raw) data are used for computing the discriminant function. The discriminant function is also referred to as the *canonical discriminant function*, because discriminant analysis is a special case of canonical correlation analysis. The estimated weights of the discriminant function given by Eq. 8.6 appear to be different from the weights of the discriminant function in Eq. 8.1. However, as shown below, the weights of the two equations, in a relative sense, are the same.

The coefficients of the discriminant function are not unique; they are unique only in a relative sense. That is, only the ratio of the coefficients is unique. For example, the ratio of the weights given in Eq. 8.6 is 2.616 (15.0919 ÷ 5.769) which, within rounding error, is the ratio of the coefficients given in Eq. 8.1 (0.934 ÷ 0.358 = 2.609). The coefficients given in Eq. 8.6 can be normalized to sum to one by dividing each coefficient by $\sqrt{w_1^2 + w_2^2}$.[6] Normalized discriminant function coefficients are

$$\begin{aligned} w_1 &= \frac{15.0919}{\sqrt{15.0919^2 + 5.769^2}} \\ &= .934 \end{aligned}$$

[6]Since normalizing is done by dividing each coefficient by a constant, it does not change the relative value of the discriminant score.

and

$$w_2 = \frac{5.769}{\sqrt{15.0919^2 + 5.769^2}}$$
$$= .357.$$

As can be seen, the normalized weights, within rounding errors, are the same as the ones used in Eq. 8.1. A constant is added to the unstandardized discriminant function so that the average of the discriminant scores is zero and, therefore, the constant simply adjusts the scale of the discriminant scores.

Statistical Significance of the Discriminant Function

Differences in the means of two groups for each discriminator variable were tested using the univariate Wilks' Λ test statistic in Section 8.3.1. However, in the case of more than one discriminator variable it is desirable to test the differences between the two groups for all the variables jointly or simultaneously. This *multivariate test of significance* has the following null and alternate hypotheses:

$$H_o : \begin{pmatrix} \mu^1_{EBITASS} \\ \mu^1_{ROTC} \end{pmatrix} = \begin{pmatrix} \mu^2_{EBITASS} \\ \mu^2_{ROTC} \end{pmatrix},$$

and

$$H_a : \begin{pmatrix} \mu^1_{EBITASS} \\ \mu^1_{ROTC} \end{pmatrix} \neq \begin{pmatrix} \mu^2_{EBITASS} \\ \mu^2_{ROTC} \end{pmatrix}.$$

The test statistic for testing these multivariate hypotheses is a direct generalization of the univariate Wilks' Λ statistic and is given by

$$\Lambda = \frac{|\mathbf{SSCP_w}|}{|\mathbf{SSCP_t}|},$$

where $|\cdot|$ represents the determinant of a given matrix.[7] Wilks' Λ can be approximated as a chi-square statistic using the following transformation:

$$\chi^2 = -[n - 1 - (p + G)/2] \ln \Lambda. \tag{8.7}$$

The χ^2 statistic is distributed as a chi-square distribution with $p(G - 1)$ degrees of freedom. The Wilks' Λ for the discriminant function is .195 [8], and its equivalent χ^2 value is

$$\chi^2 = -[24 - 1 - (2 + 2)/2] \ln(.195) = 34.330,$$

which, within rounding error, is the same as given in the output [8]. The null hypothesis can be rejected at an alpha level of 0.05, implying that the two groups are significantly different with respect to *EBITASS* and *ROTC* taken jointly.

Since the discriminant function is a linear combination of discriminator variables, it can also be concluded that the discriminant function is statistically significant. That is, the means of the discriminant scores for the two groups are significantly different. Statistical significance of the discriminant function can also be assessed by transforming

[7] Various test statistics such as the t-statistic, F-statistic, and Hotelling's T^2 are special cases of the Wilks' Λ test statistic.

Wilks' Λ into an exact F-ratio using Eq. 8.5. That is,

$$F = \left(\frac{1 - .195}{.195}\right)\left(\frac{24 - 2 - 1}{2}\right) = 43.346,$$

which is statistically significant at an alpha level of .05.

Practical Significance of the Discriminant Function

It is quite possible that the difference between the two groups is statistically significant even though, for all practical purposes, the differences between the groups may not be large. This can occur for large sample sizes. Practical significance relates to assessing how large or how meaningful the differences between the two groups are. The output reports the canonical correlation, which is equal to 0.897 [8]. As discussed below, the square of the canonical correlation can be used as a measure of the practical significance of the discriminant function.

It can be shown that the squared canonical correlation (CR^2) is equal to

$$CR^2 = \frac{SS_b}{SS_t}, \tag{8.8}$$

or

$$CR = \sqrt{\frac{SS_b}{SS_t}}. \tag{8.9}$$

From Eq. 8.8 it is obvious that CR^2 gives the proportion of the total sum of squares for the discriminant score that is due to the differences between the groups. Furthermore, as will be shown in Section 8.4, two-group discriminant analysis can also be formulated as a multiple regression problem. The corresponding multiple R that would be obtained is the same as the canonical correlation. Recall that in regression analysis R^2 is a measure of the amount of variance in the dependent variable that is accounted for by the independent variables, and therefore it is a measure of the strength of the relationship between the dependent and the independent variables. Since the discriminant score is a linear function of the discriminating variables, CR^2 gives the amount of variation between the groups that is explained by the discriminating variables. Hence, CR^2 is a measure of the strength of the discriminant function.

From the output, CR is equal to 0.897 [8] and therefore CR^2 is equal to .804. That is, about 80% of the variation between the two groups is accounted for by the discriminating variables, which appears to be quite high. Although CR^2 ranges between zero and one, there are no guidelines to suggest how high is "high." The researcher therefore should compare this CR^2 to those obtained in other similar applications and determine if the strength of the relationship is relatively strong, moderate, or weak.

Assessing the Importance of Discriminant Variables and the Meaning of the Discriminant Function

If discriminant analysis is done on standardized data then the resulting discriminant function is referred to as standardized canonical discriminant function. However, a separate analysis is not needed as standardized coefficients can be computed from the unstandardized coefficients by using the following transformation:

$$\hat{b}_j^* = \hat{b}_j \hat{s}_j,$$

where $\hat{b}_j^*$, $\hat{b}_j$, and $\hat{s}_j$ are, respectively, standardized coefficient, unstandardized coefficient, and the pooled standard deviation of variable j. The standardized coefficients for

EBITASS and *ROTC*, respectively, are equal to .743 (i.e., $15.0919\sqrt{.0024261}$) and .305 (i.e., $5.769\sqrt{0.002804}$), which are the same as given in the output [9]. Note that .0024261 and 0.002804 are the pooled variances for variables *EBITASS* and *ROTC*, respectively, and are obtained from Σ_w [2].

Standardized coefficients are normally used for assessing the relative importance of discriminator variables forming the discriminant function. The greater the standardized coefficient, the greater the relative importance of a given variable and vice versa. Therefore, it appears that *ROTC* is relatively less important than *EBITASS* in forming the discriminant function. However, caution is advised in such an interpretation when the variables are correlated among themselves. Depending on the severity of multicollinearity present in the sample data, the relative importance of the variables could change from sample to sample. Consequently, in the presence of multicollinearity in the data, it is recommended that inferences regarding the importance of the discriminator variables be avoided. The problem is similar to that posed by multicollinearity in multiple regression analysis. Since the discriminant score is a composite index or a linear combination of original variables, it might be interesting to know what exactly the discriminant score represents. In other words, just as in principal components and factor analysis, a label can be assigned to the discriminant function. The loadings or the *structure coefficients* are helpful for assigning the label and also for interpreting the contribution of each variable to the formation of the discriminant function. The loading of a given discriminator variable is simply the correlation coefficient between the discriminant score and the discriminator variable and the value of the loading will lie between +1 and −1. The closer the absolute value of the loading of a variable to 1, the more communality there is between the discriminating variable and the discriminant function and vice versa. Loadings are given in the *structure matrix* [10]. Alternatively, they can be computed using the following formula:

$$l_i = \sum_{j=1}^{p} r_{ij} b_j^*, \tag{8.10}$$

where l_i is the loading of variable i, r_{ij} is the pooled correlation between variable i with variable j, and b_j^* is the standardized coefficient of variable j. For example, the loading of *EBITASS* ($i = 1$) is given by

$$l_1 = 1.000 \times .743 + .780 \times .305,$$

and is equal to .981. Note that 0.780 is the pooled correlation between *EBITASS* and *ROTC* [3]. Since the loadings of both of the discriminator variables are high, the discriminant score can be interpreted as a measure of the financial health of a given firm. Once again, how "high" is high is a judgmental question; many researchers have used a value of 0.50 as the cutoff value. Also, the contribution of both the variables toward the formation of the discriminant function is high because both variables have high loadings.

8.3.3 Classification Methods

A number of methods are available for classifying sample and future observations. Some of the commonly used methods are:

1. Cutoff-value method.
2. Statistical decision theory method.

3. Classification function method.
4. Mahalanobis distance method.

Cutoff-Value Method

As discussed earlier, classification of observations essentially reduces to dividing the discriminant space into two regions. The value of the discriminant score that divides the space into the two regions is called the cutoff value. Following is a discussion of how the cutoff value is computed.

Table 8.3 gives the discriminant score for each observation that was formed by using Eq. 8.1. As can be seen from Eq. 8.1, the greater the values for *EBITASS* and *ROTC* the greater the value for the discriminant score and vice versa. Since financially healthy firms will have higher values for the two financial ratios, most-admired firms will have a greater discriminant score than least-admired firms. Therefore, any given firm will be classified as a most-admired firm if its discriminant score is greater than the cutoff value, and as a least-admired firm if its discriminant score is less than the cutoff value. Normally the cutoff value selected is the one that minimizes the number of incorrect classifications or misclassification errors. A commonly used cutoff value that minimizes the number of incorrect classifications for the sample data is

$$\text{cutoff value} = \frac{\bar{Z}_1 + \bar{Z}_2}{2}, \tag{8.11}$$

where $\bar{Z}_j$ is the average discriminant score for group j. This formula assumes equal sample sizes for the two groups. For unequal sample sizes the cutoff value is given by:

$$\text{cutoff value} = \frac{n_1\bar{Z}_1 + n_2\bar{Z}_2}{n_1 + n_2}, \tag{8.12}$$

where n_g is the number of observations in group g. From Table 8.3, the averages of the discriminant scores for groups 1 and 2 are, respectively, 0.244 and 0.00367, and the cutoff value will be

$$\begin{aligned}\text{cutoff value} &= \frac{0.244 + 0.00367}{2} \\ &= 0.124.\end{aligned}$$

Table 8.3 also gives the classification of the firms based on the computed discriminant score and the cutoff value of 0.124. Note that only the 20th observation is misclassified, giving a correct classification rate of 95.83% (i.e., 23 ÷ 24). Identical results are obtained if classification is done using the discriminant scores computed from the unstandardized discriminant function given by Eq. 8.6 (i.e., the one reported in the output [11]). Discriminant scores resulting from Eq. 8.6 are also given in the output [14]. The average of the discriminant scores for groups 1 and 2, respectively, are 1.944 and −1.944 [12] giving a cutoff value of zero. Once again the 20th observation is misclassified.

A summary of the classification results is provided in a matrix known as the *classification matrix* or the *confusion matrix*. The classification matrix is given at the end of the output [16]. All but one of the observations have been correctly classified.

There are other rules for computing cutoff values and for classifying future observations. Equation 8.11 assumes equal misclassification costs and equal priors. Equal

misclassification costs implies that the penalty or the cost of misclassifying observations in groups 1 or 2 is the same. That is, the cost of misclassifying a most-admired firm is the same as misclassifying a least-admired firm. Equal priors imply that the prior probabilities are equal. That is, any given firm selected at random will have an equal chance of being either a most- or a least-admired firm. Alternative classification procedures or rules that relax these assumptions are discussed below.

Statistical Decision Theory

SPSS uses the statistical decision theory method for classifying sample observations into various groups. This method minimizes *misclassification errors*, taking into account prior probabilities and misclassification costs. For example, the data set given in Table 8.1 consists of an equal number of most- and least-admired firms. This does not imply that the population also has an equal number of most- and least-admired firms. It is quite possible that, say, 70% of the firms in the population are most-admired firms and only 30% of firms are least-admired. That is, the probability of any given firm being most-admired is .7. This probability is known as the prior probability.

Misclassification costs also may not be equal. For example, in the case of a jury verdict the "social" cost of finding an innocent person guilty might be much more than finding a guilty person innocent. Or, in the case of studies dealing with bankruptcy prediction, it might be more costly to classify a healthy firm as a potential candidate for bankruptcy than a potentially bankrupt firm as healthy.

Classification rules that incorporate prior probabilities and misclassification costs are based on *Bayesian theory*. Bayesian theory essentially revises prior probabilities based on additional available information. That is, if nothing is known about a given firm, then the probability that it belongs to the most-admired group is p_1, where p_1 is the prior probability. Based on additional information about the firm (i.e., its values for *EBITASS* and *ROTC*) the prior probability can be revised to q_1. Revising the prior probability p_1 to a posterior or revised probability q_1, based on additional information, is the whole essence of Bayesian theory. A classification rule incorporating prior probabilities is given by:

Assign the observation to group 1 if

$$Z \geq \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln\left[\frac{p_2}{p_1}\right], \tag{8.13}$$

and assign to group 2 if

$$Z < \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln\left[\frac{p_2}{p_1}\right], \tag{8.14}$$

where Z is the discriminant score for a given observation, $\bar{Z}_j$ is the average discriminant score for group j, and p_j is the prior probability of group j.

Misclassification costs also can be incorporated into the above classification rule. For example, consider the 2×2 misclassification cost table given in Table 8.6. In the table, $C(i/j)$ is the cost of misclassifying into group i an observation that belongs to group j. The rule for classifying observations which incorporate prior probabilities and misclassification costs is given by:

Assign the observation to group 1 if

$$Z \geq \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln\left[\frac{p_2 C(1/2)}{p_1 C(2/1)}\right], \tag{8.15}$$

Table 8.6 **Misclassification Costs**

	Actual Membership	
Predicted Membership	Group 1	Group 2
Group 1	Zero cost	$C(1/2)$
Group 2	$C(2/1)$	Zero cost

and assign to group 2 if

$$Z < \frac{\bar{Z}_1 + \bar{Z}_2}{2} + \ln\left[\frac{p_2 C(1/2)}{p_1 C(2/1)}\right]. \tag{8.16}$$

Equations 8.15 and 8.16 give the general classification rule based on statistical decision theory and this rule minimizes misclassification errors. The classification rule is derived assuming that the discriminator variables have a multivariate normal distribution (see the Appendix). From Eqs. 8.15 and 8.16, it is obvious that the cutoff value is shifted toward the group that has a lower prior or a lower cost of misclassification. Or geometrically, the classification region or space increases for groups that have a higher prior or a higher misclassification cost.

Note that for equal misclassification costs and equal priors, Eqs. 8.15 and 8.16 reduce to:
Assign the observation to group 1 if

$$Z \geq \frac{\bar{Z}_1 + \bar{Z}_2}{2},$$

and assign the observation to group 2 if

$$Z < \frac{\bar{Z}_1 + \bar{Z}_2}{2}.$$

The right-hand side of the above equations is the cutoff value used in the cutoff-value classification method. Therefore, the cutoff-value classification method is the same as the statistical decision theory method with equal priors, equal misclassification costs, and assuming that the data come from a multivariate normal distribution.

In many instances the researcher is interested not in classification of the observations but in their posterior probabilities. SPSS computes the posterior probabilities under the assumption that the data come from a multivariate normal distribution and that the covariance matrices of the two groups are equal. The interested reader is referred to the Appendix for further details and for the formula used for computing posterior probabilities. The posterior probabilities (given by $P(G/D)$, where G represents the group and D is the discriminant score) are given in the output [14]. The posteriors can be used for classifying observations. An observation is assigned to the group with the highest posterior probability. For example, the posterior probabilities of Observation 20 for groups 1 and 2 are, respectively, 0.573 and 0.427. Therefore, once again, the observation is misclassified into group 1.

Classification Functions

Classification can also be done by using the classification functions computed for each group. Classifications based on classification functions are identical to those given by

Eqs. 8.15 and 8.16. SPSS computes the classification functions. The classification functions reported in the output are [7]:

$$C_1 = -8.481 + 61.237 \times EBITASS + 21.0269 \times ROTC$$

for group 1, and

$$C_2 = -0.697 + 2.551 \times EBITASS - 1.404 \times ROTC$$

for group 2. Observations are assigned to the group with the largest classification score. The coefficients of the classification functions are not interpreted. These functions are used solely for classification purposes. Furthermore, as will be seen later, prior probabilities and misclassification costs only affect the constant of the preceding equations; the coefficients of the classification functions are not affected.

Mahalanobis Distance Method

Observations can also be classified using Mahalanobis or the statistical distance (computed by employing original variables) of each observation from the centroid of each group. The observation is assigned to the group to which it is the closest as measured by the Mahalanobis distance. For example, using Eq. 3.9 of Chapter 3, the Mahalanobis distance for Observation 1 from the centroid of group 1 (i.e., most-admired firms) is equal to

$$\begin{aligned} MD^2 &= \frac{1}{1-.780^2}\left[\frac{(.158-.191)^2}{.00243} + \frac{(.182-.184)^2}{.0028} \right. \\ &\quad \left. - 2 \times .780 \frac{(.158-.191)(.182-.184)}{\sqrt{.00243}\sqrt{.0028}}\right] \\ &= 1.047. \end{aligned}$$

Table 8.7 gives the Mahalanobis distance for each observation and its classification into the respective group (see the Appendix for further details). Again, only the 20th observation is misclassified. Classifications employing the Mahalanobis distance method assume equal priors, equal misclassification costs, and multivariate normality for the discriminating variables.

How Good Is the Classification Rate?

The correct classification rate is 95.83% [16]. How good is this classification rate? Huberty (1984) has proposed approximate test statistics that can be used to evaluate the statistical and the practical significance of the overall classification rate and the classification rate for each group.

STATISTICAL TESTS. The test statistics for assessing the statistical significance of the classification rate for any group and for the overall classification rate are given by

$$Z_g^* = \frac{(o_g - e_g)\sqrt{n_g}}{\sqrt{e_g(n_g - e_g)}} \tag{8.17}$$

$$Z^* = \frac{(o - e)\sqrt{n}}{\sqrt{e(n - e)}} \tag{8.18}$$

Table 8.7 **Classification Based on Mahalanobis Distance**

Group 1				Group 2			
	Mahalanobis Distance from				Mahalanobis Distance from		
Firm Number	Group 1	Group 2	Classification	Firm Number	Group 1	Group 2	Classification
1	1.047	12.240	1	13	18.808	0.440	2
2	0.182	18.494	1	14	9.888	0.982	2
3	0.186	17.310	1	15	10.994	0.524	2
4	3.722	31.509	1	16	28.424	2.165	2
5	0.029	16.234	1	17	33.437	6.113	2
6	2.077	20.807	1	18	15.784	0.015	2
7	2.862	13.556	1	19	14.342	1.207	2
8	2.192	25.858	1	20	4.546	5.207	1
9	8.102	8.548	1	21	25.171	2.119	2
10	1.122	8.753	1	22	8.777	1.423	2
11	1.607	16.148	1	23	20.010	0.355	2
12	0.104	15.062	1	24	12.723	0.248	2

$$e_g = \frac{n_g^2}{n} \tag{8.19}$$

$$e = \frac{1}{n}\sum_{g=1}^{G} n_g^2, \tag{8.20}$$

where o_g is the number of correct classifications for group g; e_g is the expected number of correct classifications due to chance for group g; n_g is the number of observations in group g; o is the total number of correct classifications; e is the expected number of correct classifications due to chance for the total sample; and n is the total number of observations. The test statistics, Z_g^* and Z^*, follow an approximately normal probability distribution. From Eqs. 8.19 and 8.20, $e_1 = e_2 = 6$ and $e = 12$, and from Eqs. 8.17 and 8.18:

$$Z_1^* = \frac{(12-6)\sqrt{12}}{\sqrt{6(12-6)}} = 3.464$$

$$Z_2^* = \frac{(11-6)\sqrt{12}}{\sqrt{6(12-6)}} = 2.887$$

$$Z^* = \frac{(23-12)\sqrt{24}}{\sqrt{12(24-12)}} = 4.491.$$

The statistics, Z_1^*, Z_2^*, and Z^* are significant at an alpha level of .05, suggesting that the number of correct classifications is significantly greater than due to chance.

An alternative definition of total classifications due to chance is based on the use of a *naive* prediction rule. In the naive prediction rule, all observations are classified into the largest group. If the sizes of the groups are equal then the naive prediction rule would give n/G correct classifications due to chance where n is the number of observations and G is the number of groups. In the present case, use of the naive classification rule would result in 12 correct classifications due to chance and $Z = 4.491$ which is significant at $p < .05$.

PRACTICAL SIGNIFICANCE. The practical significance of classification is the extent to which the classification rate obtained via the various classification techniques is better than the classification rate obtained due to chance alone. That is, to what extent is the improvement over classification due to chance if one uses the various classification techniques? The index to measure the improvement is given by (Huberty 1984)

$$I = \frac{o/n - e/n}{1 - e/n} \times 100. \tag{8.21}$$

The I in the above equation gives the percent reduction in error over chance classification that would result if a given classification method is used. Using Eq. 8.21 [16],

$$I = \frac{23/24 - 12/24}{1 - 12/24} \times 100$$
$$= 91.667.$$

That is, by using the classification method in the discriminant analysis procedure a 91.667% reduction in error over chance is obtained. In other words, 11 (i.e., 0.91667×12) observations over and above chance classifications are obtained by using the classification method in discriminant analysis.

Using Misclassification Costs and Priors for Classification in SPSS

Unequal priors can be specified in the SPSS using the PRIORS subcommand, but misclassification costs cannot be directly specified in the SPSS program. However, the program can be tricked into considering misclassification costs. The trick is to incorporate the misclassification costs into the priors. Suppose that the prior probabilities for the data set in Table 8.1 are $p_1 = .7$ and $p_2 = .3$, and it is four times as costly to misclassify a most-admired firm as a least-admired firm than it is to misclassify a least-admired firm as a most-admired firm. These costs of misclassification can be stated as

$$C(1/2) = 1 \qquad C(2/1) = 4.$$

Using the above information, first compute

$$p_1 \cdot C(2/1) = .7 \times 4 = 2.800$$

and

$$p_2 \cdot C(1/2) = .300 \times 1 = .300.$$

New prior probabilities are computed by normalizing the above equations to sum to one. That is,

$$\text{new } p_1 = \frac{2.8}{2.8 + 0.3} = .90, \qquad \text{new } p_2 = \frac{.3}{2.8 + 0.3} = .10.$$

Discriminant analysis can be rerun with the above new priors by including the following PRIORS subcommand, PRIORS = .9 .1. The resulting classification functions are

$$C_1 = -7.893 + 61.237 \times EBITASS + 21.027 \times ROTC$$

for group 1 and

$$C_2 = -2.306 + 2.551 \times EBITASS - 1.404 \times ROTC$$

for group 2. Note that only the constant has changed; the coefficients or weights of $EBITASS$ and $ROTC$ have not changed. Table 8.8 gives that part of the SPSS output which contains discriminant scores, classification, and posterior probabilities. Once again, note that discriminant scores have not changed even though the posterior probabilities have changed.

Summary of Classification Methods

In the present case all the methods provided the same classification. However, this may not always be the case. All four methods will give the same results: (1) if the data come from a multivariate normal distribution; (2) if the covariance matrices of the two groups are equal; and (3) if the misclassification costs and priors are equal. In general the results may be different depending upon which of the assumptions are satisfied. The effect of the first two assumptions on classification is discussed in Section 8.5.

Note that classification methods based on the Mahalanobis distance do not employ the discriminant function or the discriminant scores. That is, classification by the Mahalanobis method can be viewed as a separate technique that is independent of discriminant analysis. In fact, as discussed earlier, many textbooks prefer to treat discriminant analysis and classification as separate problems because all the classification methods discussed above can be shown to be independent of the discriminant function.

Table 8.8 **Discriminant Scores, Classification, and Posterior Probability for Unequal Priors**

Case Number	Actual Group	Highest Probability			2nd Highest		Discrim Scores
		Group	P(D/G)	P(G/D)	Group	P(G/D)	
1	1	1	0.6086	0.9996	2	0.0004	1.4326
2	1	1	0.6821	1.0000	2	0.0000	2.3543
3	1	1	0.7955	1.0000	2	0.0000	2.2039
4	1	1	0.1008	1.0000	2	0.0000	3.5859
5	1	1	0.8862	1.0000	2	0.0000	2.0878
6	1	1	0.6334	1.0000	2	0.0000	2.4216
7	1	1	0.5633	0.9995	2	0.0005	1.3668
8	1	1	0.2675	1.0000	2	0.0000	3.0535
9	1	1	0.0570	0.9135	2	0.0865	0.0413
10	1	1	0.3371	0.9976	2	0.0024	0.9848
11	1	1	0.9497	0.9999	2	0.0001	1.8817
12	1	1	0.9823	0.9999	2	0.0001	1.9225
13	2	2	0.6774	0.9991	1	0.0009	−2.3607
14	2	2	0.4278	0.9074	1	0.0926	−1.1517
15	2	2	0.4699	0.9280	1	0.0720	−1.2221
16	2	2	0.1510	1.0000	1	0.0000	−3.3808
17	2	2	0.1184	1.0000	1	0.0000	−3.5064
18	2	2	0.9329	0.9966	1	0.0034	−2.0289
19	2	2	0.8083	0.9881	1	0.0119	−1.7020
20	2**	1	0.0615	0.9234	2	0.0766	0.0750
21	2	2	0.3085	0.9999	1	0.0001	−2.9632
22	2	2	0.3216	0.8193	1	0.1807	−0.9536
23	2	2	0.5561	0.9995	1	0.0005	−2.5334
24	2	2	0.7370	0.9830	1	0.0170	−1.6089

**Misclassified observations.

8.3.4 Histograms for the Discriminant Scores

This section of the output gives a histogram for both the groups [15]. The histogram provides a visual display of group separation with respect to the discriminant score. As can be seen, there appears to be virtually no overlap between the two groups.

8.4 REGRESSION APPROACH TO DISCRIMINANT ANALYSIS

As mentioned earlier, two-group discriminant analysis can also be formulated as a multiple regression problem. The dependent variable is the group membership and is binary (i.e., 0 or 1). For the present example, we can arbitrarily code 0 for least-admired firms and 1 for most-admired firms. The independent variables are the discriminator variables. Exhibit 8.2 gives the multiple regression output for the data set given in Table 8.1.

Multiple R is equal to .897 [1], and it is the same as the canonical correlation reported in the discriminant analysis output given in Exhibit 8.1. The multiple regression equation is given by

$$Y = 0.086 + 3.124 \times EBITASS + 1.194 \times ROTC.$$

Exhibit 8.2

Multiple regression approach to discriminant analysis

```
(1) Multiple R            .89713
    R Square              .80484
    Adjusted R Square     .78625
    Standard Error        .23614

    Analysis of Variance
                        DF       Sum of Squares       Mean Square
    Regression           2              4.82903           2.41451
    Residual            21              1.17097            .05576

    F =       43.30142        Signif F =  .0000
    ------------------ Variables in the Equation ------------------
    Variable                B         SE B        Beta          T  Sig T

    EBITASS          3.123638     1.483193     .657003      2.106  .0474
    ROTC             1.193931     1.495806     .249005       .798  .4337
    (Constant)        .085677      .065683                  1.304  .2062
```

As stated previously, coefficients of the discriminant function are not unique. As before, ignoring the constant, the ratio of the coefficients is equal to 2.616 (3.124/1.194). The normalized coefficients for

$$EBITASS = \frac{3.124}{\sqrt{3.124^2 + 1.194^2}} = .934$$

$$ROTC = \frac{1.194}{\sqrt{3.124^2 + 1.194^2}} = .357,$$

within rounding error, are the same as those given by Eq. 8.1 and the normalized coefficients reported in Section 8.3.2. However, caution is advised in interpreting the statistical significance tests from regression analysis as the multivariate normality and homoscedasticity assumptions will be violated due to the binary nature of the dependent variable.

8.5 ASSUMPTIONS

Discriminant analysis assumes that data come from a multivariate normal distribution and that the covariance matrices of the groups are equal. Procedures to test for the violation of these assumptions are discussed in Chapter 12. In the following section we discuss the effect of the violation of these assumptions on the results of discriminant analysis.

8.5.1 Multivariate Normality

The assumption of multivariate normality is necessary for the significance tests of the discriminator variables and the discriminant function. If the data do not come from a multivariate normal distribution then, in theory, none of the significance tests are valid. Classification results, in theory, are also affected if the data do not come from a multivariate normal distribution. The real issue is the degree to which data can

deviate from nonnormality without substantially affecting the results. Unfortunately, there is no clear-cut answer as to "how much" of nonnormality is acceptable. However, the researcher should be aware that studies have shown that, although the overall classification error is not affected, the classification error of some groups might be overestimated and for other groups it might be underestimated (Lachenbruch, Sneeringer, and Revo 1973).

If there is reason to believe that the multivariate normality assumption is clearly being violated then one can use logistic regression analysis because it does not make any distributional assumptions for the independent variables. Logistic regression analysis is discussed in Chapter 10.

8.5.2 Equality of Covariance Matrices

Linear discriminant analysis assumes that the covariance matrices of the two groups are equal. Violation of this assumption affects the significance tests and the classification results. Research studies have shown that the degree to which they are affected depends on the number of discriminator variables and the sample size of each group (Holloway and Dunn 1967, Gilbert 1969, and Marks and Dunn 1974). Specifically, the null hypothesis of equal mean vectors is rejected more often than it should be when the number of discriminator variables is large or the sample sizes of the groups are different. That is, the significance level is inflated. Furthermore, as the number of variables increases the significance level becomes more sensitive to unequal sample sizes.

The classification rate is also affected and various rules do not result in a minimum amount of misclassification error. If the assumption of equal covariance matrices is rejected, one could use a quadratic discriminant function for classification purposes. However, it has been found that for small sample sizes the performance of linear discriminant function is superior to quadratic discriminant function, as the number of parameters that need to be estimated for the quadratic discriminant function is nearly doubled.

A statistical test is available for testing the equality of the covariance matrices. The null and alternate hypotheses for the statistical test are

$$H_o : \Sigma_1 = \Sigma_2$$
$$H_a : \Sigma_1 \neq \Sigma_2$$

where Σ_g is the covariance matrix for group g. The appropriate test statistic is Box's M and can be approximated as an F-statistic. SPSS reports the test statistic and, as can be seen, the null hypothesis is rejected (see the part circled 13 in Exhibit 8.1). However, the preceding test is sensitive to sample sizes in that for a large sample even small differences between the covariance matrices will be statistically significant.

To summarize, violation of the assumptions of equality of covariance matrices and nonnormality affects the statistical significance tests and classification results. As indicated previously, it has been shown that discriminant analysis is quite robust to the violations of these assumptions. Nevertheless, when interpreting results the researcher should be aware of the possible effects due to violation of these assumptions.

8.6 STEPWISE DISCRIMINANT ANALYSIS

Until now it was assumed that the best set of discriminator variables is known, and the known discriminator variables are used to form the discriminant function. Situations

do arise when a number of potential discriminator variables are known, but there is no indication as to which would be the best set of variables for forming the discriminant function. Stepwise discriminant analysis is a useful technique for selecting the best set of discriminating variables to form the discriminant function.

8.6.1 Stepwise Procedures

The best set of variables for forming the discriminant function can be selected using a forward, a backward, or a stepwise procedure. Each of these procedures is discussed below.

Forward Selection

In forward selection, the variable that is entered first into the discriminant function is the one that provides the most discrimination between the groups as measured by a given statistical criterion. In the next step, the variable that is entered is the one that adds the maximum amount of additional discriminating power to the discriminant function, as measured by the statistical criterion. The procedure continues until no additional variables are entered into the discriminant function.

Backward Selection

The backward selection begins with all the variables in the discriminant function. At each step, one variable is removed, that one being the one that provides the least amount of decrease in the discriminating power, as measured by the statistical criterion. The procedure continues until no more variables can be removed.

Stepwise Selection

Stepwise selection is a combination of the forward and backward elimination procedures. It begins with no variables in the discriminant function, then at each step a variable is either added or removed. A variable already in the discriminant function is removed if it does not significantly lower the discriminating power, as measured by the statistical criterion. If no variable is removed at a given step then the variable that significantly adds the most discriminating power, as measured by the statistical criterion, is added to the discriminant function. The procedure stops when at a given step no variable is added or removed from the discriminant function.

Each of the three procedures gives the same discriminant function if the variables are not correlated among themselves. However, the results could be very different if there is a substantial amount of multicollinearity in the data. Consequently, the researcher should exercise caution in the use of stepwise procedures if a substantial amount of multicollinearity is suspected in the data. The problem of multicollinearity in discriminant analysis is similar to that in the case of multiple regression analysis. The problem of multicollinearity and its effect on the results is further discussed in Section 8.6.4.

8.6.2 Selection Criteria

As mentioned previously, a statistical criterion is used for determining the addition or removal of variables in discriminant function. A number of criteria have been suggested. A discussion of commonly used criteria follows.

Wilks' Λ

Wilks' Λ is the ratio of the within-group sum of squares to the total sum of squares. At each step the variable that is included is the one with the smallest Wilks' Λ after the effect of variables already in the discriminant function is removed or *partialled out*. Since the Wilks' Λ can be approximated by the F-ratio, the rule is tantamount to entering the variable that has the highest partial F-ratio. Because Wilks' Λ is equal to

$$\Lambda = \frac{SS_w}{SS_t} = \frac{SS_w}{SS_b + SS_w},$$

minimizing Wilks' Λ implies that the within-group sum of squares is minimized and the between-groups sum of squares is maximized. That is, the Wilks' Λ selection criterion considers between-groups separation and within-group homogeneity.

Rao's V

Rao's V is based on the Mahalanobis distance, and concentrates on separation between the groups, as measured by the distance of the centroid of each group from the centroid of the total sample. Rao's V and the change in it while adding or deleting a variable can be approximated as a χ^2 statistic that follows a χ^2 distribution. However, although it maximizes between-groups separation, Rao's V does not take into consideration group homogeneity. Therefore, the use of Rao's V may produce a discriminant function that does not have maximum within-group homogeneity.

Mahalanobis Squared Distance

Wilks' Λ and Rao's V maximize the *total* separation among all the groups. In the case of more than two groups, the result could be that all pairs of groups may not have an optimal separation. The Mahalanobis squared distance tries to ensure that there is separation among all pairs of groups. At each step the procedure enters (removes) the variable that provides the maximum increase (minimum decrease) in separation, as measured by the Mahalanobis squared distance, between the pairs of groups that are closest to each other.

Between-Groups F-ratio

In computing the Mahalanobis distance all groups are given equal weight. To overcome this limitation, the Mahalanobis distance is converted into an F-ratio. The formula used to compute the F-ratio takes into account the sizes of the groups such that larger groups receive more weight than smaller groups. The between-groups F-ratio measures the separation between a given pair of groups.

Each of the above-mentioned criteria may result in a discriminant function with a different subset of the potential discriminating variables. Although there are no absolute rules regarding the best statistical criterion, the researcher should be aware of the objectives of the various criteria in selecting the criterion for performing stepwise discriminant analysis. Wilks' Λ is the most commonly used statistical criterion.

8.6.3 Cutoff Values for Selection Criteria

As discussed above, the stepwise procedures select a variable such that there is an increase in discriminating power of the function. Therefore, a cutoff value needs to be specified, below which the discriminant power, as measured by the statistical criterion,

is considered to be insignificant. That is, minimum conditions for selection of variables need to be specified.

If the objective is to include only those variables that improve the discriminating power of the discriminant function at a given significance level, then the normal procedure is to specify a significance level (e.g., $p = 0.05$) for inclusion and/or removal of variables. However, a much lower significance level than that desired should be specified because the overall probability of rejecting the inclusion or deletion of any given variable may be much less than 0.05, as such tests are performed for many variables. For example, if 10 independent hypotheses are tested, each at a significance level of .05, the overall significance level (i.e., probability of Type I error) is 0.40 (i.e., $1 - .95^{10}$).

If the objective is to maximize *total* discriminating power of the discriminant function irrespective of how small the discriminating power of each variable is, then a moderate significance level should be specified. Costanza and Affifi (1979) recommend a p-value between .1 and .25.

8.6.4 Stepwise Discriminant Analysis Using SPSS

Consider the case of most- and least-admired firms. Assume that in addition to *EBITASS* and *ROTC*, the following financial ratios for the firms are also available: *ROE*, return on equity; *REASS*, return on assets; and *MKTBOOK*, market to book value. Table 8.9 gives the five financial ratios for the 24 firms. Our objective is to select the *best set* of discriminating variables for forming the discriminant function. As

Table 8.9 **Financial Data for Most-Admired and Least-Admired Firms**

Firm	Group	*MKTBOOK*	*ROTC*	*ROE*	*REASS*	*EBITASS*
1	1	2.304	0.182	0.191	0.377	0.158
2	1	2.703	0.206	0.205	0.469	0.210
3	1	2.385	0.188	0.182	0.581	0.207
4	1	5.981	0.236	0.258	0.491	0.280
5	1	2.762	0.193	0.178	0.587	0.197
6	1	2.984	0.173	0.178	0.546	0.227
7	1	2.070	0.196	0.178	0.443	0.148
8	1	2.762	0.212	0.219	0.472	0.254
9	1	1.345	0.147	0.148	0.297	0.079
10	1	1.716	0.128	0.118	0.597	0.149
11	1	3.000	0.150	0.157	0.530	0.200
12	1	3.006	0.191	0.194	0.575	0.187
13	2	0.975	−0.031	−0.280	0.105	−0.012
14	2	0.945	0.053	0.019	0.306	0.036
15	2	0.270	0.036	0.012	0.269	0.038
16	2	0.739	−0.074	−0.150	0.204	−0.063
17	2	0.833	−0.119	−0.358	0.155	−0.054
18	2	0.716	−0.005	−0.305	0.027	0.000
19	2	0.574	0.039	−0.042	0.268	0.005
20	2	0.800	0.122	0.080	0.339	0.091
21	2	2.028	−0.072	−0.836	−0.185	−0.036
22	2	1.225	0.064	−0.430	−0.057	0.045
23	2	1.502	−0.024	−0.545	−0.050	−0.026
24	2	0.714	0.026	−0.110	0.021	0.016

Table 8.10 **SPSS Commands for Stepwise Discriminant Analysis**

```
DISCRIMINANT GROUPS=EXCELL(1,2)
 /VARIABLES=MKTBOOK ROTC ROE REASS EBITASS
 /ANALYSIS=MKTBOOK ROTC ROE REASS EBITASS
 /METHOD=WILKS
 /PIN=.15
 /POUT=.15
 /STATISTICS=ALL
 /PLOT=ALL
```

mentioned previously, only variables that meet a given statistical criterion are selected to form the discriminant function. A stepwise discriminant analysis will be employed with the following criteria:

1. Wilks' Λ is used as the selection criterion. That is, at each step either a variable is added or deleted from the discriminant function according to the value of Wilks' Λ.
2. A tolerance level of .001 is used.
3. Priors are assumed to be equal.
4. A .15 probability level is used for entering and removing of variables.

The necessary SPSS commands for a stepwise discriminant analysis are given in Table 8.10. The PIN and POUT subcommands specify the significance levels that should be used, respectively, for entering and removing variables. Exhibit 8.3 gives the relevant part of the resulting output. Once again, the discussion is keyed to the circled numbers in the output.

The output gives the means of all the variables along with the necessary statistics for univariate significance tests [1, 2]. As can be seen, the means of all the variables are significantly different for the two groups of firms [2].

For each variable not included in the discriminant function, the Wilks' Λ and its significance level are also printed [3]. Since the significance level and the tolerance value of all variables are above the cutoff value, all the variables are candidates for inclusion in the discriminant function. In the first step, *EBITASS* is included in the discriminant function because it provides the maximum discrimination as evidenced by the selection criterion, Wilks' Λ. Also the corresponding F-ratio for evaluating the overall discriminant function is provided [4a]. An F-value of 87.408 is significant ($p < .0000$), indicating that the discriminant function is statistically significant [4a]. Statistics for evaluating the extent to which the discriminant function formed so far differentiates between pairs of groups are also provided. The value of 87.408 [4d] for the F-statistic indicates that the difference between group 1 (most-admired) and group 2 (least-admired) with respect to the discriminant score is statistically significant ($p < .0000$) [4d]. Since there are only two groups, the F-statistic measuring the separation of groups 1 and 2 is the same as the F-value of 87.408 for the overall discriminant function.[8]

[8]This is not the case when there are more than two groups. For example, in the case of three groups there are three pairwise F-ratios for comparing differences between groups 1 and 2, groups 1 and 3, and groups 2 and 3. The equivalent F-ratio for the Wilks' Λ of the overall discriminant function measures the overall significance of *all the groups and not pairs of groups*.

Exhibit 8.3

Stepwise discriminant analysis

(1) Group means

EXCELL	MKTBOOK	ROTC	ROE	REASS	EBITASS
1	2.75150	.18350	.18383	.49708	.19133
2	.94342	.00125	-.24542	.11683	.00333
Total	1.84746	.09238	-.03079	.30696	.09733

(2) Wilks' Lambda (U-statistic) and univariate F-ratio with 1 and 22 degrees of freedom

Variable	Wilks' Lambda	F	Significance
MKTBOOK	.46135	25.6866	.0000
ROTC	.23638	71.0699	.0000
ROE	.42427	29.8532	.0000
REASS	.31570	47.6858	.0000
EBITASS	.20108	87.4076	.0000

---------- Variables not in the Analysis after Step 0 ----------

(3)

Variable	Tolerance	Minimum Tolerance	Signif. of F to Enter	Wilks' Lambda
MKTBOOK	1.0000000	1.0000000	.0000447	.4613459
ROTC	1.0000000	1.0000000	.0000000	.2363816
ROE	1.0000000	1.0000000	.0000173	.4242745
REASS	1.0000000	1.0000000	.0000006	.3157028
EBITASS	1.0000000	1.0000000	.0000000	.2010830

(4a) At step 1, EBITASS was included in the analysis.

		Degrees of Freedom		Signif.	Between Groups
Wilks' Lambda	.20108	1	1 22.0		
Equivalent F	87.40757		1 22.0	.0000	

(4b) ---------- Variables in the Analysis after Step 1 ----------

Variable	Tolerance	Signif. of F to Remove	Wilks' Lambda
EBITASS	1.0000000	.0000	

(4c) ---------- Variables not in the Analysis after Step 1 ----------

Variable	Tolerance	Minimum Tolerance	Signif. of F to Enter	Wilks' Lambda
MKTBOOK	.7541675	.7541675	.8293020	.2006277
ROTC	.3920789	.3920789	.4336968	.1951621
ROE	.8187493	.8187493	.4804895	.1962610
REASS	.8453627	.8453627	.1388271	.1807110

(continued)

Exhibit 8.3 (continued)

```
F statistics and significances between pairs of groups after step 1
Each F statistic has 1 and 22 degrees of freedom.

(4d)                    Group        1
   Group
       2                         87.4076
                                   .0000

(5a) At step 2, REASS    was included in the analysis.

                                Degrees of Freedom  Signif. Between Groups
Wilks' Lambda    .18071    2     1        22.0
Equivalent F   47.60384          2        21.0    .0000

(5b) ------------- Variables in the Analysis after Step 2 -------------

                        Signif. of
Variable  Tolerance  F to Remove   Wilks' Lambda
REASS      .8453627       .1388       .2010830
EBITASS    .8453627       .0007       .3157028

------------- Variables not in the Analysis after Step 2 -------------

                        Minimum   Signif. of
Variable  Tolerance  Tolerance  F to Enter    Wilks' Lambda
MKTBOOK    .6162599   .5323841       .3804981      .1737252
ROTC       .3918749   .3690001       .4882375      .1763149
ROE        .3722500   .3722500       .5727441      .1777879

F statistics and significances between pairs of groups after step 2
Each F statistic has 2 and 21 degrees of freedom.

                        Group        1
   Group
       2                         47.6038
                                   .0000
F level or tolerance or VIN insufficient for further computation.

(6)                                Summary Table

          Action       Vars  Wilks'
Step Entered Removed    in   Lambda   Sig.  Label
  1  EBITASS            1   .20108  .0000
  2  REASS              2   .18071  .0000

Classification function coefficients
(Fisher's linear discriminant functions)

EXCELL  =          1              2
REASS          18.9244282      7.3632759
EBITASS        58.4844420     -6.5551023
(Constant)    -10.9916677     -1.1123600
```

(continued)

Exhibit 8.3 (continued)

```
                    Canonical Discriminant Functions

                  Pct of   Cum  Canonical  After  Wilks'
  Fcn Eigenvalue Variance  Pct     Corr      Fcn  Lambda  Chi-square  df  Sig

                                         :    0 .180711     35.928     2  .0000
   1*    4.5337  100.00  100.00    .9051 :

    * Marks the 1 canonical discriminant functions remaining in the analysis.

 Standardized canonical discriminant function coefficients

              Func  1
 REASS         .38246
 EBITASS       .78573

(7) Structure matrix:

 Pooled within-groups correlations between discriminating variables
                                  and canonical discriminant functions
 (Variables ordered by size of correlation within function)

              Func  1
 EBITASS       .93613
 ROTC          .73492
 REASS         .69144
 ROE           .63352
 MKTBOOK       .33356

 Classification results -

(8)                        No. of    Predicted Group Membership
   Actual Group             Cases          1          2
 --------------------      ------    --------   --------
 Group       1               12         11          1
                                      91.7%       8.3%

 Group       2               12          0         12
                                       .0%     100.0%
 Percent of "grouped" cases correctly classified:  95.83%
```

The output also gives the significance level of the partial F-ratio and the tolerance for variables that form the discriminant function [4b]. If the p-value of the partial F-ratio or the tolerance level of any variable does not meet specified cutoff values, then that variable is removed from the function. Since this is not the case, no variables are removed at this step. A list of variables, along with the p-values of the partial F-ratios and tolerance levels, that have not been included in the discriminant function is provided next [4c]. This table is used to determine which variable will enter the discriminant function in the next step.

Based on the tolerance level and the partial F-ratios of variables that are not in the discriminant function, the variable *REASS* is entered in Step 2 [5a]. After Step 2, none of the variables can be removed from the function, and of the variables that are not in the function none are candidates for inclusion in the function [5b]. Consequently, the stepwise procedure is terminated, with the final discriminant function composed of variables *REASS* and *EBITASS*. The output also gives the summary of all the steps

and other pertinent statistics for evaluating the final discriminant function [6]. These statistics have been discussed previously.

A few remarks regarding the structure matrix [7] are in order. The structure matrix gives the loadings or the correlations between original variables and the discriminant score; SPSS gives the loadings for all the potential variables. One would have expected the loadings for the selected variables *EBITASS* and *REASS* to be the highest, but this is not the case. The loading of *ROTC* is higher than that of *REASS* due to the presence of multicollinearity in the data. This point is further discussed in the next section.

The summary of classification results indicates that one firm belonging to group 1 has been misclassified into group 2, giving an overall classification rate of 95.83% [8].

Multicollinearity and Stepwise Discriminant Analysis

Based on the univariate Wilks' Λ's and the equivalent F-ratios given in Exhibit 8.3 [2], it appears that *EBITASS* and *ROTC* would provide the best discrimination because they have the lowest values for Wilks' Λ. However, stepwise discriminant analysis did not select *ROTC*.[9] Why? The reason is that the variables are correlated among themselves. In other words, there is multicollinearity in the data. Table 8.11 gives the correlation matrix among the independent variables. It is clear from this table that the correlations among the independent variables are substantial. For example, the correlation between *ROTC* and *EBITASS* is .951, suggesting that one of the two ratios is redundant. Consequently, it may not be necessary to include both variables in the discriminant function. This does not imply that just because *ROTC* is not included in the discriminant function it is not an important variable. All that is being implied is that one of them is redundant.

Also, the correlation between *EBITASS* and *REASS* is .839, implying that these two variables have a lot in common, and one variable may dominate the other or might be suppressed by the other. Therefore, just because the standardized coefficient of *EBITASS* is much larger than that of *REASS* does not mean that *EBITASS* is more important than *REASS*.

The important points to remember from the preceding discussion are that in the presence of multicollinearity:

1. The selection of discriminator variables is affected by multicollinearity. Just because a given variable is not included does not imply that it is not important. It may very well be that this variable is important and *does* discriminate between the groups, but is not included due to its high correlation with other variables.

Table 8.11 **Correlation Matrix for Discriminating Variables**

Variable	*MKTBOOK*	*ROTC*	*ROE*	*REASS*	*EBITASS*
MKTBOOK	1.000				
ROTC	.691	1.000			
ROE	.458	.848	1.000		
REASS	.551	.810	.914	1.000	
EBITASS	.807	.951	.802	.839	1.000

[9]Note that the variables used to form the discriminant function in Eq. 8.1 were *EBITASS* and *ROTC*.

2. Use of standardized coefficients or other measures to determine the importance of variables in the discriminant function is not appropriate. It is quite possible, as in our example, that one variable may lessen the importance of variables with which it is correlated. Consequently, it is recommended that statements concerning the importance of each variable should not be made when multicollinearity is present in the data.
3. In the presence of multicollinearity in the data, stepwise discriminant analysis may or may not be appropriate depending on the source of multicollinearity. In a *population-based multicollinearity* the pattern of correlations among the independent variables, within sampling errors, is the same from sample to sample. In such a case, use of stepwise discriminant analysis is appropriate, as the relationship among variables is a population characteristic and the results will not change from sample to sample. On the other hand, the results of stepwise discriminant analysis could differ from sample to sample if the pattern of correlations among the independent variables varies across samples. This is called *sample-based multicollinearity*. In such a case stepwise discriminant analysis is not appropriate.

8.7 EXTERNAL VALIDATION OF THE DISCRIMINANT FUNCTION

If discriminant analysis is used for classifying observations then the external validity needs to be examined. External validity refers to the accuracy with which the discriminant function can classify observations that are from another sample. The error rate obtained by classifying observations that have also been used to estimate the discriminant function is biased and, therefore, should not be used as a measure for validating the discriminant function. Following are three suggested techniques that are commonly employed to validate the discriminant function.

8.7.1 Holdout Method

The sample is randomly divided into two groups. The discriminant function is estimated using one group and the function is then used to classify observations of the second group. The result will be an unbiased estimate of the classification rate. One could also use the second group to estimate the discriminant function and classify observations in the first group. This is referred to as double cross-validation. Obviously this technique requires a large sample size. The **SPSS** commands to implement the holdout method are given in Table 8.12. The first COMPUTE command creates a new variable *SAMP* whose values come from a uniform distribution and range between 0 and 1. The second COMPUTE command rounds the values of *SAMP* to the nearest whole integer. Consequently, *SAMP* will have a value of 0 or 1. The SELECT subcommand requests discriminant analysis using only those observations whose values for the *SAMP* variable are equal to 1. Classification is done separately for observations whose values for the *SAMP* variable are 0 and 1.

8.7.2 U-Method

The U-method, proposed by Lachenbruch (1967), holds out one observation at a time, estimates the discriminant function using the remaining $n - 1$ observations, and classifies

Table 8.12 **SPSS Commands for Holdout Validation**

```
SET WIDTH=80
TITLE VALIDATION USING HOLDOUT METHOD
 /MKTBOOK ROTC ROE REASS EBITASS EXCELL
COMPUTE SAMP=UNIFORM(1)
COMPUTE SAMP=RND(SAMP)
BEGIN DATA
insert data here
END DATA
DISCRIMINANT GROUPS=EXCELL(1,2)
 /VARIABLES=EBITASS ROTC
 /SELECT=SAMP(1)
 /ANALYSIS=EBITASS ROTC
 /METHOD=DIRECT
 /STATISTICS=ALL
 /PLOT=ALL
FINISH
```

the held-out observation. This is equivalent to running n discriminant analyses and classifying the n held-out observations. This method gives an almost unbiased estimate of the classification rate. However, the U-method has been criticized in that it does not provide an error rate that has the smallest variance or mean-squared error (see Glick 1978; McLachlan 1974; Toussant 1974).

8.7.3 Bootstrap Method

Bootstrapping is a procedure where repeated samples are drawn from the sample, discriminant analysis is conducted on the samples drawn, and an error rate is computed. The overall error rate and its sampling distribution are obtained from the error rates of the repeated samples that are drawn. Bootstrapping techniques require a considerable amount of computer time. However, with the advent of fast and cheap computing, they are gaining popularity as a viable procedure for obtaining sampling distributions of statistics whose theoretical sampling distributions are not known. See Efron (1987) for an in-depth discussion of various bootstrapping techniques and see Bone, Sharma, and Shimp (1989) for their application in covariance structure analysis.

8.8 SUMMARY

This chapter discussed two-group discriminant analysis. Discriminant analysis is a technique for first identifying the "best" set of variables, known as the discriminator variables, that provide the best discrimination between the two groups. Then a discriminant function is estimated, which is a linear combination of the discriminator variables. The values resulting from the discriminant function are known as discriminant scores. The discriminant function is estimated such that the ratio of the between-groups sum of squares to the within-group sum of squares for the discriminant scores is maximum. The final objective of discriminant analysis is to classify future observations into one of the two groups, based on the values of their discriminant scores. In the next chapter we discuss discriminant analysis for more than two groups. This is known as multiple-group discriminant analysis.

QUESTIONS

8.1 Tables Q8.1 (a), (b), and (c) present data on two variables, X_1 and X_2. In each case plot the data in two-dimensional space and comment on the following:

(i) Discrimination provided by X_1
(ii) Discrimination provided by X_2
(iii) Joint discrimination provided by X_1 and X_2.

Table Q8.1(a)

Observation	X_1	X_2
1	5	8
2	6	7
3	5	2
4	6	3
5	6	8
6	5	4

Table Q8.1(b)

Observation	X_1	X_2
1	7	4
2	3	3
3	6	3
4	2	4
5	8	5
6	1	3

Table Q8.1(c)

Observation	X_1	X_2
1	2	3
2	6	8
3	1	2
4	5	8
5	3	3
6	7	6

8.2 In a consumer survey 200 respondents evaluated the taste and health benefits of 20 brands of breakfast cereals. Table Q8.2 presents the average ratings on 4 brands that consumers rated *most likely to buy* and 4 brands that they rated *least likely to buy.*

Table Q8.2

Brand	Overall Rating	Taste Rating	Health Rating
A	0	5	3
B	1	6	8
C	1	7	9
D	0	3	2
E	1	7	7
F	0	4	6
G	0	3	3
H	1	8	9

Notes:
1. Overall rating: 0 = least likely to buy; 1 = most likely to buy.
2. Taste and health attributes of the cereals were rated on a 10-point scale with 1 = extremely poor and 10 = extremely good.

(a) Plot the data in two-dimensional space and comment on the discrimination provided by: (i) taste, (ii) health, and (iii) taste and health.
(b) Consider a new axis in two-dimensional space that makes an angle θ with the "taste" axis. Derive an equation to compute the projection of the points on the new axis.
(c) Calculate the between-groups, within-group, and total sums of squares for various values of θ.
(d) Define λ as the ratio of the between-groups to within-group sum of squares. Calculate λ for each value of θ.
(e) Plot the values of λ against those of θ. Use this plot to find the value of θ that gives the maximum value of λ.
(f) Use the λ-maximizing value of θ from (e) to derive the linear discriminant function.
Hint: A spreadsheet may be used to facilitate the calculations considerably.

8.3 Consider Question 8.2.

(a) Use the discriminant function derived in Question 8.2(f) to calculate the discriminant scores for each brand of cereal.
(b) Plot the discriminant scores in one-dimensional space.
(c) Use the plot of the discriminant scores to obtain a suitable cutoff value.
(d) Comment on the accuracy of classification provided by the discriminant function and the cutoff value.

8.4 Use SPSS (or any other software) to perform discriminant analysis on the data shown in Table Q8.2. Compare the results with those obtained in Questions 8.2 and 8.3.

8.5 Cluster analysis and discriminant analysis can both be used for purposes of classification. Discuss the similarities and differences in the two approaches. Under what circumstances is each approach more appropriate?

8.6 Two-group discriminant analysis and multiple regression share a lot of similarities, and are in fact computationally indistinguishable. However, there are some fundamental differences between these two approaches. Discuss the differences between these two methods.

8.7 File DEPRES.DAT gives data from a study in which the respondents were adult residents of Los Angeles County.[a] The major objectives of the study were to provide estimates of the prevalence and incidence of depression and to identify causal factors and outcomes associated with this condition. The major instrument used for classifying depression is the Depression Index (CESD) of the National Institute of Mental Health, Center for Epidemiologic Studies. A description of the relevant variables is provided in file DEPRES.DOC.

(a) Use discriminant analysis to estimate whether an individual is likely to be depressed (use the variable "CASES" as the dependent variable) based on his/her income and education.
(b) Include the variables ACUTEILL, SEX, AGE, HEALTH, BEDDAYS, and CHRONILL as additional independent variables and check whether there is an improvement in the prediction.
Hint: Since the number of independent variables is large you may want to use stepwise discriminant analysis.
(c) Interpret the discriminant analysis solution.

8.8 Refer to the mass transportation data in file MASST.DAT (for a description of the data refer to file MASST.DOC). Use the cluster analysis results from Question 7.3 (Chapter 7) to form two groups: "users" and "nonusers" of mass transportation. Using this new grouping as the dependent variable and determinant attributes of mass transportation as independent variables, perform a discriminant analysis and discuss the differences between the two groups.
Note: Use variables ID, V_1 to V_{18}, for this question. Ignore the other variables.

8.9 For the two groups case show that:

$$\mathbf{B} = \frac{n_1 n_2}{(n_1 + n_2)}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)',$$

where $\mathbf{B}$ = between-groups SSCP matrix for p variables, $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the $p \times 1$ vectors of means for group 1 and group 2, and n_1 and n_2 are the number of observations in group 1 and group 2.

[a]Afifi, A.A., and Virginia Clark (1984), *Computer Aided Multivariate Analysis*, Lifetime Learning Publications, Belmont, California, pp.30–39.

Appendix

A8.1 FISHER'S LINEAR DISCRIMINANT FUNCTION

Let $\mathbf{X}$ be a $p \times 1$ random vector of p variables whose variance–covariance matrix is given by $\boldsymbol{\Sigma}$ and the total SSCP matrix by $\mathbf{T}$. Let $\boldsymbol{\gamma}$ be a $p \times 1$ vector of weights. The discriminant function will be given by

$$\xi = \mathbf{X}'\boldsymbol{\gamma}. \tag{A8.1}$$

The sum of squares for the resulting discriminant scores will be given by

$$\begin{aligned} \xi'\xi &= (\mathbf{X}'\boldsymbol{\gamma})'(\mathbf{X}'\boldsymbol{\gamma}) \\ &= \boldsymbol{\gamma}'\mathbf{X}\mathbf{X}'\boldsymbol{\gamma} \\ &= \boldsymbol{\gamma}'\mathbf{T}\boldsymbol{\gamma} \end{aligned} \tag{A8.2}$$

where $\mathbf{T} = \mathbf{X}\mathbf{X}'$ and is the total SSCP matrix for the p variables. Since

$$\mathbf{T} = \mathbf{B} + \mathbf{W},$$

where $\mathbf{B}$ and $\mathbf{W}$ are, respectively, between-groups and within-group SSCP matrices for the p variables, Eq. A8.2 can be written as

$$\begin{aligned} \xi'\xi &= \boldsymbol{\gamma}'(\mathbf{B} + \mathbf{W})\boldsymbol{\gamma} \\ &= \boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma} + \boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma}. \end{aligned} \tag{A8.3}$$

In Eq. A8.3, $\boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma}$ and $\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma}$ are, respectively, the between-groups and within-group sum of squares for the discriminant score ξ.

The objective of discriminant analysis is to estimate the weight vector, $\boldsymbol{\gamma}$, of the discriminant function given by Eq. A8.1 such that

$$\lambda = \frac{\boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma}}{\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma}} \tag{A8.4}$$

is maximized. The vector of weights, $\boldsymbol{\gamma}$, can be obtained by differentiating λ with respect to $\boldsymbol{\gamma}$, and equating to zero. That is

$$\frac{\partial\lambda}{\partial\boldsymbol{\gamma}} = \frac{2(\mathbf{B}\boldsymbol{\gamma})(\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma}) - 2(\boldsymbol{\gamma}'\mathbf{B}\boldsymbol{\gamma})(\mathbf{W}\boldsymbol{\gamma})}{(\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma})^2} = 0.$$

Or, dividing through by $\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma}$,

$$\begin{aligned} \frac{2(\mathbf{B}\boldsymbol{\gamma} - \lambda\mathbf{W}\boldsymbol{\gamma})}{\boldsymbol{\gamma}'\mathbf{W}\boldsymbol{\gamma}} &= \mathbf{0} \\ (\mathbf{B} - \lambda\mathbf{W})\boldsymbol{\gamma} &= \mathbf{0} \\ (\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I})\boldsymbol{\gamma} &= \mathbf{0}. \end{aligned} \tag{A8.5}$$

Equation A8.5 is a system of homogeneous equations and for a nontrivial solution

$$|\mathbf{W}^{-1}\mathbf{B} - \lambda\mathbf{I}| = \mathbf{0}. \tag{A8.6}$$

That is, the problem reduces to finding eigenvalues and eigenvectors of the nonsymmetric matrix, $\mathbf{W}^{-1}\mathbf{B}$, with the eigenvectors giving the weight matrix for forming the discriminant function. For the two-groups case, Eq. A8.5 can be further simplified. It can be shown that for two groups $\mathbf{B}$

is equal to

$$\begin{aligned} \mathbf{B} &= \frac{n_1 n_2}{n_1 + n_2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \\ &= C(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \end{aligned} \tag{A8.7}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, respectively, are $p \times 1$ vectors of means for group 1 and group 2; n_1 and n_2 are number of observations in group 1 and group 2, respectively; and $C = n_1 n_2/(n_1 + n_2)$, which is a constant.

Therefore, Eq. A8.5 can be written as

$$\begin{aligned} [\mathbf{W}^{-1}C(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' - \lambda\mathbf{I}]\boldsymbol{\gamma} &= 0 \\ C\mathbf{W}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\gamma} &= \lambda\boldsymbol{\gamma} \\ \frac{C}{\lambda}[\mathbf{W}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\gamma}] &= \boldsymbol{\gamma}. \end{aligned} \tag{A8.8}$$

Since $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\gamma}$ is a scalar, Eq. A8.8 can be written as

$$\boldsymbol{\gamma} = K\mathbf{W}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{A8.9}$$

where $K = C(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\gamma}/\lambda$ is a scalar and therefore is a constant. Since the within-group variance–covariance matrix, $\boldsymbol{\Sigma}_w$, is proportional to $\mathbf{W}$ and it is assumed that $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_w = \boldsymbol{\Sigma}$, Eq. A8.9 can also be written as

$$\boldsymbol{\gamma} = K\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2). \tag{A8.10}$$

Assuming a value of one for the constant K, Eq. A8.10 can be written as

$$\boldsymbol{\gamma} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

or

$$\boldsymbol{\gamma}' = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{\Sigma}^{-1}. \tag{A8.11}$$

The discriminant function given by Eq. A8.11 is the Fisher's discriminant function. It is obvious that different values of the constant K give different values for $\boldsymbol{\gamma}$ and therefore the absolute weights of the discriminant function are not unique. The weights are unique only in a relative sense; that is, only the ratio of the weights is unique.

A8.2 CLASSIFICATION

Geometrically, classification involves the partitioning of the discriminant or the variable space into two mutually exclusive regions. Figure A8.1 gives a hypothetical plot of two groups measured on only one variable, X. The classification problem reduces to determining a cutoff value that divides the space into two regions, $R1$ and $R2$. Observations falling in region $R1$ are clas-

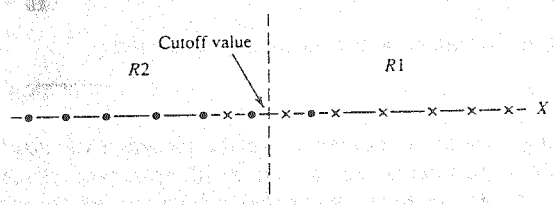


Figure A8.1 Classification in one-dimensional space.

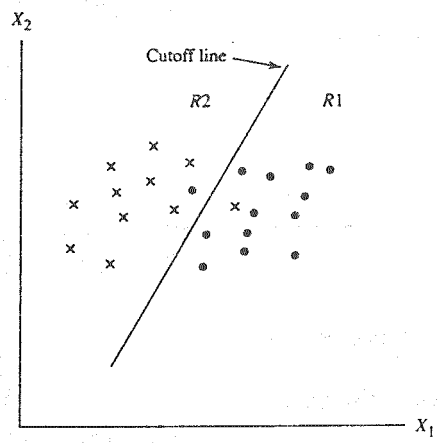


Figure A8.2 Classification in two-dimensional space.

sified into group 1 and those falling in region $R2$ are classified into group 2. Similarly, Figure A8.2 gives a hypothetical plot of observations measured on two variables, X_1 and X_2. The classification problem now reduces to identifying a line that divides the two-dimensional space into two regions. If the observations are measured on three variables then the classification problem reduces to identifying a plane that divides the three-dimensional space into two regions. In general, for a p-dimensional space the problem reduces to finding a $(p-1)$-dimensional hyperplane that divides the p-dimensional space into two regions.

In classifying observations, two types of errors can occur, as given in Table A8.1. An observation coming from group 1 can be misclassified into group 2. Let $c(2/1)$ be the cost of this misclassification. Similarly, $c(1/2)$ is the cost of misclassifying an observation from group 2 into group 1. Obviously, one would prefer to use the criterion that minimizes misclassification costs. The following section discusses the statistical decision theory method for dividing the total space into the classification regions, i.e., developing classification rules.

A8.2.1 Statistical Decision Theory Method for Developing Classification Rules

Consider the case where we have only one discriminating variable, X. Let π_1 and π_2 represent the populations for the two groups and $f_1(x)$ and $f_2(x)$ be the respective probability density functions for the 1×1 random vector, X. Figure A8.3 depicts the two density functions and the cutoff value, c, for one discriminating variable (i.e., $p = 1$). The conditional probability of correctly classifying observations in group 1 is given by

$$P(1/1) = \int_{-\infty}^{c} f_1(x)dx,$$

Table A8.1 **Misclassification Costs**

Group Assignment by Decision Maker	Actual Group Membership 1	Actual Group Membership 2
1	No cost	C(1/2)
2	C(2/1)	No cost

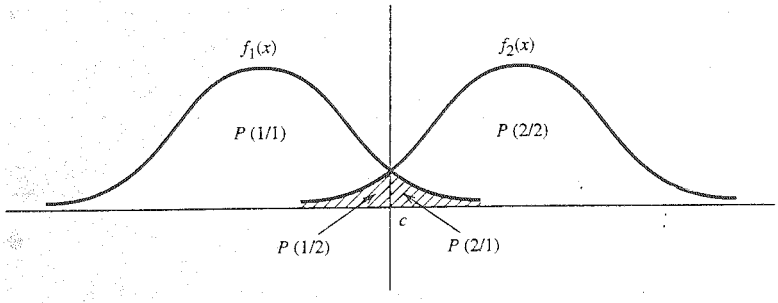


Figure A8.3 Density functions for one discriminating variable.

and that of correctly classifying observations in group 2 is given by

$$P(2/2) = \int_c^{\infty} f_2(x)dx,$$

where $P(i/j)$ is the conditional probability of classifying observations to group i given that they belong to group j.

The conditional probability of misclassification is given by

$$P(1/2) = \int_{-\infty}^{c} f_2(x)dx \tag{A8.12}$$

and

$$P(2/1) = \int_c^{\infty} f_1(x)dx. \tag{A8.13}$$

Now assume that the prior probabilities of a given observation belonging to group 1 and group 2, respectively, are p_1 and p_2. The various classification probabilities are given by

$$P(\text{correctly classified into } \pi_1) = P(1/1)p_1$$
$$P(\text{correctly classified into } \pi_2) = P(2/2)p_2$$
$$P(\text{misclassified into } \pi_1) = P(1/2)p_2 \tag{A8.14}$$
$$P(\text{misclassified into } \pi_2) = P(2/1)p_1. \tag{A8.15}$$

The *total probability of misclassification* (TPM) is given by the sum of Eqs. A8.14 and A8.15. That is,

$$TPM = P(1/2)p_2 + P(2/1)p_1. \tag{A8.16}$$

Equation A8.16 does not consider the misclassification costs. Given the misclassification costs specified in Table A8.1, the *total cost of misclassification* (TCM) is then given by

$$TCM = C(1/2)P(1/2)p_2 + C(2/1)P(2/1)p_1. \tag{A8.17}$$

Classification rules can be obtained by minimizing either Eq. A8.16 or Eq. A8.17. In other words, the value c in Figure A8.3 should be chosen such that either TPM or TCM is minimized. If Eq. A8.16 is minimized then the resulting classification rule assumes equal costs of misclassification. Minimization of Eq. A8.17 will result in a classification rule that assumes unequal priors and unequal misclassification costs. Consequently, minimization of Eq. A8.17 results in a more general classification rule. It can be shown that the rule which minimizes Eq. A8.17 is given by: Assign an observation to π_1 if

$$\frac{f_1(x)}{f_2(x)} \geq \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right], \tag{A8.18}$$

Table A8.2 **Summary of Classification Rules**

Priors	Misclassification Costs	Classification Rules: Assign to Group 1 (π_1) if:	Classification Rules: Assign to Group 2 (π_2) if:
Unequal	Unequal	$\frac{f_1(x)}{f_2(x)} \geq \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]$	$\frac{f_1(x)}{f_2(x)} < \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]$
Unequal	Equal	$\frac{f_1(x)}{f_2(x)} \geq \left[\frac{p_2}{p_1}\right]$	$\frac{f_1(x)}{f_2(x)} < \left[\frac{p_2}{p_1}\right]$
Equal	Unequal	$\frac{f_1(x)}{f_2(x)} \geq \left[\frac{C(1/2)}{C(2/1)}\right]$	$\frac{f_1(x)}{f_2(x)} < \left[\frac{C(1/2)}{C(2/1)}\right]$
Equal	Equal	$\frac{f_1(x)}{f_2(x)} \geq 1$	$\frac{f_1(x)}{f_2(x)} < 1$

and assign to π_2 if

$$\frac{f_1(x)}{f_2(x)} < \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]. \tag{A8.19}$$

Table A8.2 gives the classification rules for various combinations of priors and misclassification costs. These classification rules can be readily generalized to more than one variable by defining the function $f(\mathbf{x})$ as $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_p)$.

Posterior Probabilities

It is also possible to compute the posterior probabilities based on Bayesian theory. The posterior probability, $P(\pi_1/x_i)$, that any observation i belongs to group 1 given the value of variable x_i is given by

$$\begin{aligned} P(\pi_1|x_i) &= \frac{P(\pi_1 \text{ occurs and observe } x_i)}{P(\text{observe } x_i)} \\ &= \frac{P(x_i \cap \pi_1)}{P(x_i)} \\ &= \frac{P(\pi_1)P(x_i|\pi_1)}{P(x_i|\pi_1)P(\pi_1) + P(x_i|\pi_2)P(\pi_2)} \\ &= \frac{p_1 f_1(x_i)}{p_1 f_1(x_i) + p_2 f_2(x_i)}. \end{aligned} \tag{A8.20}$$

Similarly, the posterior probability, $P(\pi_2/x_i)$ is given by

$$\frac{p_2 f_2(x_i)}{p_1 f_1(x_i) + p_2 f_2(x_i)}. \tag{A8.21}$$

It is clear that for classification purposes, one must know at a minimum, the density function, $f(x)$. In the following section the above classification rules for a multivariate normal density function are developed.

A8.2.2 Classification Rules for Multivariate Normal Distributions

The joint density function of any group i for a $p \times 1$ random vector $\mathbf{x}$ is given by

$$f_i(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}. \tag{A8.22}$$

Assuming $\Sigma_1 = \Sigma_2 = \Sigma$ and substituting Eq. A8.22 in Eqs. A8.18 and A8.19 results in the following classification rules:
Assign to π_1 if

$$e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)} \geq \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right] \tag{A8.23}$$

and assign to π_2 if

$$e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_1)+\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_2)} < \left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]. \tag{A8.24}$$

Taking the natural log of both sides of Eqs. A8.23 and A8.24 and rearranging items, the rule becomes:
Assign to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\mathbf{x} \geq \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \ln\left\{\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]\right\} \tag{A8.25}$$

and assign to π_2 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\mathbf{x} < \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + \ln\left\{\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]\right\}. \tag{A8.26}$$

Further simplification of the preceding equations results in:
Assign to π_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\mathbf{x} \geq .5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{\mu}_1 + .5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{\mu}_2 + \ln\left\{\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]\right\} \tag{A8.27}$$

and assign to π_2 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\mathbf{x} < .5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{\mu}_1 + .5(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}\boldsymbol{\mu}_2 + \ln\left\{\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]\right\}. \tag{A8.28}$$

The quantity $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\Sigma^{-1}$ in Eqs. A8.27 and A8.28 is the discriminant function and, therefore, the equations can be written as:
Assign to π_1 if

$$\xi \geq .5\bar{\xi}_1 + .5\bar{\xi}_2 + \ln\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right] \tag{A8.29}$$

and assign to π_2 if

$$\xi < .5\bar{\xi}_1 + .5\bar{\xi}_2 + \ln\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right], \tag{A8.30}$$

where ξ is the discriminant score and $\bar{\xi}_j$ is the mean discriminant score for group j.

It can be clearly seen that for equal misclassification costs and equal priors, the preceding classification rules reduce to:
Assign to π_1 if

$$\xi \geq .5\bar{\xi}_1 + .5\bar{\xi}_2 \tag{A8.31}$$

and assign to π_2 if

$$\xi < .5\bar{\xi}_1 + .5\bar{\xi}_2 \tag{A8.32}$$

because $\ln(1) = 0$.

Therefore, the cutoff value used in the cutoff-value method presented in the chapter implicitly assumes a multivariate normal distribution, equal priors, equal misclassification costs, and equal covariance matrices.

Posterior Probabilities

Substituting Eq. A8.22 in Eq. A8.20 and assuming $\Sigma_1 = \Sigma_2 = \Sigma$ gives the posterior probability, $p(\pi_1/x)$, of classifying an observation in group 1. That is,

$$p(\pi_1/x) = \frac{p_1 e^{-1/2(x-\mu_1)'\Sigma^{-1}(x-\mu_1)}}{p_1 e^{-1/2(x-\mu_1)'\Sigma^{-1}(x-\mu_1)} + p_2 e^{-1/2(x-\mu_2)'\Sigma^{-1}(x-\mu_2)}}. \tag{A8.33}$$

Equation A8.33 can be further simplified as

$$\begin{aligned} p(\pi_1/x) &= \frac{1}{1 + \dfrac{p_2 e^{-1/2(x-\mu_2)'\Sigma^{-1}(x-\mu_2)}}{p_1 e^{-1/2(x-\mu_1)'\Sigma^{-1}(x-\mu_1)}}} \\ &= \frac{1}{1 + \dfrac{p_2}{p_1} e^{-1/2[(x-\mu_2)'\Sigma^{-1}(x-\mu_2)-(x-\mu_1)'\Sigma^{-1}(x-\mu_1)]}} \\ &= \frac{1}{1 + \dfrac{p_2}{p_1} e^{[-(\mu_1-\mu_2)'\Sigma^{-1}x + \frac{(\mu_1-\mu_2)'\Sigma^{-1}\mu_1}{2} + \frac{(\mu_1-\mu_2)'\Sigma^{-1}\mu_1}{2}]}} \\ &= \frac{1}{1 + \dfrac{p_2}{p_1} e^{(-\xi + \frac{\xi_1+\xi_2}{2})}} \end{aligned} \tag{A8.34}$$

as $(\mu_1 - \mu_2)'\Sigma^{-1}$ is the discriminant function.

We have so far assumed that the population parameters μ_1, μ_2, and $\Sigma_1 = \Sigma_2$ are known. Furthermore, it has been assumed that $\Sigma_1 = \Sigma_2 = \Sigma$. If the population parameters are not known then sample-based estimates $\bar{x}_j$ and $\hat{S}_w^{-1}$, respectively, of population parameters μ_j and Σ^{-1} are used. However, using the sample estimates in the classification equations does not ensure that the TCM will be minimized. But, as sample size increases the bias induced in TCM by the use of sample estimates is reduced. If $\Sigma_1 \neq \Sigma_2$ then the procedure for developing classification rules is the same, however the rules become quite cumbersome.

A8.2.3 Mahalanobis Distance Method

Once again assuming $\Sigma_1 = \Sigma_2 = \Sigma$, the Mahalanobis, or statistical distance, of any observation i from group 1 is given by

$$(x_i - \mu_1)'\Sigma^{-1}(x_i - \mu_1),$$

and from group 2 it is given by

$$(x_i - \mu_2)'\Sigma^{-1}(x_i - \mu_2).$$

That is, an observation i is assigned to group 1 if

$$(x_i - \mu_1)'\Sigma^{-1}(x_i - \mu_1) \leq (x_i - \mu_2)'\Sigma^{-1}(x_i - \mu_2) \tag{A8.35}$$

and to group 2 otherwise. Equation A8.35 can be rewritten as

$$x_i'\Sigma^{-1}(\mu_1 - \mu_2) \geq .5(\mu_1'\Sigma^{-1}\mu_1 - \mu_2'\Sigma^{-1}\mu_2)$$

or

$$(\mu_1 - \mu_2)'\Sigma^{-1}x_i \geq .5(\mu_1 - \mu_2)'\Sigma^{-1}\mu_1 + .5(\mu_1 - \mu_2)'\Sigma^{-1}\mu_2. \tag{A8.36}$$

Equation A8.36 results in the following classification rule:
Assign to π_1 if

$$\xi \geq .5\bar{\xi}_1 + .5\bar{\xi}_2 \tag{A8.37}$$

and assign to π_2 if

$$\xi < .5\bar{\xi}_1 + .5\bar{\xi}_2. \tag{A8.38}$$

As can be seen the rule given by Eqs. A8.37 and A8.38 is the same as that of the cutoff-value method. It is also the same as that given by the statistical decision theory method under the assumption of equal priors, equal misclassification costs, a multivariate normal distribution for the discriminator variables, and equal covariance matrices. The only difference between Mahalanobis distance method and the cutoff-value method employing discriminant scores is that in the former, classification regions are formed in the original variable space, and in the latter, classification regions are formed in the discriminant score space. Obviously, forming classification regions in the discriminant score space gives a more parsimonious representation of the classification problem.

A8.3 ILLUSTRATIVE EXAMPLE

We use a numerical example to illustrate some of the concepts presented in the Appendix. First, we illustrate the procedures for any known distribution and then for normally distributed variables.

A8.3.1 Any Known Distribution

Given the following information

$$C(2/1) = 5, \quad C(1/2) = 10, \quad p_1 = .8, \quad p_2 = .2, \quad f_1(x_i) = .3, \quad f_2(x_i) = .4,$$

classify the observation i.

According to Eq. A8.18 the cutoff value is equal to

$$\frac{10 \times 0.2}{5 \times 0.8} = 0.50,$$

and

$$\frac{f_1(x)}{f_2(x)} = \frac{.3}{.4} = 0.75,$$

and therefore observation i is assigned to group 1.

If the misclassification costs are equal, then the cutoff value is equal to $0.2/0.8 = 0.25$ (see Table A8.2) and the observation is once again assigned to group 1. For equal priors the cutoff value is equal to $10/5 = 2$ and the observation is assigned to group 2 (see Table A8.2). If the priors and the misclassification costs are equal, then the observation will be assigned to group 2 as the cutoff value is equal to one (see Table A8.2).

The posterior probabilities can be computed from Eqs. A8.20 and A8.21 and are equal to

$$P(\pi_1/x_i) = \frac{0.8 \times 0.3}{0.8 \times 0.3 + 0.2 \times 0.4} = 0.75,$$

and

$$P(\pi_2/x_i) = \frac{0.2 \times 0.4}{0.8 \times 0.3 + 0.2 \times 0.4} = .25.$$

Therefore, the observation will be assigned to group 1.

A8.3.2 Normal Distribution

For simplicity, assume that the one variable distribution shown in Figure A8.3 is normal. Assume further that $f_1(x) \sim N(2, 1)$, $f_2(x) \sim N(6, 1)$, $p_1 = .7$, $p_2 = .3$, $C(1/2) = 3$, and $C(2/1) = 5$.

Cutoff Value

The shaded area in Figure A8.3 gives the misclassification probabilities. Suppose that we select the cutoff value, c, to be equal to 3. The conditional probability $P(1/2)$ given by Eq. A8.12 for misclassifying an observation belonging to group 2 into group 1 can be computed from the cumulative standard unit normal distribution table. The corresponding z-value will be $(3 - 6) = -3$ and the area under the curve from $-\infty$ to -3 (which is $P(1/2)$) will be 0.0013. Similarly one can compute $P(2/1)$, which is equal to 0.1587. From Eq. A8.17, the total cost of misclassification (TCM) will be

$$\text{TCM} = 3 \times 0.0013 \times 0.3 + 5 \times 0.1587 \times 0.7 = 0.5566.$$

Table A8.3 gives $P(1/2)$, $P(2/1)$, and TCM for different values of c and Figure A8.4 gives the plot of c and TCM. As can be seen, TCM decreases and then increases for various values of c, and TCM is the lowest when c is approximately equal to 4.4. That is, a cutoff value of about 4.4 will give the lowest TCM. Table A8.4 gives the TCMs for various combinations of priors and misclassification costs and the corresponding cutoff values. It can be seen that as the priors and misclassification costs change, the cutoff value also changes. Specifically, the higher the prior for a given group the larger is its classification region and vice versa. Similarly, the higher the misclassification costs for a group, the larger is its classification region and vice versa.

Classification using the cutoff value can also be done using Eqs. A8.27 and A8.28. Substituting in Eq. A8.27 we get:
Assign to group 1 if

$$(2 - 6)x \geq .5(2 - 6) \times 2 + .5(2 - 6) \times 6 + \ln\left\{\left[\frac{C(1/2)}{C(2/1)}\right]\left[\frac{p_2}{p_1}\right]\right\}$$

$$-4x \geq -16 + \ln\left[\frac{3 \times .3}{5 \times .7}\right]$$

$$x \leq 4 - \frac{1}{4}\ln(.257)$$

$$\leq 4.340.$$

Table A8.3 **Illustrative Example**

Cutoff Value c	Conditional Probability $P(1/2)$	Conditional Probability $P(2/1)$	Total Cost of Misclassification
3.0	0.0013	0.1587	.5566
3.2	0.0026	0.1151	.4052
3.4	0.0047	0.0808	.2870
3.6	0.0082	0.0548	.1992
3.8	0.0139	0.0359	.1382
4.0	0.0228	0.0228	.1003
4.2	0.0359	0.0139	.0810
4.4	0.0548	0.0082	.0780
4.6	0.0808	0.0047	.0892
4.8	0.1151	0.0026	.1127
5.0	0.1587	0.0013	.1474

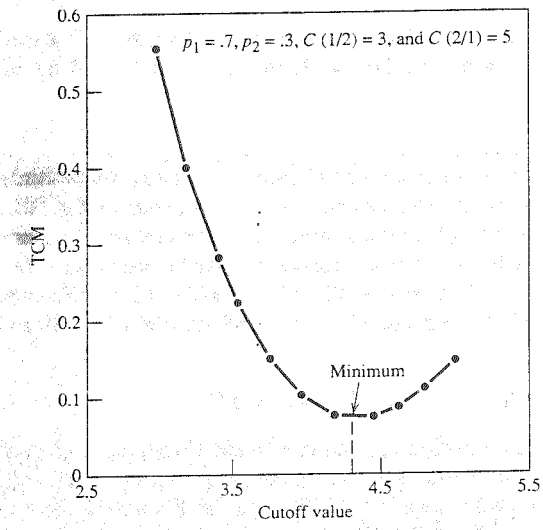


Figure A8.4 TCM as a function of cutoff value.

Table A8.4 **TCM for Various Combinations of Misclassification Costs and Priors**

Cutoff Value	Conditional Probability		Total Cost of Misclassification		
c	$P(1/2)$	$P(2/1)$	TCM_1	TCM_2	TCM_3
3.0	0.0013	0.1587	0.3344	0.3987	0.2400
3.2	0.0026	0.1151	0.2441	0.2917	0.1766
3.4	0.0047	0.0808	0.1739	0.2091	0.1283
3.6	0.0082	0.0548	0.1225	0.1493	0.0945
3.8	0.0139	0.0359	0.0879	0.1106	0.0747
4.0	0.0228	0.0228	0.0684	0.0912	0.0684
4.2	0.0359	0.0139	0.0615	0.0886	0.0747
4.4	0.0548	0.0082	0.0665	0.1027	0.0945
4.6	0.0808	0.0047	0.0826	0.1330	0.1283
4.8	0.1151	0.0026	0.1091	0.1792	0.1766
5.0	0.1587	0.0013	0.1456	0.2413	0.2400

Notes: TCM_1: $p_1 = 0.7$, $p_2 = 0.3$, $C(2/1) = 3$ and $C(1/2) = 3$.
TCM_2: $p_1 = 0.5$, $p_2 = 0.5$, $C(2/1) = 5$ and $C(1/2) = 3$.
TCM_3: $p_1 = 0.5$, $p_2 = 0.5$, $C(2/1) = 3$ and $C(1/2) = 3$.

That is, the cutoff value is equal to 4.340 which, within rounding error, is the same as that obtained previously. The assignment rule is: Assign the observation to group 1 if the value of the observation is less than or equal to 4.340 and assign to group 2 if its value is greater than 4.340.

CHAPTER 9

Multiple-Group Discriminant Analysis

In the previous chapter we discussed discriminant analysis for two groups. In many instances, however, one might be interested in discriminating among more than two groups. For example, consider the following situations:

- A marketing manager is interested in determining factors that best discriminate among heavy, medium, and light users of a given product category.
- Management of a telephone company is interested in identifying characteristics that best discriminate among households that have one, two, and three or more phone lines.
- Management of a multinational firm is interested in identifying salient attributes that differentiate successful product introductions in Latin American, European, Far Eastern, and Middle Eastern countries.

Each of these examples involves discrimination among three or more groups. Multiple-group discriminant analysis (MDA) is a suitable technique for such purposes. The objectives of MDA are the same as those for two-group discriminant analysis, except for the following difference. In the case of two-group discriminant analysis, only one discriminant function is required to represent all of the differences between the two groups. In the case of more than two groups, however, it may not be possible to represent or account for all of the differences among the groups by a single discriminant function, making it necessary to identify additional discriminant function(s). That is, an additional objective in multiple-group discriminant analysis is to identify the *minimum* number of discriminant functions that will provide most of the discrimination among the groups. The following section provides a geometrical view of MDA.

9.1 GEOMETRICAL VIEW OF MDA

The issue of identifying the set of variables that best discriminate among the groups is the same as in two-group discriminant analysis, therefore we do not provide a geometrical view of this objective. An additional objective in multiple-group discriminant analysis is to identify the number of discriminant functions needed to best represent the difference among the groups, so we begin by giving a geometrical view of this objective.

9.1.1 How Many Discriminant Functions Are Needed?

Panel I of Figure 9.1 gives the scatterplot of a hypothetical set of observations in the variable space. The observations come from four groups and are measured on two variables, X_1 and X_2. Therefore, one can have a maximum of two discriminant functions.[1] It appears that the means of the two variables, X_1 and X_2, are different across the four groups. Let Z be the axis representing the discriminant function. As discussed in Chapter 8, projection of the points onto the discriminant function, Z, gives the discriminant scores. The discriminant scores provide a reasonably good separation among the four groups. That is, the discriminant scores resulting from the discriminant function Z are sufficient to represent most of the differences among the four groups.

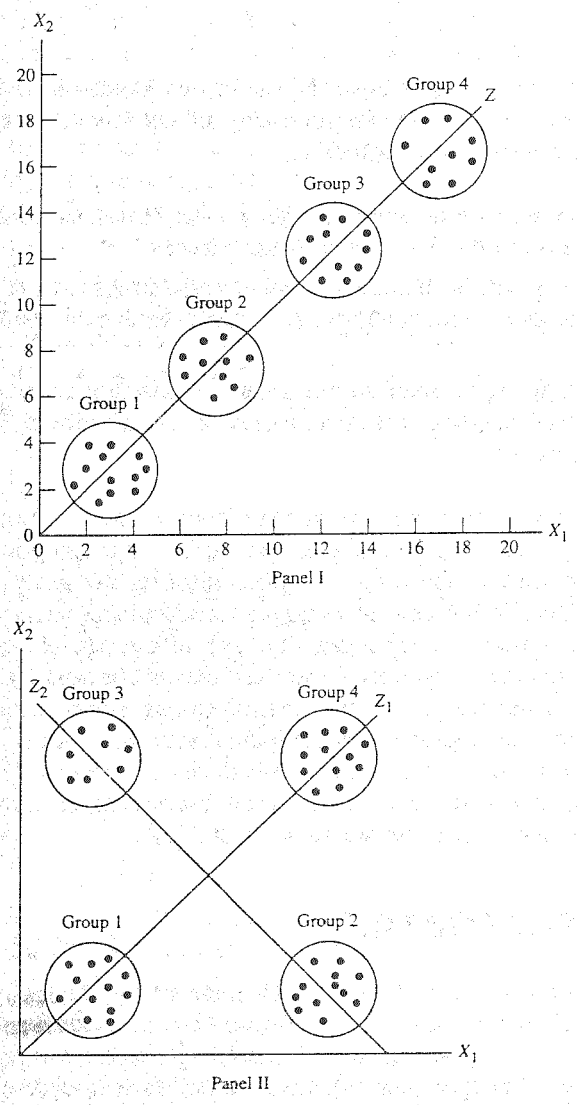


Figure 9.1 Hypothetical scatterplot.

[1]Recall that the number of discriminant functions is equal to the $min(G-1, p)$ where G and p are, respectively, the number of groups and the number of discriminating variables.

Panel II of Figure 9.1 gives a plot for another hypothetical set of observations belonging to four groups. Again, it is apparent that the means of variables X_1 and X_2 for the four groups are different. Let Z_1 be the axis that represents the discriminant function. The discriminant scores, given by the projection of the points onto the discriminant function, Z_1, appear to provide good discrimination between all pairs of groups except groups 2 and 3. Therefore, it is necessary to identify another discriminant function for discriminating between groups 2 and 3. Let Z_2 be the axis representing the second discriminant function. This second discriminant function discriminates between groups 2 and 3 as well as other pairs of groups; however, it does not discriminate between groups 1 and 4. Therefore, in order to account for *all the possible* discrimination or differences among all pairs of the four groups we need both discriminant functions, Z_1 and Z_2. In the present case, more than one discriminant function is required to adequately represent the differences among the four groups. The two axes (i.e., the discriminant functions), Z_1 and Z_2, are not constrained or required to be orthogonal to each other. The only requirement is that the two sets of discriminant scores be uncorrelated.

In the preceding example we did not gain much in terms of data reduction; we could as well have used the original two variables, X_1 and X_2, for discriminating purposes. But suppose that the spatial configuration of the observations in the four groups given in Panel II of Figure 9.1 is the same in, say, 20 dimensions (i.e., $p = 20$). For such a spatial configuration, most of the differences among the four groups can be represented in a two-dimensional discriminant space defined by the two discriminant functions, Z_1 and Z_2, as opposed to representing the differences in 20 dimensions. Obviously, this gives a substantial amount of parsimony in representing the data. Because the number of discriminating variables is usually much larger than the number of groups, a substantial amount of parsimony can be obtained by representing the differences among groups in an r-dimensional discriminant space where $r \leq G - 1$.[2]

9.1.2 Identifying New Axes

Consider the hypothetical data given in Table 9.1. Figure 9.2 gives a plot of the data. Let Z_1 be a new axis that makes an angle of $\theta = 46.115°$ with the X_1 axis. Projection of the points onto Z_1 gives the new variable Z_1. Table 9.2 gives the total, between-groups, and within-group sums of squares and λ_1, the ratio of between-groups to within-group sums of squares, for various angles between Z_1 and X_1. Figure 9.3 gives a plot of λ_1 and θ. From the table and the figure we can see that the maximum value of λ_1 is 19.250 when θ is equal to 46.115° or 226.115° (46.115° + 180°). The equation for obtaining the projection of points onto Z_1 is:

$$\begin{aligned} Z_1 &= \cos 46.115 \times X_1 + \sin 46.115 \times X_2 \\ &= 0.693 \times X_1 + 0.721 \times X_2, \end{aligned} \tag{9.1}$$

which can be used to compute the discriminant scores for each observation.

However, note from Figure 9.2 that Z_1 cannot differentiate between groups 2 and 3. Therefore, we have to search for another axis that would differentiate between these two groups. The first discriminant function accounts for the maximum differences among the groups and corresponds to the maximum value of λ, so the second discriminant function would naturally correspond to the other extreme value of λ. From Table 9.2 and Figure 9.3 we see that the second extreme point corresponds to $\theta = 136.115°$ or

[2] All of the $G - 1$ dimensions may not be necessary. That is, only r of the $G - 1$ dimensions may be necessary to adequately account for an acceptable amount of the differences among the groups.

Table 9.1 **Hypothetical Data for Four Groups**

	Group 1		Group 2		Group 3		Group 4	
Observation	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
1	1	3	11	3	1	13	11	13
2	2	1	12	1	2	11	12	11
3	4	1	14	1	4	11	14	11
4	5	3	15	3	5	13	15	13
5	4	4.5	14	4.5	4	14.5	14	14.5
6	2	4.5	12	4.5	2	14.5	12	14.5
7	3	1.5	13	1.5	3	11.5	13	11.5
8	4.5	1.5	14.5	1.5	4.5	11.5	14.5	11.5
9	4.5	3	14.5	3	4.5	13	14.5	13
10	3	4	13	4	3	14	13	14
11	2	3.5	12	3.5	2	13.5	12	13.5
12	2	2	12	2	2	12	12	12
13	3	3	13	3	3	13	13	13
Mean	3.077	2.731	13.077	2.731	3.077	12.731	13.077	12.731
Std. Dev.	1.239	1.235	1.239	1.235	1.239	1.235	1.239	1.235

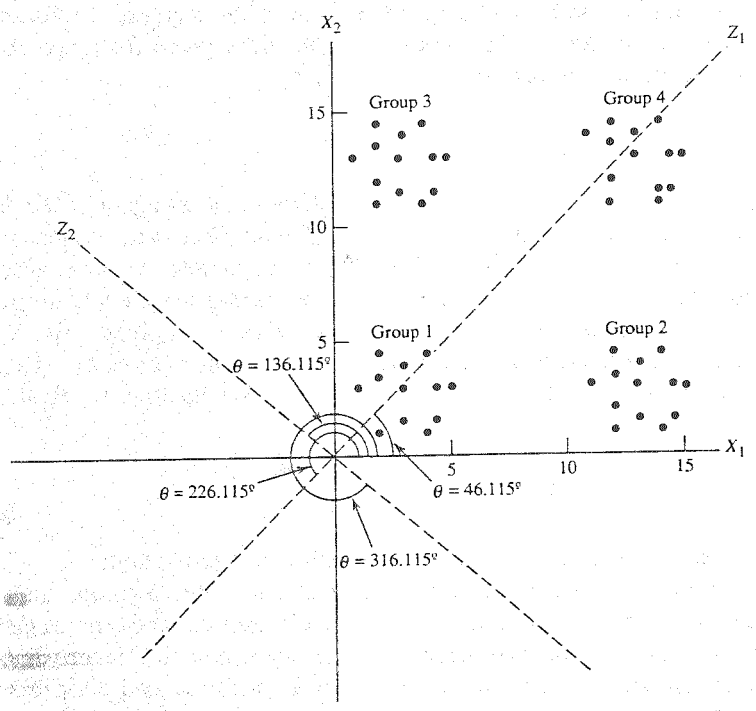


Figure 9.2 Plot of data in Table 9.1.

Table 9.2 **Lambda for Various Angles between Z and X_1**

Rotation, θ_1 (deg)	Weights		Sums of Squares			
	w_1	w_2	SS_t	SS_w	SS_b	λ
0	1.000	0.000	1373.692	73.692	1300.000	17.641
40	0.766	0.643	1367.682	67.671	1300.011	19.211
46.115	0.693	0.721	1367.534	67.534	1300.000	19.250
80	0.174	0.985	1371.203	71.217	1299.986	18.254
120	−0.500	0.866	1378.415	78.472	1299.943	16.566
136.115	−0.721	0.693	1379.411	79.396	1300.015	16.374
160	−0.940	0.342	1377.457	77.448	1300.009	16.786
200	−0.940	−0.342	1369.858	69.837	1300.021	18.615
240	−0.500	−0.866	1368.156	68.214	1299.942	19.057
280	0.174	−0.985	1375.254	75.271	1299.983	17.271
316.115	0.721	−0.693	1379.418	79.388	1300.030	16.376
320	0.766	−0.643	1379.298	79.330	1299.968	16.387
360	1.000	0.000	1373.692	73.692	1300.000	17.641

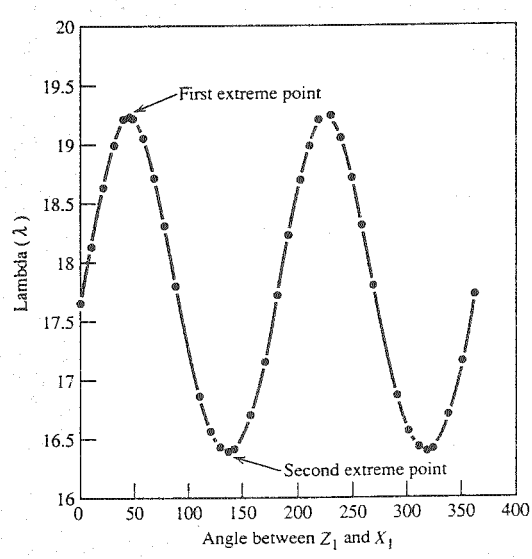


Figure 9.3 Plot of rotation angle versus lambda.

316.115° (136.115° + 180°), giving a value of 16.374 for λ_2. The equation giving the projection of points onto Z_2 is:

$$\begin{aligned} Z_2 &= \cos 136.115 \times X_1 + \sin 136.115 \times X_2 \\ &= -0.721 \times X_1 + 0.693 \times X_2, \end{aligned} \tag{9.2}$$

which can be used to compute the second set of discriminant scores for each observation. Note that in the present case the two axes, Z_1 and Z_2, are orthogonal. This will not necessarily be the case for other data sets. That is, the discriminant functions are not constrained to be orthogonal to each other. The only constraint is that the resulting discriminant scores be uncorrelated.

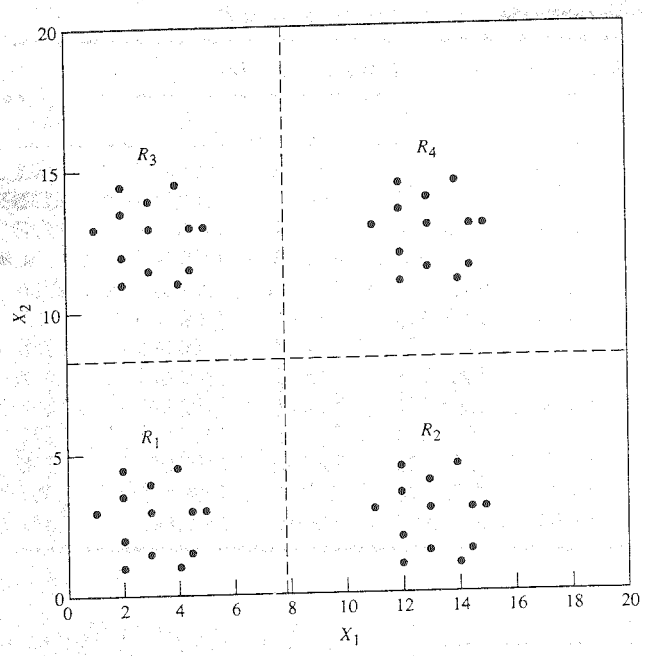


Figure 9.4 Classification in variable space.

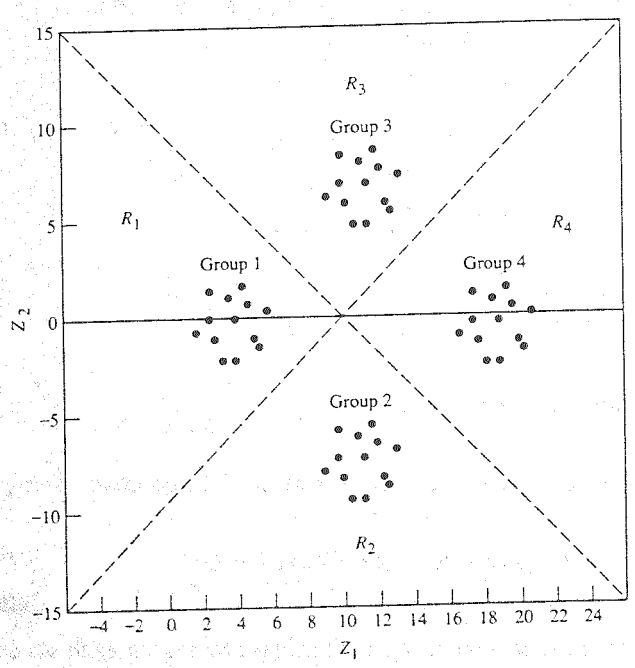


Figure 9.5 Classification in discriminant space.

9.1.3 Classification

As pointed out in Chapter 8, classification can be viewed as the division of the total discriminant or variable space into $R_1, R_2, \ldots, R_G$ mutually exclusive and exhaustive regions. Any given observation is classified into the group in whose region the observation falls. Figure 9.4 gives the four classification regions in the variable space (i.e., the original data). Note that two straight lines are needed to divide the two-dimensional space into the four regions. The straight lines can be referred to as the cutoff lines. A number of criteria or rules can be used for identifying the cutoff lines to obtain the classification regions. These rules are generalizations of the rules discussed in Chapter 8 and its Appendix, and are discussed in detail in the Appendix of this chapter.

Figure 9.5 gives the four classification regions in the discriminant space. Again, two cutoff lines are needed to obtain the four regions. However, if only one discriminant function were needed to adequately represent the differences among the four groups then the discriminant space plot would be a one-dimensional plot, and three points (i.e., cutoff values) would be needed to divide the one-dimensional space into four regions.

9.2 ANALYTICAL APPROACH

The objectives and mechanics of multiple-group discriminant analysis are quite similar to those of two-group discriminant analysis. First, a univariate analysis can be done to determine if each of the discriminating variables significantly discriminates among the four groups. This can be achieved by an overall F-test. The overall F-test would be significant if the mean of at least one pair of groups is significantly different.

Having identified the discriminating variables, the next step is to estimate the discriminant function. Suppose the first discriminant function is

$$Z_1 = w_{11}X_1 + w_{12}X_2 + \cdots + w_{1p}X_p$$

where w_{ij} is the weight of the jth variable for the ith discriminant function. The weights of the discriminant function are estimated such that the

$$\lambda_1 = \frac{\text{between-groups SS of } Z_1}{\text{within-group SS of } Z_1}$$

is maximized.

Suppose that the second discriminant function is given by

$$Z_2 = w_{21}X_1 + w_{22}X_2 + \cdots + w_{2p}X_p.$$

The weights of the above discriminant function are estimated such that the

$$\lambda_2 = \frac{\text{between-groups SS of } Z_2}{\text{within-group SS } Z_2}$$

is maximized subject to the constraint that the discriminant scores Z_1 and Z_2 are uncorrelated. The procedure is repeated until all the possible discriminant functions are identified. This is clearly an optimization problem and, as discussed in the Appendix to Chapter 8, the solution is to find the eigenvalues and eigenvectors of the nonsymmetric matrix $\mathbf{W}^{-1}\mathbf{B}$ where $\mathbf{W}$ and $\mathbf{B}$ are, respectively, the within-group and between-group $SSCP$ matrices of the p variables. Note that since the matrix $\mathbf{W}^{-1}\mathbf{B}$ is nonsymmetric, the eigenvectors may not be orthogonal. That is, the discriminant functions will not be orthogonal. However, the resulting discriminant scores will be uncorrelated.

Once the discriminant function(s) have been identified, the next step is to determine a rule for classifying future observations. Classification procedures in MDA are generalizations of the procedures discussed in the two-group case. As discussed previously, all classification procedures involve the division of the discriminant space, or the variable space, into G mutually exclusive and collectively exhaustive regions. For example, to classify any given observation using the discriminant scores, the discriminant scores are computed, then the observation is plotted in the discriminant space. The observation is classified into the group in whose region it falls. The various classification procedures are discussed in Section 9.3.3 and in the Appendix.

9.3 MDA USING SPSS

The data given in Table 9.1 are used to discuss the resulting SPSS output. Table 9.3 gives the SPSS commands and Exhibit 9.1 gives the resulting output. Many of the computational procedures relating to the various test statistics are discussed in detail in Chapter 8, and we refer the reader to appropriate sections of that chapter.

9.3.1 Evaluating the Significance of the Variables

The output reports the means of the variables for the total sample and each group, the Wilks' Λ, and the univariate F-ratio [1a, 1b]. The transformed value of Wilks' Λ follows an exact F-distribution only for certain cases (see Table 9.4). In all other cases, the distribution of the transformed value of Wilks' Λ can only be *approximated* by an F-distribution. The F-statistic is used to test the following univariate null and alternative hypotheses for each discriminating variable, X_1 and X_2:

$$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_a : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4,$$

where μ_1, μ_2, μ_3, and μ_4 are, respectively, population means for groups 1, 2, 3, and 4. The null hypothesis will be rejected if the means of at least one pair of groups are significantly different. The null hypothesis for both the variables can be rejected at a significance level of .05. That is, at least one pair of groups is significantly different with respect to the means of X_1 and X_2. A discussion of which pair or pairs of groups are different is provided in the following section.

9.3.2 The Discriminant Function

Options for Computing the Discriminant Function

The various control parameters for estimating the discriminant functions are given in this section [2a]. Since we have four groups and only two variables, the maximum

Table 9.3 **SPSS Commands**

```
DISCRIMINANT GROUPS=GROUP(1,4)
 /VARIABLES=X1,X2
 /ANALYSIS=X1,X2
 /METHOD=DIRECT
 /STATISTICS=ALL
```

Exhibit 9.1

Discriminant analysis for data in table 9.1

```
0GROUP MEANS

(1a)0  GROUP          X1            X2
          1        3.07692       2.73077
          2       13.07692       2.73077
          3        3.07692      12.73077
          4       13.07692      12.73077
0      TOTAL       8.07692       7.73077

(1b)0WILKS' LAMBDA (U-STATISTIC) AND UNIVARIATE F-RATIO
 WITH   3 AND        48 DEGREES OF FREEDOM
0  VARIABLE   WILKS' LAMBDA        F          SIGNIFICANCE
   --------   -------------   -------------   ------------
   X1            0.05365         282.3           0.0000
   X2            0.05333         284.0           0.0000

(2a)      MINIMUM TOLERANCE LEVEL................... 0.00100
0CANONICAL DISCRIMINANT FUNCTIONS
0     MAXIMUM NUMBER OF FUNCTIONS..............        2
      MINIMUM CUMULATIVE PERCENT OF VARIANCE...   100.00
      MAXIMUM SIGNIFICANCE OF WILKS' LAMBDA....   1.0000
0
 PRIOR PROBABILITY FOR EACH GROUP IS 0.25000

(2b)0CLASSIFICATION FUNCTION COEFFICIENTS
 (FISHER'S LINEAR DISCRIMINANT FUNCTIONS)
0GROUP   =          1             2             3             4
 X1            2.162097      8.718289      2.692377      9.248569
 X2            1.964791      2.495072      8.562304      9.092584
 (CONSTANT)   -7.395294     -61.79722     -60.03077     -119.7355

-                          CANONICAL DISCRIMINANT FUNCTIONS
-                   PCT OF   CUM  CANONICAL  AFTER  WILKS'      (2d)
(2c) FCN EIGENVALUE VARIANCE  PCT     CORR     FCN  LAMBDA  CHISQUARE    DF  SIG

                                          :    0  0.0028    281.432     6 0.0000
     1*   19.2496   54.03   54.03   0.9750 :    1  0.0576    137.042     2 0.0000
     2*   16.3750   45.97  100.00   0.9708 :
0     * MARKS THE 2 CANONICAL DISCRIMINANT FUNCTIONS REMAINING IN THE ANALYSIS.
(CONSTANT)  -9.417712     -0.3595065

(2e)0UNSTANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS
0              FUNC  1        FUNC  2

 X1          0.5844154      0.5604264
 X2          0.6076282     -0.5390168
 (CONSTANT)  -9.417712     -0.3595065
```

(continued)

Exhibit 9.1 (continued)

2f 0SYMBOLS USED IN TERRITORIAL MAP

0SYMBOL	GROUP	LABEL
1	1	
2	2	
3	3	
4	4	
*		GROUP CENTROIDS

TERRITORIAL MAP * INDICATES A GROUP CENTROID

```
0                    CANONICAL DISCRIMINANT FUNCTION 1
      -12.0      -8.0      -4.0        .0       4.0       8.0      12.0
          +---------+---------+---------+---------+---------+---------+
 C  12.0 +222                                                      2244+
 A       I1122                                                    22244 I
 N       I 11222                                                 22444  I
 O       I  111222                                              2244    I
 N       I    11122                                            22244     I
 I       I      11222                                         22444      I
 C   8.0 +       11122        +         +         +        22244        +
 A       I         11222                                   22444        I
 L       I          111222              R2                22244         I
         I            11122                             22444           I
 D       I              11222           *             22244             I
 I       I               11122                      22444               I
 S   4.0 +          +      11222        +         22244       +         +
 C       I                 111222                22444                 I
 R       I                   11122            22244                   I
 I       I                    11222         22444                      I
 M       I              R1        11122     22244      R4              I
 I       I                         11222  22444                       I
 N    .0 +          +    *    +     111222244     +    *    +         +
 A       I                           113333444                        I
 N       I                          11133  33344                      I
 T       I                         11333     33444                    I
         I                       11133        33344                   I
 F       I                      11333           33444                I
 U  -4.0 +          +        11133      +       333444      +         +
 N       I                  11333                33344               I
 C       I                11133          *          33444             I
 T       I               11333                      33344             I
 I       I             11133              R3            33444         I
 O       I            11333                            333444         I
 N  -8.0 +          11133     +         +         +       33444       +
         I         11333                                   33444      I
 2       I       11133                                      33344     I
         I      11333                                         33444   I
         I     1133                                            333444 I
         I   11133                                               33344I
-12.0    +  11333                                                  334+
          +---------+---------+---------+---------+---------+---------+
      -12.0
```

(continued)

Exhibit 9.1 (continued)

2g 0CANONICAL DISCRIMINANT FUNCTIONS EVALUATED AT GROUP MEANS (GROUP CENTROIDS)

0 GROUP	FUNC 1	FUNC 2
1	-5.96022	-0.10705
2	-0.11606	5.49722
3	0.11606	-5.49722
4	5.96022	0.10705

2h

CASE NUMBER	ACTUAL GROUP	HIGHEST PROBABILITY GROUP	P(D/G)	P(G/D)	2ND HIGHEST GROUP	P(G/D)	DISCRIM SCORES
1	1	1	0.2446	1.0000	3	.0000	-7.0104 -1.4161
2	1	1	0.2306	1.0000	1	1.0000	-7.6413 0.2223
3	1	1	0.3064	1.0000	2	.0000	-6.4724 1.3432
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
50	4	4	.5878	1.0000	3	.0000	5.7983 -.9111
51	4	4	.5499	1.0000	3	.0000	4.8868 -.1026
52	4	4	.9756	1.0000	3	.0000	6.0789 -.0812

2i 0CLASSIFICATION RESULTS -

0 ACTUAL GROUP		NO. OF CASES	PREDICTED GROUP MEMBERSHIP 1	2	3	4
0GROUP	1	13	13 100.0%	0 0.0%	0 0.0%	0 0.0%
0GROUP	2	13	0 0.0%	13 100.0%	0 0.0%	0 0.0%
0GROUP	3	13	0 0.0%	0 0.0%	13 100.0%	0 0.0%
0GROUP	4	13	0 0.0%	0 0.0%	0 0.0%	13 100.0%

0PERCENT OF "GROUPED" CASES CORRECTLY CLASSIFIED: 100.00%

number of discriminant functions that can be estimated is two. Also, the priors are assumed to be equal (i.e., they are .25 for each group). That is, the probability that any given observation belongs to any one of the four groups is the same.

Estimate of the Discriminant Functions

The unstandardized discriminant functions are [2e]:

$$Z_1 = -9.418 + .584X_1 + .608X_2 \tag{9.3}$$

$$Z_2 = -0.360 + .560X_1 - .539X_2. \tag{9.4}$$

Again, ignoring the constant, the ratios of the coefficients in Eqs. 9.3 and 9.4, respectively, are the same as those reported in Eqs. 9.1 and 9.2. Note that the signs of the

Table 9.4 **Cases in Which Wilks' Λ Is Exactly Distributed as *F***

Number of Variables (p)	Number of Groups (G)	Transformation	Degrees of Freedom
Any	2	$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n-p-1}{p}\right)$	$p, n-p-1$
Any	3	$\left(\frac{1-\Lambda^{1/2}}{\Lambda^{1/2}}\right)\left(\frac{n-p-2}{p}\right)$	$2p, 2(n-p-1)$
1	Any	$\left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n-G}{G-1}\right)$	$G-1, n-G$
2	Any	$\left(\frac{1-\Lambda^{1/2}}{\Lambda^{1/2}}\right)\left(\frac{n-G-1}{G-1}\right)$	$2(G-1), 2(n-G-1)$

coefficients for the discriminant function given by Eq. 9.4 are opposite of those given in Eq. 9.2. This is not a matter of concern as the latter equation can be obtained by multiplying the former equation by minus one. Furthermore, note that in Figure 9.2 Z_2 makes an angle of 136.115° or 316.115° (i.e., 136.115° + 180°) with X_1. If one were to use an angle of 316.115° between Z_2 and X_1 then in Eq. 9.2 the weights of X_1 and X_2, respectively, would be 0.721 and −0.693, which now have the same sign as the weights of Eq. 9.4.

How Many Discriminant Functions?

The next obvious question is: How many discriminant functions should one retain or use to adequately represent the differences among the groups? This question can be answered by evaluating the statistical significance and the practical significance of each discriminant function. That is, does the discriminant score of the respective discriminant function significantly differentiate among the groups?

STATISTICAL SIGNIFICANCE. Not all of the K discriminant functions may be statistically significant. That is, only r (where $r \leq K$) discriminant functions may be necessary to represent most of the differences among the groups. The following formula is used to compute the χ^2 value for assessing the overall statistical significance of all the discriminant functions:

$$\chi^2 = [n - 1 - (p + G)/2] \sum_{k=1}^{K} \ln(1 + \lambda_k) \tag{9.5}$$

where λ_k is the eigenvalue of the kth discriminant function. Using the above formula the resulting χ^2 value is

$$\begin{aligned}\chi^2 &= [52 - 1 - (2 + 4)/2][\ln(1 + 19.24957) + \ln(1 + 16.37504)] \\ &= 281.424,\end{aligned}$$

which is the same as the χ^2 reported in the output [2d]. Notice that the preceding χ^2 value uses eigenvalues for *all* the K discriminant functions. Therefore, the χ^2 value reported in the first row of the output does not test the statistical significance of just the first function, rather it jointly tests the statistical significance of all the possible discriminant functions. A significant χ^2 value implies that *at least* the first discriminant function is significant; other discriminant functions may or may not be significant. In the present case the χ^2 value of 281.432 is statistically significant, suggesting that at least the first discriminant function is statistically significant.

Statistical significance of the remaining discriminant functions determines whether they jointly explain a significant amount of difference among the four groups that has not been explained by the first discriminant function. The statistical significance test can be accomplished by computing the χ^2 value from the following equation

$$\chi^2 = [n - 1 - (p + G)/2] \sum_{k=2}^{K} \ln(1 + \lambda_k), \tag{9.6}$$

which in the present case is equal to

$$\begin{aligned}\chi^2 &= [52 - 1 - (2/4)/2][\ln(1 + 16.37504)] \\ &= 137.040\end{aligned}$$

and is the same as that reported in the output [2d]. Notice that Eq. 9.6 is a modification of Eq. 9.5 in that the computation excludes the eigenvalue of the first discriminant function. A significant χ^2 value would imply that the second and maybe the following discriminant functions significantly explain the difference in the groups that was not explained by the first function. Because the χ^2 value of 137.040 is statistically significant, we conclude that at least the second discriminant function also explains a significant amount of difference among the four groups that was not explained by the first discriminant function.

In the case of K discriminant functions the above procedure is repeated until the χ^2 value is not significant. In general, to examine the statistical significance of the rth discriminant function the formula used for computing the χ^2 value is

$$\chi_r^2 = [n - 1 - (p + G)/2] \sum_{k=r}^{K} \ln(1 + \lambda_k) \tag{9.7}$$

with $(p - r + 1)(G - r)$ degrees of freedom.

The conclusion drawn from the preceding significance tests is that the four groups are significantly different with respect to the means of the discriminant scores of both the discriminant functions. But, which pairs of groups are different? This question can be addressed by examining the means of the discriminant scores [2g]. Note that the means of the discriminant score obtained from the first function, Z_1, appears to be different for all pairs of groups except groups 2 and 3; and the means of the discriminant score obtained from the second function, Z_2, are not different for groups 1 and 4. That is, it appears that the first discriminant function significantly discriminates between all pairs of groups except groups 2 and 3 and the second discriminant function significantly discriminates between all pairs of groups except groups 1 and 4. However, it will not always be possible to determine for each function which pairs of groups are significantly different by visually examining the means. In order to formally determine which pairs of group-means are different, one would have to resort to pairwise tests, such as LSD (least significant difference), Tukey's test, and Scheffe's test. These tests are available in the ONEWAY procedure in **SPSS**.[3] Following is a brief discussion of the output from the ONEWAY procedure.

Table 9.5 gives the **SPSS** commands. The COMPUTE statements are used for computing the discriminant scores. Note that the unstandardized discriminant functions are used for computing the discriminant scores [2e]. The ONEWAY procedure requests an analysis of variance for each of the dependent variables, Z_1 and Z_2. The RANGES = LSD(.05) subcommand requests pairwise comparison of means using the LSD test and an alpha level of .05. Other tests can be requested by specifying the name of the test. For example, Tukey's test can be requested by specifying RANGES = TUKEY. Exhibit 9.2 gives the partial output.

Table 9.5 **SPSS Commands for Range Tests**

```
COMPUTE Z1=-9.417712+.5844154*X1+.6076282*X2
COMPUTE Z2=-.3595065+.5604264*X1-.5390168*X2
ONEWAY Z1,Z2,BY,GROUP(1,4)
 /RANGES=LSD(.05)
```

[3]See Winer (1982) for a detailed discussion of these tests.

Exhibit 9.2

Range Tests for Data in Table 9.1

```
-       Variable  Z1
0                                    ANALYSIS OF VARIANCE
1
                                 SUM OF        MEAN            F       F
      SOURCE              D.F.  SQUARES      SQUARES         RATIO  PROB.

 BETWEEN GROUPS             3    923.9794    307.9931      307.9932  .0000
0WITHIN GROUPS             48     48.0000      1.0000
0TOTAL                     51    971.9794

LSD PROCEDURE  RANGES FOR THE 0.050 LEVEL -
(*) DENOTES PAIRS OF GROUPS SIGNIFICANTLY DIFFERENT AT THE 0.050 LEVEL

2                          G G G G
                           r r r r
                           p p p p

  Mean       Group         1 2 3 4

 -5.9602     Grp 1
  -.1161     Grp 2         *
   .1161     Grp 3         *
  5.9602     Grp 4         * * *

3-      Variable  Z2
0                                    ANALYSIS OF VARIANCE

                                 SUM OF        MEAN            F       F
      SOURCE              D.F.  SQUARES      SQUARES         RATIO  PROB.

 BETWEEN GROUPS             3    786.0019    262.0006      262.0007  .0000
0WITHIN GROUPS             48     48.0000      1.0000
0TOTAL                     51    834.0019

4 LSD PROCEDURE RANGES FOR THE 0.050 LEVEL -

                           G G G G
                           r r r r
                           p p p p

  Mean       Group         3 1 4 2

 -5.4972     Grp 3
  -.1070     Grp 1         *
   .1070     Grp 4         *
  5.4972     Grp 2         * * *
```

The ANOVA table gives the F-ratio for testing the following null and alternate hypotheses.

$$H_o : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_a : \mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4.$$

The overall F-ratios of 307.993 for Z_1 and 262.001 for Z_2 suggest that the null hypothesis can be rejected at an alpha of .05 [1, 3]. That is, at least one pair of groups is significantly different with respect to the two discriminant functions. This conclusion is not different from that reached previously. The additional information of interest given in the output is for pairwise tests [2, 4]. The asterisks indicate which pairs of means are significantly different at the given alpha level. It can be seen that the means of Z_1 are significantly different for all pairs of groups except groups 2 and 3 [2], and the means of Z_2 are significantly different for all pairs of groups except groups 1 and 4 [4].

PRACTICAL SIGNIFICANCE. As usual, statistical significance tests are sensitive to sample size. That is, for a large sample size a discriminant function accounting for only a small difference among the groups might be statistically significant. Therefore, one must also take into account the practical significance of a given discriminant function. The practical significance of a discriminant function is assessed by the squared canonical correlation (CR^2) and the λ's or the eigenvalues.

As discussed in Chapter 8, a two-group discriminant analysis problem can be formulated as a multiple regression problem. Similarly, a multiple-group discriminant analysis problem can be formulated as a canonical correlation problem with group membership, coded using dummy variables, as the dependent variables.[4] In the present case, three dummy variables are required to code the four groups, resulting in three dependent variables, and the canonical correlation analysis will result in two canonical functions.[5] The first and the second canonical functions, respectively, relate to the first and second discriminant functions. The resulting canonical correlations will be .975 and .971 giving CR^2s of .951 and .943, respectively, for the first and the second discriminant functions [2c, Exhibit 9.1]. High values of CR^2s suggest that the discriminant functions account for a substantial portion of the differences among the four groups.

One can also use the eigenvalues (i.e., λ's) to assess the practical significance of the discriminant functions. Recall that λ is equal to SS_b/SS_w. The greater the value of λ for a given discriminant function, the greater the ability of that discriminant function to discriminate among the groups. Therefore, the λ of a given discriminant function can also be used as a measure of its practical significance. The importance or the discriminating ability of the jth discriminant function can be assessed by the measure, *percent of variance*, which is defined as

$$\frac{\lambda_j}{\sum_{j=1}^{K} \lambda_j} \times 100,$$

where K is the maximum number of discriminant functions that can be estimated. Note that the percent of variance measure does not refer to the variance in the data; rather it represents the percent of the total differences among the groups that is accounted for by the discriminant function. The percents of variance for the two discriminant functions are equal to

$$\frac{19.250}{19.250 + 16.375} \times 100 = 54.03; \qquad \frac{16.375}{19.250 + 16.375} \times 100 = 45.97.$$

[4]Canonical correlation analysis is discussed in Chapter 13.

[5]The maximum number of canonical functions is equal to $min(p, q)$ where p and q are equal to, respectively, the number of variables in Set 1 and Set 2. Although canonical correlation analysis does not differentiate between independent and dependent variables, for this example Set 1 corresponds to independent variables and Set 2 corresponds to dependent variables.

That is, the first discriminant function accounts for 54.03% of the possible differences among the groups and the second accounts for the remaining 45.97% of the differences among the groups [2c]. Together, the discriminant functions account for all (i.e., 100%) of the possible differences among the four groups. In the present case both the discriminant functions are needed to account for a significant portion of the total differences among the groups. This assertion is also supported by the high values of CR^2s.

But, how high is high? Or, what is the cutoff value for determining how many discriminant functions should be used? The problem is similar to the problem of how many principal components or factors should be retained in principal components analysis and factor analysis. One could use a scree plot in which the X axis represents the number of discriminant functions and the Y axis represents the eigenvalues. In any case, the issue of how many functions should be retained is ultimately a judgmental issue and varies from researcher to researcher, and from situation to situation.

Assessing the Importance of Discriminant Variables and the Meaning of Discriminant Function

As discussed in Chapter 8, the standardized coefficients and the loadings can be used for assessing the importance of the variables forming the discriminant functions. Since the data are hypothetical (i.e., we do not know what X_1 and X_2 stand for) it is not possible to assign any meaningful labels to the discriminant functions. The use of loadings to assign meaning to the discriminant functions is discussed in Section 9.4.

9.3.3 Classification

A number of different rules can be used for classifying future observations. These rules are generalizations of the rules discussed in the Appendix to Chapter 8. The Appendix to this chapter provides a detailed discussion of the various rules for classifying observations into multiple groups.

Classification functions for each group are reported by **SPSS** [2b]. To classify a given observation, first the classification functions of each group are used to compute the classification scores and the observation is assigned to the group that has the highest classification score.

The posterior probability of an observation belonging to a given group can also be computed. The observation is assigned to the group with the highest posterior probability. **SPSS** reports the two highest posterior probabilities [2h]. According to the classification matrix all the observations are correctly classified [2i]. The statistical significance of the classification rate can be assessed by using the procedure described in Chapter 8. Using Eq. 8.20, the expected number of correct classifications due to chance alone is 13, and from Eq. 8.18 $Z^* = 12.49$, which is significant at $p < .01$.

As mentioned in Section 9.1.3, classification essentially involves the division of the total discriminant space into mutually exclusive and collectively exhaustive regions. A plot displaying these regions is called the *territorial map*. **SPSS** provides the territorial map [2f]. In the map the axes represent the discriminant functions and the asterisks represent the centroids of the groups. The four mutually exclusive regions are marked as R_1, R_2, R_3, and R_4. In order to classify a given observation, discriminant scores are first computed and plotted in the territorial map. The observation is then classified into the group in whose territory or region the given observation falls. For example, consider a new observation with values of 3 and 4, respectively, for X_1 and X_2. The discriminant scores, Z_1 and Z_2, respectively, will be

$$Z_1 = -9.418 + .584 \times 3 + .608 \times 4 = -5.234$$

and

$$Z_2 = -.360 + .560 \times 3 - .539 \times 4 = -.836.$$

It can be seen that the observation falls in region R_1 and, therefore, it is classified into group 1.

9.4 AN ILLUSTRATIVE EXAMPLE

Assume that the brand manager of a major brewery is interested in gaining additional insights about differences among major brands of beer, as perceived by its target market. Consumers are asked to rate four major brands of beer using the following five-point semantic differential scale:

Heavy	___	___	___	___	___	Light
Mellow	___	___	___	___	___	Not Mellow
Not Filling	___	___	___	___	___	Filling
Good Flavor	___	___	___	___	___	Bad Flavor
No Aftertaste	___	___	___	___	___	Aftertaste
Foamy	___	___	___	___	___	Not Foamy

The data are coded such that higher numbers represent positive attributes. For example, *Heavy* was coded as 1 and *Light* as 5, and *Good Flavor* was coded as 5 and *Bad Flavor* as 1.

Table 9.6 gives the SPSS commands. The FUNCTIONS subcommand is used to specify the number of functions that are to be retained. The first option gives the maximum number of functions that can be retained, the second specifies the maximum percent of variance that can be accounted for, and the third gives the p-value of the retained functions. In the present case, out of the three possible functions that would account for 100% of the differences, only those functions that are statistically significant at an alpha of .05 will be retained. The ROTATE subcommand specifies that the STRUCTURE matrix should be rotated. Exhibit 9.3 gives the relevant portion of the discriminant analysis output. The Wilks' Λ and the univariate F-ratio indicate that the means of all but the sixth attribute (*Foamy, Not Foamy*) are different among the four brands [1]. However, management desired to include all six variables for computing the respective discriminant functions. Of the three possible discriminant functions, only the first two are statistically significant and account for most (i.e., more than 95%) of the possible differences among the four brands [2].

The output also gives average discriminant scores for the four groups [6]. These are obtained by substituting the group centroids or group means in the unstandardized

Table 9.6 **SPSS Commands for Beer Example**

```
DISCRIMINANT GROUPS=BRAND(1,4)
 /VARIABLES=LIGHT MELLOW FILLING FLAVOR TASTE FOAMY
 /METHOD=DIRECT
 /FUNCTIONS=3,100,.05
 /ROTATE=STRUCTURE
 /STATISTICS=ALL
```

Exhibit 9.3

SPSS Output for the Beer Example

```
(1) 0WILKS' LAMBDA (U-STATISTIC) AND UNIVARIATE F-RATIO
     WITH   3 AND        79 DEGREES OF FREEDOM
   0  VARIABLE   WILKS' LAMBDA        F          SIGNIFICANCE
      --------   -------------   -------------   ------------

      LIGHT         0.72870         9.804          0.0000
      MELLOW        0.77068         7.836          0.0001
      FILLING       0.90586         2.737          0.0490
      FLAVOR        0.73708         9.393          0.0000
      TASTE         0.87292         3.834          0.0128
      FOAMY         0.97821        0.5865          0.6256


(2) -                      CANONICAL DISCRIMINANT FUNCTIONS
   -                  PCT OF   CUM  CANONICAL  AFTER  WILKS'
     FCN EIGENVALUE VARIANCE  PCT     CORR      FCN  LAMBDA  CHISQUARE  DF  SIG
                                                :   0  0.4655     58.875  18 0.0000
      1*    0.5461   59.26   59.26   0.5943 :    1  0.7197     25.323  10 0.0048
      2*    0.3333   36.17   95.43   0.5000 :    2  0.9596      3.173   4 0.5293
      3     0.0421    4.57  100.00   0.2009 :
   0    * MARKS THE 2 CANONICAL DISCRIMINANT FUNCTIONS REMAINING IN THE ANALYSIS.

(3) 0STRUCTURE MATRIX:
   0POOLED WITHIN-GROUPS CORRELATIONS BETWEEN DISCRIMINATING VARIABLES
                                     AND CANONICAL DISCRIMINANT FUNCTIONS
    (VARIABLES ORDERED BY SIZE OF CORRELATION WITHIN FUNCTION)
   0              FUNC  1     FUNC  2
    LIGHT         0.72663*    0.48139
    FLAVOR        0.72630*   -0.44991
    MELLOW        0.71650*    0.08120
    TASTE         0.49455*   -0.18393

    FILLING       0.20731     0.48099*
    FOAMY        -0.11984     0.15391*

(4) 0ROTATED CORRELATIONS BETWEEN DISCRIMINATING VARIABLES
                          AND CANONICAL DISCRIMINANT FUNCTIONS
    (VARIABLES ORDERED BY SIZE OF CORRELATION WITHIN FUNCTION)
   0              FUNC  1     FUNC  2
    FLAVOR        0.84897*    0.09578
    MELLOW        0.51274*    0.50701
    TASTE         0.50235*    0.16140
    FOAMY        -0.18936*    0.04680

    LIGHT         0.27315     0.82772*
    FILLING      -0.13464     0.50616*

(5) 0UNSTANDARDIZED CANONICAL DISCRIMINANT FUNCTION COEFFICIENTS
   0               FUNC  1         FUNC  2
    LIGHT        0.5326631E-01   0.4133531
    MELLOW      -0.2814968E-03   0.3395821
    FILLING     -0.2546457       0.2142430
    FLAVOR       0.4967704      -0.1624510
    TASTE        0.2686773      -0.2402488
    FOAMY        0.4381629E-01   0.2124775E-01
    (CONSTANT)  -2.742837        -2.102164
```

(continued)

Exhibit 9.3 (continued)

```
6 0CANONICAL DISCRIMINANT FUNCTIONS EVALUATED AT GROUP MEANS (GROUP CENTROIDS)
  0   GROUP        FUNC   1     FUNC   2

        1          0.57699      0.71465
        2          0.44436      0.41718
        3         -1.11255     -0.01062
        4          0.15853     -0.93050
```

discriminant function [5]. Exhibit 9.4 gives the relevant output from the ONEWAY procedure to test which pairs of groups (i.e., brands) are different with respect to the two discriminant functions. It can be seen that with respect to the first discriminant function, groups 1, 2, and 4 (i.e., brands A, B, and D) are not significantly different from each other [1], and each of these three groups is significantly different from group 3 (i.e., brand C). Groups 1 and 2 (i.e., brands A and B) and groups 2 and 3 (i.e., brands B and C) are not significantly different with respect to the second discriminant function. The next obvious question is: In what respect are these brands different or similar?

Exhibit 9.4

Range Tests for the Beer Example

```
    LSD PROCEDURE RANGES FOR THE 0.050 LEVEL -

  0    (*) DENOTES PAIRS OF GROUPS SIGNIFICANTLY DIFFERENT
           AT THE 0.050 LEVEL

1                                G G G G
                                 r r r r
                                 p p p p

         Mean        Group       3 4 2 1

       -1.1125       Grp 3
         .1585       Grp 4       *
         .4444       Grp 2       *
         .5770       Grp 1       *

-        Variable Z2

    LSD PROCEDURE RANGES FOR THE 0.050 LEVEL -

2                                G G G G
                                 r r r r
                                 p p p p

         Mean        Grpup       4 3 2 1

        -.9305       Grp 4
        -.0106       Grp 3       *
         .4172       Grp 2       *
         .7147       Grp 1       * *
```

This question can be answered by assigning labels to the discriminant functions.[6] As discussed in the following section, the loadings can be used to label the discriminant functions and plot the attributes in the discriminant space.

9.4.1 Labeling the Discriminant Functions

The structure matrix gives the simple correlation between the attributes and the discriminant scores [3, Exhibit 9.3]. The higher the loading of a given attribute on a function, the more representative a function is of that attribute. However, just as in the case of factor analysis, the structure matrix can be rotated to obtain a simple structure. Only the varimax rotation is available in the discriminant analysis procedure. As discussed in Chapter 5, varimax rotation attempts to obtain loadings such that each variable loads primarily on only one discriminant function. As per the rotated loadings [4], *Flavor* and *Taste* have a high loading on the first discriminant function and therefore this function is labeled "Quality" to represent the quality of the beer. *Filling* and *Light* load more on the second discriminant function so it is labeled as "Lightness" to represent the lightness of beer. Attribute *Mellow* loads equally well on both the dimensions, implying that this attribute is partially represented by both the functions. That is, it is possible that the mellowness of a beer may be implying both Quality and Lightness. The foaminess attribute does not load highly on any function and, therefore, is not very helpful in assigning labels to the functions. It should be noted that based on the univariate F-ratio the mean of this attribute was not significantly different across the four brands. Notice that the rotated loadings [4] are quite different from the unrotated loadings [3].

9.4.2 Examining Differences in Brands

In many applications MDA is used mostly to examine differences among groups and not to classify future observations. Also, in most of the applications the group differences are typically represented by two discriminant functions. Consequently, one can plot the group centroids [6] to provide further insights about group differences. Such a plot is commonly referred to as a *perceptual map*. A perceptual map gives a visual representation of the differences among groups with respect to the key dimensions (i.e., the discriminant functions).

A plot of the group centroids is shown in Figure 9.6. It was previously concluded that brands A, B, and D do not differ with respect to quality (i.e., the first discriminant function), and brands A and B, and brands B and C are not different with respect to lightness (i.e., the second discriminant function). This leads to the following conclusions: (1) Brands A and B are not different from each other with respect to both lightness and quality, but are different from the other brands. That is, consumers perceive brands A and B as light, high quality beers and consider them quite different from the rest of the brands; (2) Brand D is perceived to be a quality beer that is not light; and (3) Brand C, a private label beer, is perceived to have the lowest quality.

One can also plot the attributes in the perceptual map. The loadings are essentially coordinates of the attributes with respect to the discriminant functions. Consequently, they can be represented as vectors in the discriminant space. Figure 9.7 gives such a plot. This plot can be used to determine the rating or ranking of each brand on each of the attributes. This can be done by dropping a perpendicular from the stimulus (i.e., brand) onto the attributes. For example, rankings of brands with respect to the *Aftertaste* attribute are: brands A, B, D, and C.

[6]The problem of assigning labels to the discriminant functions is the same as that of assigning labels to the principal components and the factors.

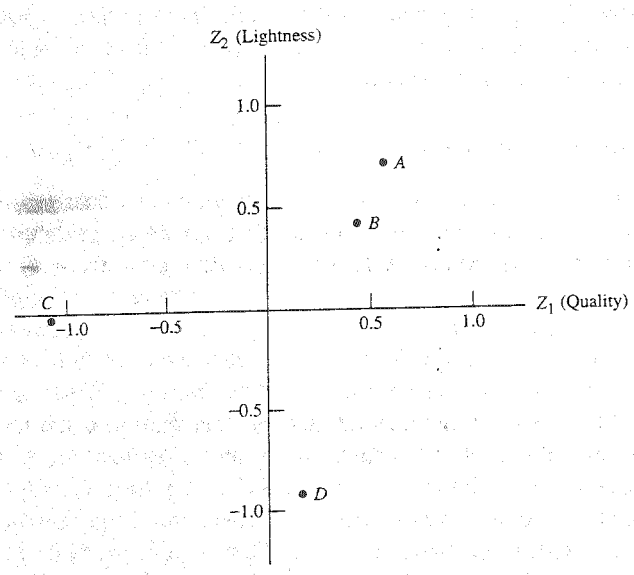


Figure 9.6 Plot of brands in discriminant space.

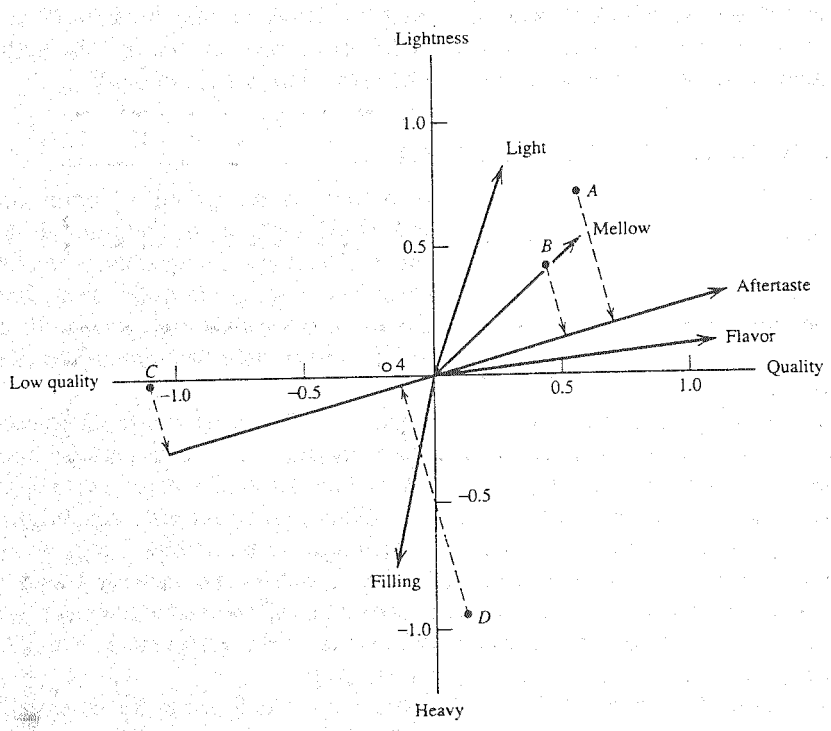


Figure 9.7 Plot of brands and attributes.

9.5 SUMMARY

In this chapter we discussed MDA, which is a generalization of two-group discriminant analysis. It was seen that geometrically MDA reduced to identifying a set of new axes such that the projection of the points onto the first axis accounted for the maximum differences among the groups. The projection of the points onto the second axis accounted for the maximum of what

was not accounted for by the first axis and so on until all the axes [i.e., $min\ (G-1, p)$] were identified.

Classification of future observations into different groups was treated as a separate procedure and an extension of the MDA procedure. Geometrically, classification reduced to dividing the variable space or the discriminant space into mutually exclusive and collectively exhaustive regions. Any given observation is classified into the group into whose region it falls. A marketing example was used to illustrate the use of MDA to assess differences among brands. Furthermore, it was also shown how MDA can be used to develop perceptual maps. In the next chapter we discuss logistic regression as an alternative procedure to two-group discriminant analysis.

QUESTIONS

9.1 Tables Q9.1(*a*) and (*b*) present data on two variables X_1 and X_2. In each case plot the data in two-dimensional space and visually inspect the plots to discuss the following:

(i) Number of distinct groups
(ii) Discrimination provided by X_1
(iii) Discrimination provided by X_2
(iv) Discrimination provided by X_1 and X_2
(v) Minimum number of discriminant functions required to adequately represent the differences between the groups.

Table Q9.1(a)

Obs.	1	2	3	4	5	6	7	8	9	10	11	12
X_1	0.4	3.3	3.0	2.0	4.0	0.9	5.2	2.5	0.5	4.8	2.0	3.4
X_2	1.1	6.0	6.4	4.5	8.2	1.0	8.4	4.0	1.5	7.9	4.0	6.5

Table Q9.1(b)

Obs.	1	2	3	4	5	6	7	8	9	10	11	12
X_1	1.9	1.8	2.0	6.2	2.0	6.4	1.6	6.2	7.0	6.7	7.0	2.4
X_2	1.0	6.1	6.0	4.1	1.8	1.7	6.6	4.6	4.7	1.0	1.8	1.5

9.2 Refer to Tables Q9.1(*a*) and (*b*). In each case compute the discriminant functions (minimum number required to adequately represent the differences between the groups).

Hint: To compute the discriminant functions, follow the steps shown in Question 8.2.

Plot the discriminant scores and use suitable cutoff points/lines to divide the discriminant space into distinct regions. Comment on the accuracy of classification.

9.3 Use **SPSS** (or any other software) to perform discriminant analysis on the data shown in Tables Q9.1(*a*) and (*b*).

Compare the results to those obtained in Question 9.2.

Hint: Use the plots of the data to create a grouping variable.

9.4 Perform discriminant analysis on the data given in file FIN.DAT. Use the industry type as the grouping variable. Interpret the solution and discuss the differences between the different types of industries. Comment on the classification accuracy.

9.5 Refer to Question 7.6 and the data in file FIN.DAT. Using the cluster memberships as the grouping variable, perform discriminant analysis on the data. Discuss the differences between the segments and comment on the classification accuracy.

9.6 The two-group discriminant analysis problem can be formulated as a multiple regression problem. Can the multiple-group discriminant analysis problem be similarly formulated? Justify.

9.7 The matrix of discriminant coefficients is usually rotated using the varimax criterion, to improve the interpretability of the discriminant analysis solution. In what ways does rotation change/not change the unrotated solution? How does rotation improve the interpretability of the discriminant analysis solution?

9.8 File PHONE.DAT gives data on the attitudinal responses of families owning one, two, or three or more telephones. 250 respondents were asked to indicate their extent of agreement with the six attitudinal statements using a 0–10 scale where 0 = do not agree at all and 10 = agree completely. The statements are given in file PHONE.DOC.

Use discriminant analysis to determine the relative importance of the above six attitudes in differentiating between families owning different numbers of telephones. Discuss the differences between the three types of families (those owning one, two, or three or more telephones).

9.9 File ADMIS.DAT gives admission data for a graduate school of business.[a] Analyze the data using discriminant analysis. Interpret the solution and comment on the admission policy of the business school.

9.10 Note the following:

$$\boldsymbol{\mu}_1 = (2\ 3)' \qquad \boldsymbol{\mu}_2 = (4\ 5)' \qquad \boldsymbol{\mu}_3 = (6\ 7)'$$

where $\boldsymbol{\mu}_i$ is the vector of means for group i. Divide the two-dimensional space into three classification regions (assume $\boldsymbol{\Sigma} = \mathbf{I}$).

Hint: Estimate the equations for the cutoff lines first.

9.11 Assume a single independent variable with group means given by:

$$\mu_1 = 2 \qquad \mu_2 = 4 \qquad \mu_3 = 6$$

and group variances given by:

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1.$$

Also, assume that the independent variable has a normal distribution and that the priors and misclassification costs are equal.

If the cutoff values are given by $c_1 = 3$ and $c_2 = 5$, compute the following probabilities of misclassification:

(a) $P(2 \mid 1)$; (b) $P(3 \mid 1)$; (c) $P(1 \mid 2)$;
(d) $P(3 \mid 2)$; (e) $P(1 \mid 3)$; (f) $P(2 \mid 3)$.

What effect will unequal misclassification costs and unequal priors have on the cutoff values?

Appendix

Estimation of discriminant functions and classification procedures are direct generalizations of the concepts discussed in the Appendix to Chapter 8. In this Appendix we provide further discussion of classification to show how the variable space can be divided into multiple classification regions.

[a]Johnson, R. A., and D. W. Wichern (1988), *Applied Multivariate Statistical Analysis*, Prentice Hall, Englewood Cliffs, New Jersey, Table 11.5, p. 539.

A9.1 CLASSIFICATION FOR MORE THAN TWO GROUPS

Classifying observations into G groups is a direct generalization of the two-group case, and entails the division of the discriminant or the discriminating variable space into G mutually exclusive and collectively exhaustive regions.

Let $f_i(\mathbf{x})$ be the density function for population π_i, $i = 1, \ldots, G$, where G is the number of groups; p_i be the prior probability of population π_i; $C(j/i)$ be the cost of misclassifying an observation to group j that belongs to group i; and R_j the region in which the observation should fall for it to be classified into group j.

The probability $P(j/i)$ of misclassifying an observation belonging to group i into group j is given by

$$P(j/i) = \int_{R_j} f_i(\mathbf{x})\, d\mathbf{x}. \tag{A9.1}$$

The total cost of misclassifying an observation belonging to group i is given by

$$\sum_{j=1, j\neq i}^{G} P(j/i) \cdot C(j/i), \tag{A9.2}$$

and the expected total cost of misclassifying observations belonging to group i will be

$$\mathrm{TCM}_i = p_i \left[\sum_{j=1, j\neq i}^{G} P(j/i) \cdot C(j/i) \right]. \tag{A9.3}$$

The resulting total expected cost of misclassification, TCM, for all groups will be

$$\begin{aligned} \mathrm{TCM} &= \sum_{i=1}^{G} \mathrm{TCM}_i \\ &= \sum_{i=1}^{G} p_i \sum_{j=1, j\neq i}^{G} P(j/i) \cdot C(j/i). \end{aligned} \tag{A9.4}$$

The classification problem reduces to choosing $R_1, R_2, \ldots, R_j$ such that Eq. A9.4 is minimized. It can be shown that the solution of Eq. A9.4 results in the following allocation rule (see Anderson (1984), p. 224):

Allocate to π_j (i.e., group j), $j = 1, 2, \ldots, G$, for which

$$\sum_{i=1, i\neq j}^{G} p_i \cdot f_i(\mathbf{x}) \cdot C(j/i) \tag{A9.5}$$

is smallest.

A9.1.1 Equal Misclassification Costs

If the misclassification costs are equal, Eq. A9.5 reduces to:

Allocate to π_j, $j = 1, 2, \ldots, G$, for which

$$\sum_{i=1, i\neq j}^{G} p_i \cdot f_i(\mathbf{x}) \tag{A9.6}$$

is the smallest.

Table A9.1 **Illustrative Example**

	True Group Membership		
Classify into Group	**1**	**2**	**3**
1	$C(1/1) = 0$	$C(1/2) = 30$	$C(1/3) = 300$
2	$C(2/1) = 5$	$C(2/2) = 0$	$C(2/3) = 25$
3	$C(3/1) = 100$	$C(3/2) = 250$	$C(3/3) = 0$
Prior probability (p_i)	.10	.70	.20
Density function ($f_i(\mathbf{x})$)	.10	.70	.40

It can be clearly seen that the value obtained from Eq. A9.6 will be smaller if $p_j \cdot f_j(\mathbf{x})$ is largest. Therefore, the allocation rule can be restated as:
Allocate a given observation, $\mathbf{x}$, to π_j if

$$p_j f_j(\mathbf{x}) > p_i f_i(\mathbf{x}) \qquad \text{for all } i \neq j \tag{A9.7}$$

or

$$\ln[p_j f_j(\mathbf{x})] > \ln[p_i f_i(\mathbf{x})] \qquad \text{for all } i \neq j. \tag{A9.8}$$

The classification rule given by Eqs. A9.7 and A9.8 is the same as:
Assign any given observation, $\mathbf{x}$, to π_j such that its posterior probability $p(\pi_j/\mathbf{x})$, given by

$$\frac{p_j f_j(\mathbf{x})}{\sum_{j=1}^{G} p_j f_j(\mathbf{x})}, \tag{A9.9}$$

is the largest.

Observe that posterior probabilities given by Eq. A9.9 are the same as posterior probabilities given by Eq. A8.20 of the Appendix to Chapter 8.

A9.1.2 Illustrative Example

Classification rules discussed in the previous section are illustrated in this section using a numerical example. Consider the information given in Table A9.1. For group 1, the resulting value for Eq. A9.5, obtained by substituting the appropriate numbers from Table A9.1, is

$$\sum_{j=2}^{3} p_j \cdot f_j(\mathbf{x}) \cdot C(j/1) = .70 \times .70 \times 5 + .20 \times .40 \times 100 = 10.45.$$

Similarly, the values for groups 2 and 3 are 20.3 and 15.25, respectively. Therefore, the observation is assigned to group 1. If equal misclassification costs are assumed then the resulting values obtained from Eq. A9.6 for groups 1, 2, and 3, respectively, are .57, .09, and .500. Therefore, for equal misclassification costs the observation is assigned to group 2. The posterior probabilities can be easily computed from Eq. A9.9 and are equal to 0.017, .845, and .138 for groups 1, 2, and 3, respectively.

A9.2 MULTIVARIATE NORMAL DISTRIBUTION

In the previous section, classification rules were developed for any density function. In this section classification rules will be developed by assuming that the discriminating variables come from a multivariate normal distribution. Furthermore, equal misclassification costs will be assumed.

The classification rule given by Eq. A9.8 can be restated as:
Assign x to π_j if

$$\ln[p_j f_j(\mathbf{x})] \tag{A9.10}$$

is maximum for all j. Since x is assumed to come from a multivariate normal distribution, this rule reduces to:
Allocate to π_j if

$$-p/2\ln(2\pi) - 1/2\ln|\Sigma_j| + \ln p_j - 1/2(\mathbf{x}-\boldsymbol{\mu}_j)'\Sigma_j^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) \tag{A9.11}$$

is maximum for all j.

The first term of Eq. A9.11 is a constant and can be ignored. Under the assumption that

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_G,$$

the second term in Eq. A9.11 can also be ignored resulting in the following rule:
Allocate to π_j if

$$\ln p_j - 1/2(\mathbf{x}-\boldsymbol{\mu}_j)'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}_j) \tag{A9.12}$$

is maximum for all j.

Equation A9.12 can be further simplified to give the following rule:
Allocate to π_j if

$$d_j = \boldsymbol{\mu}_j'\Sigma^{-1}\mathbf{x} - 1/2\boldsymbol{\mu}_j'\Sigma^{-1}\boldsymbol{\mu}_j + \ln p_j \tag{A9.13}$$

is maximum for all j.

The term $\boldsymbol{\mu}_j'\Sigma^{-1}$ in this equation contains the coefficients of the classification function, and the constant of the classification function is given by the second and third terms. That is, d_j from Eq. A9.13 is the score obtained from the classification function of group j. Note that prior probabilities affect only the constant and not the coefficients of the classification function. Furthermore, classification functions of linear discriminant analysis assume equal variance–covariance matrices for all the groups and a multivariate normal distribution for the discriminating variables. As shown in the following section, these classification functions can be used to develop classification regions.

A9.2.1 Classification Regions

As mentioned previously, the classification problem can also be viewed as a partitioning problem. That is, the total discriminant or the discriminating variable space is partitioned into G mutually exclusive and collectively exhaustive regions. This section presents a procedure for obtaining the G classification regions under the assumption that the data come from a multivariate distribution and the misclassification costs are equal.

Any given observation x can be assigned to group j if the following conditions are satisfied

$$d_j(\mathbf{x}) > d_i(\mathbf{x})$$

for all $i \neq j$. That is, the classification rule is
Allocate x to π_j if

$$d_j(\mathbf{x}) - d_i(\mathbf{x}) > 0 \qquad \text{for all } i \neq j \tag{A9.14}$$

or

$$\boldsymbol{\mu}_j'\Sigma^{-1}\mathbf{x} - 1/2\boldsymbol{\mu}_j'\Sigma^{-1}\boldsymbol{\mu}_j + \ln p_j - \boldsymbol{\mu}_i'\Sigma^{-1}\mathbf{x} + 1/2\boldsymbol{\mu}_i'\Sigma^{-1}\boldsymbol{\mu}_i - \ln p_i > 0 \tag{A9.15}$$

or

$$(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)'\Sigma^{-1}\mathbf{x} - 1/2(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)'\Sigma^{-1}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_i) > \ln\left(\frac{p_i}{p_j}\right) \tag{A9.16}$$

or

$$d_{ji} > \ln\left(\frac{p_i}{p_j}\right) \qquad \text{for all } i \neq j \tag{A9.17}$$

where

$$d_{ji} = (\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}\mathbf{x} - 1/2(\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_i). \tag{A9.18}$$

Equation A9.17 can be used for dividing the space into the G regions.

For example, consider the case of three groups and two variables. The discriminating variable space can be divided into three regions by solving the following equations:[1]

$$\begin{aligned} d_{12}(\mathbf{x}) &\geq \ln\left(\frac{p_2}{p_1}\right) \\ d_{13}(\mathbf{x}) &\geq \ln\left(\frac{p_3}{p_1}\right) \\ d_{23}(\mathbf{x}) &\geq \ln\left(\frac{p_3}{p_2}\right). \end{aligned} \tag{A9.19}$$

If the centroids of the three groups do not lie in a straight line, then the intersection of the lines defined by Eqs. A9.19 will result in the three regions depicted in Panel I of Figure A9.1. On the other hand, if the centroids of the three groups lie in a straight line then the lines defined by these three equations will be parallel, as depicted in Panel II of Figure A9.1. In the case of p variables, Eqs. A9.19 define hyperplanes that partition the p-dimensional space into G regions.

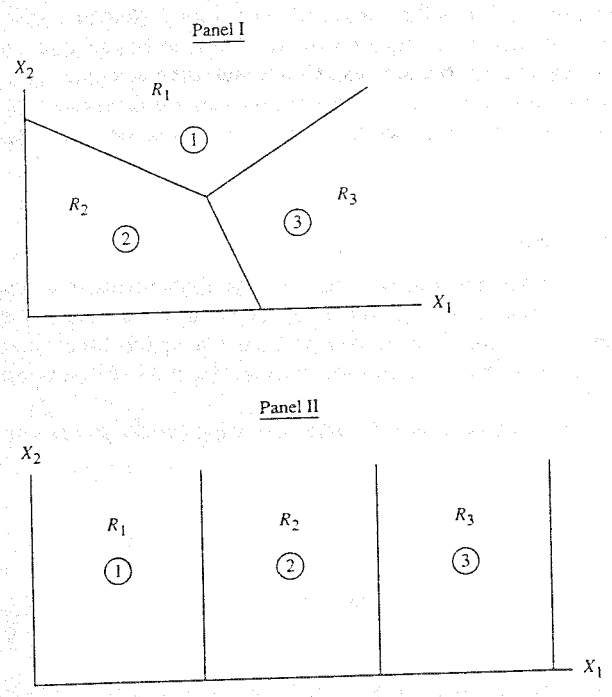


Figure A9.1 Classification regions for three groups.

[1]In order to facilitate the solution, the > is changed to ≥.

Illustrative Example

Consider the case of four groups and two discriminating variables. Let $\boldsymbol{\mu}_1' = (3\,3)$, $\boldsymbol{\mu}_2' = (13\,3)$, $\boldsymbol{\mu}_3' = (3\,13)$, and $\boldsymbol{\mu}_4' = (13\,13)$. Figure A9.2 shows the group centroids. For computational ease assume equal priors and $\boldsymbol{\Sigma} = \mathbf{I}$. According to Eq. A9.17 any observation $\mathbf{x}$ will be assigned to group 1 if $d_{12} \geq 0, d_{13} \geq 0$, and $d_{14} \geq 0$. From Eq. A9.19

$$d_{12} = (3-13 \quad 3-3)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 0.5(3-13 \quad 3-3)\begin{pmatrix} 3+13 \\ 3+3 \end{pmatrix},$$

which reduces to

$$\begin{aligned} -10x_1 + 80 &\geq 0 \\ x_1 &\leq 8. \end{aligned} \tag{A9.20}$$

Similarly, d_{13} and d_{14} will, respectively, result in the following equations

$$x_2 \leq 8 \tag{A9.21}$$

and

$$x_1 + x_2 \leq 16. \tag{A9.22}$$

That is, the observation will be classified into group 1 if Eqs. A9.20, A9.21, and A9.22 are satisfied. Figure A9.3 also gives the lines representing the preceding equations, and their graphical solution to obtain region R_1 is given by the shaded area. Table A9.2 gives the equations for obtaining all the classification regions (i.e., R_1, R_2, R_3, and R_4) and Figure A9.3 gives the classification regions resulting from the graphical solution of the equations.

A9.2.2 Mahalanobis Distance

For equal priors, the classification rule given by Eq. A9.12 can be restated as:
Assign to π_j if

$$-1/2(\mathbf{x} - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_j) \tag{A9.23}$$

is maximum, or

$$(\mathbf{x} - \boldsymbol{\mu}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_j) \tag{A9.24}$$

is minimum.

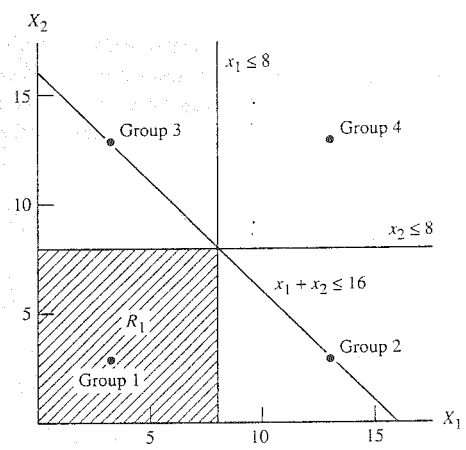


Figure A9.2 Group centroids.

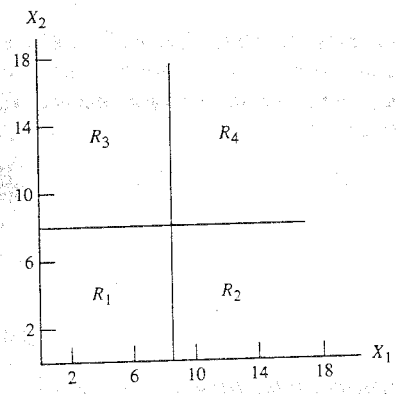


Figure A9.3 Classification regions R_1 to R_4.

Table A9.2 **Conditions and Equations for Classification Regions**

Classification Region	Conditions	Equations
R_1	$d_{12} \geq 0$	$x_1 \leq 8$
	$d_{13} \geq 0$	$x_2 \leq 8$
	$d_{14} \geq 0$	$x_1 + x_2 \leq 16$
R_2	$d_{21} \geq 0$	$x_1 \geq 8$
	$d_{23} \geq 0$	$x_1 \geq x_2$
	$d_{24} \geq 0$	$x_2 \leq 8$
R_3	$d_{31} \geq 0$	$x_2 \geq 8$
	$d_{32} \geq 0$	$x_2 \leq x_1$
	$d_{34} \leq 0$	$x_1 \leq 8$
R_4	$d_{41} \geq 0$	$x_1 + x_2 \geq 16$
	$d_{42} \geq 0$	$x_2 \geq 8$
	$d_{43} \geq 0$	$x_1 \geq 8$

Equation A9.24 gives the statistical or Mahalanobis distance of observation $\mathbf{x}$ from the centroid of the group j. Therefore, any given observation, $\mathbf{x}$, is assigned to the group to which it is closest as measured by the Mahalanobis distance. That is, classification based on Mahalanobis distance assumes equal priors and equal misclassification costs, and a multivariate normal distribution for the discriminating variables.

CHAPTER 10

Logistic Regression

Consider the following scenarios:

- A medical researcher is interested in determining whether the probability of a heart attack can be predicted given the patient's blood pressure, cholesterol level, calorie intake, gender, and lifestyle.
- The marketing manager of a cable company is interested in determining the probability that a household would subscribe to a package of premium channels given the occupant's income, education, occupation, age, marital status, and number of children.
- An auditor is interested in determining the probability that a firm will fail given a number of financial ratios and the size of the firm (i.e., large or small).

Discriminant analysis could be used for addressing each of the above problems. However, because the independent variables are a mixture of categorical and continuous variables, the multivariate normality assumption will not hold. In these cases one could use logistic regression as it does not make any assumptions about the distribution of the independent variables. That is, logistic regression is normally recommended when the independent variables do not satisfy the multivariate normality assumption.

In this chapter we discuss the use of logistic regression. Unfortunately, logistic regression does not lend itself well to a geometric illustration. Consequently, we use simple data sets to illustrate the technique and once again the treatment will be nonmathematical. We first discuss the case when the only independent variable is categorical and show that logistic regression in this case reduces to a contingency table analysis. The illustration is then extended to the case where the independent variables are a mixture of categorical and continuous variables and the best variable(s) are selected via stepwise logistic regression analysis.

10.1 BASIC CONCEPTS OF LOGISTIC REGRESSION

10.1.1 Probability and Odds

Consider the data given in Table 10.1 for a sample of 12 most-successful (*MS*) and 12 least-successful (*LS*) financial institutions (FI). Table 10.2 gives a contingency table between success and the size of the FI, and from that table the following probabilities can be computed:

Table 10.1 **Data for Most Successful and Least Successful Financial Institutions**

Most Successful			Least Successful		
SUCCESS	*SIZE*	FP	*SUCCESS*	*SIZE*	FP
1	1	0.58	2	1	2.28
1	1	2.80	2	0	1.06
1	1	2.77	2	0	1.08
1	1	3.50	2	0	0.07
1	1	2.67	2	0	0.16
1	1	2.97	2	0	0.70
1	1	2.18	2	0	0.75
1	1	3.24	2	0	1.61
1	1	1.49	2	0	0.34
1	1	2.19	2	0	1.15
1	0	2.70	2	0	0.44
1	0	2.57	2	0	0.86

Notes: The value of SUCCESS is equal to 1 for the most successful financial institution and is equal to 2 for the least successful financial institution. The value of SIZE is equal to 1 for the large financial institution and is equal to 0 for the small financial institution.

Table 10.2 **Contingency Table for Type and Size of Financial Institution**

	Size		
Type of Financial Institution	**Large**	**Small**	**Total**
Most successful (MS)	10	2	12
Least successful (LS)	1	11	12
Total	11	13	24

1. Probability that any given FI will be *MS* is

$$P(MS) = 12/24 = .50.$$

2. Probability that any given FI will be *MS* given that the FI is large (L)

$$P(MS|L) = 10/11 = .909.$$

3. Probability that any FI is *MS* given that the FI is small (S)

$$P(MS|S) = 2/13 = .154.$$

In many instances probabilities are stated as *odds*. For example, we frequently hear about the odds of a given football team winning the Super Bowl, or the odds of smokers getting lung cancer, or the odds of winning a state lottery. From Table 10.2 the following odds can be computed:

1. Odds of a FI being *MS* are

$$odds(MS) = 12/12 = 1,$$

implying that the odds of any given FI being most or least successful are equal, or the odds are 1 to 1.

2. Odds of a FI being *MS* given that it is large are

$$odds(MS|L) = 10/1 = 10, \tag{10.1}$$

implying that the odds of a large FI being most successful are 10 to 1. That is, the odds of large FI being most successful are 10 times than its being least successful.

3. Odds of a FI being most successful given that it is a small FI are

$$odds(MS|S) = 2/11 = .182, \tag{10.2}$$

implying that the odds of a small FI being most successful are 2 to 11 or .182 to 1.

Odds and probabilities provide the same information, but in different forms. It is easy to convert odds into probabilities and vice versa. For example,

$$\begin{aligned} P(MS|L) &= \frac{odds(MS|L)}{1 + odds(MS|L)} \\ &= \frac{10}{1+10} = .909 \end{aligned}$$

and

$$\begin{aligned} odds(MS|L) &= \frac{P(MS|L)}{1 - P(MS|L)} \\ &= \frac{.909}{1 - .909} = 10. \end{aligned}$$

10.1.2 The Logistic Regression Model

Taking the natural log of the odds given by Eqs. 10.1 and 10.2 we get

$$\begin{aligned} \ln[odds(MS|L)] &= \ln(10) = 2.303 \\ \ln[odds(MS|S)] &= \ln(0.182) = -1.704. \end{aligned}$$

These two equations can be combined into the following equation to give the log of the odds as a function of the size (i.e., *SIZE*) of the FI:

$$\ln[odds(MS|SIZE)] = -1.704 + 4.007 \times SIZE, \tag{10.3}$$

where $SIZE = 1$ if the FI is large and $SIZE = 0$ if the FI is small. It is clear from Eq. 10.3 that the log of the odds is a *linear* function of the independent variable $SIZE$, the size of the FI. The coefficient of the independent variable, $SIZE$, can be interpreted like the coefficient in regression analysis. The positive sign of the $SIZE$ coefficient means that the log of the odds increases as $SIZE$ increases; that is, the log of the odds of a large FI being most successful is greater than that of a small FI. In general, Eq. 10.3 for k independent variables can be written as

$$\ln[odds(MS|X_1, X_2, \ldots, X_k)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{10.4}$$

or

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{10.5}$$

where

$$odds(MS|X_1, X_2, \ldots, X_k) = \frac{p}{1-p},$$

and p is the probability of a FI being most successful given the independent variables, $X_1, X_2, \ldots, X_k$. Equation 10.5 models the log of the odds as a linear function of the independent variables, and is equivalent to a multiple regression equation with log of the odds as the dependent variable. The independent variables can be a combination of continuous and categorical variables. Since the log of the odds is also referred to as *logit*, Eq. 10.5 is commonly referred to as *multiple logistic regression* or in short as *logistic regression*. The following discussion provides further justification for referring to Eq. 10.5 as logistic regression.

For simplicity, assume that there is only one independent variable. Equation 10.5 can be rewritten as

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 X_1 \tag{10.6}$$

or

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1)}}. \tag{10.7}$$

Figure 10.1 gives the relationship between probability, p, and the independent variable, X_1. It can be seen that the relationship between probability and the independent variable is represented by a logistic curve that asymptotically approaches one as X_1 approaches positive infinity and zero as X_1 approaches negative infinity. The function that gives the relationship between probability and the independent variables is known as the linking function, which is logit for the above model. Other linking functions such as *normit* or *probit* (i.e., the inverse of the cumulative standard normal distribution function) and complementary log-log function (i.e., inverse of the Gompertz function) can also be used. In this chapter we use the logit function as it is the most popular linking function. For further information the interested reader is referred to Agresti (1990), Cox and Snell (1989), Freeman (1987), and Hosmer and Lemeshow (1989).

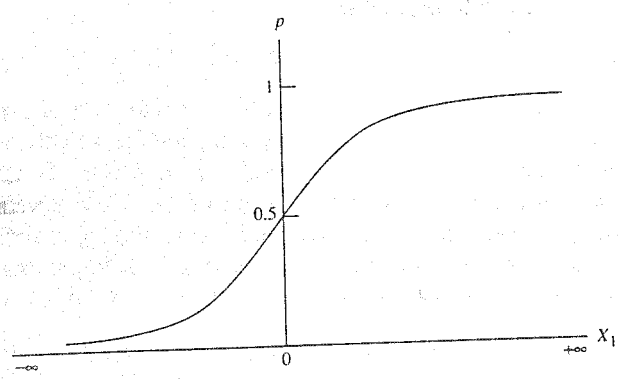


Figure 10.1 The logistic curve.

Note that the relationship between probability p and the independent variable is *nonlinear,* whereas the relationship between the log of the odds and the independent variable is *linear*. Consequently, the interpretation of the coefficients of the independent variables should be with respect to their effects on the log of odds and *not* on the probability, p.

A maximum likelihood estimation procedure can be used to obtain the parameter estimates. Since no analytical solutions exist, an iterative procedure is employed to obtain the estimates. The Appendix provides a brief discussion of the maximum likelihood estimation technique for logistic regression analysis. In the following sections we illustrate the use of the PROC LOGISTIC procedure in SAS for obtaining the estimates of the logistic regression model. The data given in Table 10.1 are used for illustration purposes. The data in Table 10.1 give the size of the FI and a measure of financial performance (FP).[1] First we consider the output when the only independent variable is categorical (i.e., *SIZE*). Next we illustrate that when the only independent variable is categorical, logistic regression analysis reduces to an analysis of the contingency or cross-tabulation table. This is followed by a discussion of the output containing a mixture of categorical and continuous independent variables.

10.2 LOGISTIC REGRESSION WITH ONLY ONE CATEGORICAL VARIABLE

Logistic regression is first run using *SIZE* as the only independent variable. Table 10.3 gives the SAS commands and Exhibit 10.1 gives the resulting output. The MODEL subcommand specifies the logit model along with the option that requests the classification table and a detailed set of measures for evaluating model fit. The OUTPUT subcommand specifies that an output data set be created that in addition to the original variables contains the new variable *PHAT*, which gives the predicted probabilities. To facilitate discussion, the circled numbers in the exhibit correspond to the bracketed numbers in the text.

10.2.1 Model Information

The basic information about the logit model is printed [1]. The response or the dependent variable SUCCESS is assumed to be ordered (i.e., ordinal) with two response levels.[2] Response levels of 1 and 2, respectively, correspond to *MS* and *LS* financial institutions. Logit is the link function used for logistic regression analysis.

Table 10.3 **SAS Commands for Logistic Regression**

```
PROC LOGISTIC;
 MODEL SUCCESS = SIZE
  /CTABLE;
 OUTPUT OUT=PRED P=PHAT;
PROC PRINT;
 VAR SUCCESS SIZE PHAT;
```

[1]FP could be any of the financial performance measures (such as the spread) used to evaluate the performance of financial institutions.

[2]Strictly speaking, the notion of ordered response variable comes into play only when there are more than two outcomes for the response variable.

Exhibit 10.1

Logistic regression analysis with one categorical variable as the independent variable

1 Response Variable: SUCCESS
Link Function: Logit

Number of Observations: 24
Response Levels: 2

Response Profile

Ordered Value	SUCCESS	Count
1	1	12
2	2	12

2 Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
2a AIC	35.271	21.864	.
2b SC	36.449	24.221	.
2c -2 LOG L	33.271	17.864	15.407 with 1 DF (p=0.0001)
2d Score	.	.	13.594 with 1 DF (p=0.0002)

3 Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	-1.7047	0.7687	4.9181	0.0266	.
SIZE	4.0073	1.3003	9.4972	0.0021	1.124514

4a Association of Predicted Probabilities and Observed Responses

Concordant	= 76.4%	Somers' D	= 0.750
Discordant	= 1.4%	Gamma	= 0.964
Tied	= 22.2%	Tau-a	= 0.391
(144 pairs)		c	= 0.875

4b Classification Table

		Predicted EVENT	Predicted NO EVENT	Total
Observed	EVENT	10	2	12
	NO EVENT	1	11	12
	Total	11	13	24

Sensitivity= 83.3% Specificity= 91.7% Correct= 87.5%
False Positive Rate= 9.1% False Negative Rate= 15.4%

NOTE: An EVENT is an outcome whose ordered response value is 1.

(continued)

Exhibit 10.1 (continued)

5

OBS	SUCCESS	SIZE	PHAT	OBS	SUCCESS	SIZE	PHAT
1	1	1	0.90909	13	2	1	0.90909
2	1	1	0.90909	14	2	0	0.15385
3	1	1	0.90909	15	2	0	0.15385
4	1	1	0.90909	16	2	0	0.15385
5	1	1	0.90909	17	2	0	0.15385
6	1	1	0.90909	18	2	0	0.15385
7	1	1	0.90909	19	2	0	0.15385
8	1	1	0.90909	20	2	0	0.15385
9	1	1	0.90909	21	2	0	0.15385
10	1	1	0.90909	22	2	0	0.15385
11	1	0	0.15385	23	2	0	0.15385
12	1	0	0.15385	24	2	0	0.15385

10.2.2 Assessing Model Fit

The logistic regression model is formed using *SIZE* as the predictor or independent variable. The first step is to assess the overall fit of the model to the data. A number of statistics are provided for this purpose [2]. The null and the alternative hypotheses for assessing overall model fit are given by

$$H_o : \text{The hypothesized model fits the data.}$$
$$H_a : \text{The hypothesized model does not fit the data.}$$

These hypotheses are similar to ones used in Chapter 6 for testing the overall fit of a confirmatory factor model to the sample data. Obviously, nonrejection of the null is desired, as it leads to the conclusion that the model fits the data. The statistic used is based on the likelihood function. The likelihood, L, of a model is defined as the probability that the estimated hypothesized model represents the input data. To test the null and alternative hypotheses, L is transformed to $-2LogL$. The $-2LogL$ statistic, sometimes referred to as the likelihood ratio χ^2 statistic, has a χ^2 distribution with $n - q$ degrees of freedom where q is the number of parameters in the model.

The output provides two $-2LogL$ statistics: one for a model that includes *only the intercept* (i.e., the model does not include any independent variables) and the other for a model that includes the *intercept and the covariates*. Note that in the logistic regression procedure **SAS** refers to the independent variables as covariates. From the output the value of $-2LogL$ for the model with only the intercept is 33.271 and it has a χ^2 distribution with 23 *df* (i.e., 24 − 1) [2c]. Although not reported in the output, the value of 33.271 is significant at an alpha of .05 and the null hypothesis is rejected, implying that the hypothesized model with only the intercept does not fit the data. The $-2LogL$ value of 17.864 with 22 *df* (i.e., 24 − 2) for the model that includes the intercept and the independent variable is not significant at an alpha of .05, suggesting that the null model cannot be rejected. That is, the model containing the intercept and the independent variable *SIZE* does fit the data.

The $-2LogL$ statistic can also be used to determine if the addition of the independent variables significantly improves model fit. This is equivalent to testing whether the coefficients of the independent variables are significantly different from zero. The corresponding null and alternative hypotheses are:

$$H_o : \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \qquad H_a : \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} \neq \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

These hypotheses can be tested by using the χ^2 difference test—a test similar to the one discussed in Chapter 6. The difference between the $-2LogL$ for the model with the intercept and the independent variables, and the $-2LogL$ for the model with only the intercept is distributed as a χ^2 distribution with the *df* equal to the difference in the respective degrees of freedom. From the output we see that the difference between the two $-2LogL$'s is equal to 15.407 (i.e., 33.271 − 17.864) with 1 *df* (i.e., 23 − 22) and is statistically significant [2c]. Therefore, the null hypothesis can be rejected, implying that the coefficient of the *SIZE* variable is significantly different from zero. That is, the inclusion of the independent variable, *SIZE*, significantly improves model fit. In other words, the independent variable *SIZE* does contribute to predicting the success of the FI.

The hypotheses pertaining to whether the coefficients are significantly different from zero can also be established using the χ^2 statistic reported in the row labeled *Score* [2d]. This statistic, which is not based on the likelihood function, has an asymptotic χ^2 distribution with p degrees of freedom, where p is the number of independent variables. The estimated coefficient of *SIZE* is significantly different from zero, as the χ^2 value of 13.594 with 1 *df* is statistically significant at an alpha of .05 [2d]. That is, there is a relationship between the dependent variable (*SUCCESS*) and the independent variable (*SIZE*).

Other measures of goodness-of-fit are the Akaike's information criterion (AIC) and Schwartz's criterion (SC), and they are essentially $-2LogL$'s adjusted for degrees of freedom [2a, 2b]. These two statistics do not have a sampling distribution and are normally used as heuristics for comparing fit of different models estimated using the same data set. Lower values of these statistics imply a better fit. For example, during a stepwise logistic regression analysis the researcher can use these heuristics to determine when to stop including variables in the model. However, there are no specific guidelines regarding how low is "low." We suggest using the likelihood ratio χ^2 test (i.e., $-2LogL$ test statistic) as it is based on widely accepted maximum likelihood estimation theory.

10.2.3 Parameter Estimates and Their Interpretation

The maximum likelihood estimates of model parameters are reported next [3]. The logistic regression model can be written as

$$\ln \frac{p}{1-p} = -1.705 + 4.007 \times SIZE. \tag{10.8}$$

Note that the coefficients of this equation, within rounding errors, are the same as those of Eq. 10.3. The standard error of the coefficients can be used to compute the t-values, which are −2.218 (−1.7047/.7687) and 3.082 (4.0073/1.3003), respectively, for *INTERCEPT* and *SIZE*. The squares of these t-values give the Wald χ^2 statistic, which can be used to assess the statistical significance of each independent variable. As can be seen, both coefficients are statistically significant at an alpha of .05.

The estimates of the coefficients of the independent variable are interpreted just like the regression coefficients in multiple regression. The coefficient of the independent variable gives the amount by which the dependent variable will increase if the independent variable changes by one unit. In Eq. 10.8 the value of 4.007 for the *SIZE* coefficient indicates that the log of the odds of being most successful would increase by 4.007 if the value of the independent variable increases by 1. Since the value of *SIZE* is 1 for a large FI and 0 for a small FI, the log of odds of being most successful increases by

4.007 for a large FI. It should be noted that the relationship between the log of odds and the independent variables is linear; however, as we will see, the relationship between odds and the independent variable is nonlinear. Consequently, the interpretation of the effect of independent variables on the odds also changes.

Equation 10.8 can be rewritten as

$$\begin{aligned}\frac{p}{1-p} &= e^{(-1.705+4.007\times SIZE)} \\ &= e^{(-1.705)} \times e^{(4.007\times SIZE)}.\end{aligned} \tag{10.9}$$

From this equation we see that the effect of the independent variables on the dependent variable is nonlinear or multiplicative. For a unit increase in *SIZE* the odds of being most successful increases by a factor of 54.982 (i.e., $e^{4.007}$). In other words, the odds of a FI being most successful are 54.982 times higher for a large FI than for a small FI.

The probability of being most successful can be calculated by rewriting Eq. 10.9 as follows:

$$p = \frac{1}{1 + e^{-(-1.705+4.007\times SIZE)}}.$$

From this equation, the estimate of the probability of a FI being most successful given that it is small (i.e., value of $SIZE = 0$) is

$$p = \frac{1}{1 + e^{-}(-1.705)} = .154,$$

and the probability that a FI is most successful given that it is a large FI (i.e., value of $SIZE = 1$) is

$$p = \frac{1}{1 + e^{-}(-1.705 + 4.007)} = .909.$$

10.2.4 Association of Predicted Probabilities and Observed Responses

The association of predicted probabilities and observed responses can be assessed by a number of statistics, such as Somers's D, Gamma, and Tau-a and c. These statistics assess the rank order correlation between *PHAT* and the observed responses [4a]. These correlations are obtained by first determining the total number of pairs, the number of *concordant*, *discordant*, and *tied* pairs, and then transforming them into rank order correlations to give a measure of the association between observed responses for the dependent variable and *PHAT*.

In logistic regression terminology an *event* is defined as an outcome whose response value is 1 and a *no-event* as an outcome whose response value is other than 1 (e.g., 2). In this particular case, the most successful FI is defined as an event and the least successful FI as a no-event. The total number of pairs is the product of the number of events and no-events. A *concordant* pair is defined as that pair formed by an event and a no-event such that the *PHAT* of the event is higher than the *PHAT* of the no-event. A *discordant* pair is one in which the *PHAT* for an event is less than the *PHAT* for a no-event. Tied pairs are ones that are neither concordant nor discordant. The total number of pairs is equal to 144. It can be seen from the predicted probabilities (i.e., PHATs) [5] that for a total of two pairs (Obs 13 & 12, and Obs 13 & 11) the *PHAT* of no-event is greater than the *PHAT* of event, and therefore these two pairs or 1.4% (2/144)

of the pairs are discordant. Similarly, it can be seen that a total of 110 pairs (76.4%) are concordant and a total of 32 pairs (22.2%) are tied. These statistics are reported in the output [4a]. Obviously, the higher the number of *concordant pairs* the greater the association between observed responses and the predicted probabilities. Exact formulae for converting the different types of pairs into rank order correlations are given in the SAS manual. The rank order correlations do not have a sampling distribution and also there is no clear guidance as to which one is preferred. Furthermore, for Tau-a the maximum value is not one and is dependent on the number of pairs in the data set. Therefore, it is normally recommended that these measures be used to compare the correlations of different models fitted to the same data set.

10.2.5 Classification

Classification of observations is done by first estimating the probabilities. The output reports the estimated probability, *PHAT*, of each observation belonging to a given group [5]. For example for observation 1 the estimated probability that it is a successful FI is

$$p = \frac{1}{1 + e^{-2.302}} = .909.$$

Note that *PHAT*s for all the large FIs are 0.909 and for small FIs they are equal to 0.154. This is because all the large FIs have the same value for *SIZE* and all the small FIs have the same value for *SIZE*. These probabilities can be used to classify observations into the two groups. Classification of observations into groups is based on a cutoff value for *PHAT*, which is usually assumed to be 0.5. All observations whose *PHAT* is greater than or equal to 0.5 are classified as most successful and those whose value is less than 0.5 are classified as least successful. Table 10.4 gives the resulting classification table. In the present case, an 87.5% (21 ÷ 24) classification is substantially greater than the naive classification rate of 50%, suggesting that the model has good predictive validity.[3] The statistical significance of the classification rate can be assessed by using Huberty's procedure discussed in Section 8.3.3. Using Eq. 8.20 the expected number of correct classifications due to chance is equal to

$$e = \frac{1}{24}(12^2 + 12^2) = 12,$$

and from Eq. 8.18

$$Z^* = \frac{(21 - 12)\sqrt{24}}{\sqrt{12(24 - 12)}} = 3.674,$$

Table 10.4 **Classification Table**

	Predicted		
Actual	**Most Successful**	**Least Successful**	**Total**
Most successful	10	2	12
Least successful	1	11	12
Total	11	13	24

[3]The naive classification rate is defined as that which is obtained if one were to classify all observations into one of the categories.

which is statistically significant at an alpha of .05. That is, the classification rate of 87.5% is significantly higher than that expected by chance alone.

Normally the classification of observations is based on a cutoff value of 0.5 for *PHAT*. The program gives the researcher the option of specifying the cutoff value. Misclassification costs could also be incorporated in computing the cutoff value. Although SAS does not give the user the option of directly specifying misclassification costs, the program can be tricked to incorporate them. The procedure is the same as discussed in Chapter 8.

Classification of data using models whose parameters are estimated using the same data is biased. Ideally one would like to use a fresh sample to obtain an unbiased estimate of the classification rate. Alternatively, one could use the holdout method discussed in Chapter 8, but only if the sample is sufficiently large. In cases where a fresh sample is not available or the holdout method is not possible due to small sample sizes, one can use the jackknife estimate of the classification rate. In the jackknife method an observation is deleted, the model is estimated using the remaining observations, and the estimated model is used to predict the holdout observation. This procedure is continued for all the n observations. It is clear that computationally obtaining the jackknife estimate could be quite cumbersome. Approximate procedures, which are computationally efficient, have been developed for obtaining pseudo-jackknife estimates of classification rates for logistic regression models. The classification table and the classification rates reported by SAS are obtained by using one such pseudo-jackknife estimation procedure (for further details see the SAS manual).

It can be seen that the classification table and classification rate reported in the output are the same as Table 10.4 [4b]. However, there can be situations where the two may not be the same. Sensitivity is the percentage of correct classifications for events, i.e., the percent of the most-successful financial institutions that have been classified correctly by the model, and specificity is the percentage of correct classifications for no-events. Similarly, the false positive and false negative rates are, respectively, the percentage of incorrect classifications for an event and no-event.

10.3 LOGISTIC REGRESSION AND CONTINGENCY TABLE ANALYSIS

As mentioned earlier, in the case of only one independent categorical variable, logistic regression essentially reduces to a contingency table analysis. Exhibit 10.2 gives a partial SAS output for the contingency table analysis for the data in Table 10.1. Note that the cross-tabulation table is exactly the same as the one given in Table 10.2. The usual null and alternative hypotheses for the contingency table are

H_o : There is no relationship between *SUCCESS* and *SIZE*.

H_a : There is a relationship between *SUCCESS* and *SIZE*.

All the χ^2 statistics indicate that the null hypothesis can be rejected at an alpha of .05 [1]. That is, there is a relationship between *SUCCESS* and *SIZE*. Since we know *a priori* that *SIZE* is the independent variable, we can conclude that the size of the FI does have an effect on its performance. Note that the χ^2 values reported are the same as the ones reported in Exhibit 10.1 [2c, 2d]. The cross-tabulation analysis reports a number of correlations to assess the predictive ability of the model, whereas logistic regression analysis reports only a few. Note that the correlations reported by the two techniques are the same [2].

Exhibit 10.2

Contingency analysis output

```
TABLE OF SUCCESS BY SIZE
SUCCESS      SIZE
Frequency|
Percent  |
Row Pct  |
Col Pct  |        1|        2|  Total
---------+---------+---------+
       1 |      10 |       2 |     12
         |   41.67 |    8.33 |  50.00
         |   83.33 |   16.67 |
         |   90.91 |   15.38 |
---------+---------+---------+
       2 |       1 |      11 |     12
         |    4.17 |   45.83 |  50.00
         |    8.33 |   91.67 |
         |    9.09 |   84.62 |
---------+---------+---------+
Total           11        13        24
             45.83     54.17    100.00

STATISTICS FOR TABLE OF SUCCESS BY SIZE

Statistic                          DF      Value         Prob
(1)-----------------------------------------------------------
Chi-Square                          1     13.594        0.000
Likelihood Ratio Chi-Square         1     15.407        0.000
Continuity Adj. Chi-Square          1     10.741        0.001

Statistic                                  Value          ASE
(2)-----------------------------------------------------------
Gamma                                      0.964        0.046
Kendall's Tau-b                            0.753        0.133
Stuart's Tau-c                             0.750        0.134
Somers' D C|R                              0.750        0.134
Somers' D R|C                              0.755        0.132
```

In short, the preceding analysis suggests that a 2×2 contingency table can be analyzed using logistic regression. In fact, one can use logistic regression to analyze a $2 \times j$ table. In cases where the dependent and the independent variables are categorical one would typically use categorical data analytic methods, which are beyond the scope of the present text. Further details on categorical data analysis can be found in Freeman (1987).

10.4 LOGISTIC REGRESSION FOR COMBINATION OF CATEGORICAL AND CONTINUOUS INDEPENDENT VARIABLES

In this section we consider the case where the set of independent variables is a combination of categorical and continuous variables. In addition, we illustrate the use of

Table 10.5 **SAS Commands for Stepwise Logistic Regression**

```
PROC LOGISTIC;
 MODEL SUCCESS = FP SIZE
  /CTABLE SELECTION=S SLENTRY=0.15 SLSTAY=0.15 DETAILS;
 OUTPUT OUT=PRED P=PHAT;
PROC PRINT;
 VAR SUCCESS FP SIZE PHAT;
```

stepwise logistic regression. In order to illustrate stepwise logistic regression analysis, assume that we would like to develop a logistic regression model that includes the *best* set of independent variables from *SIZE* and FP. Stepwise logistic regression analysis is similar to stepwise regression analysis, or stepwise discriminant analysis discussed in Chapter 8, and consequently the usual caveats apply. That is, in the presence of multicollinearity the "best" set of variables may differ from sample to sample, and therefore caution is advised in interpreting the results of stepwise logistic regression analysis. Table 10.5 gives necessary commands for stepwise logistic regression. The SELECTION=S option requests a stepwise analysis. SLENTRY and SLSTAY specify, respectively, the p-values for entering and removing a variable in the model. DETAILS requests a detailed output of the stepwise process. The OUTPUT subcommand requests the creation of a new data set PRED that, in addition to the original variables, includes the predicted probability *PHAT* of a given FI as being most successful. Exhibit 10.3 gives the SAS output.

10.4.1 Stepwise Selection Procedure

In the first step, labeled Step 0, the intercept is entered into the model [1]. The residual χ^2 value of 16.551 reported in the output is the incremental χ^2 that would result if all the independent variables that have not yet been included are included in the model [1a]. Since at Step 0 only the intercept is included in the model, the residual χ^2 value at this step essentially tests the joint statistical significance of the independent variables.

The important question is: Which independent variable should be included in the next step? This question can be answered by examining the χ^2 values of the variables not included in the model. For each variable the increase in the overall χ^2 value, if this variable is included in the model, is examined [2]. The reported χ^2 value should be interpreted just like the partial F-value in stepwise regression or stepwise discriminant analysis. The stepwise procedure selects the variable that has the highest χ^2 value and meets the p-value criterion set for inclusion in the model. Therefore at Step 1, the procedure selects FP to be included in the model [3].

At Step 2, the variable *SIZE* entered since its partial χ^2 value meets the significance criterion set for including variables in the model [3a, 4]. Since there are no additional variables to be entered, the procedure stops and the final model includes all the variables. A summary of the stepwise procedure is provided in the output [4d]. Interpretation of the fit statistics and parameter estimates have been discussed earlier. All the fit statistics indicate good model fit and a statistically significant relationship between the independent and the dependent variables [4a].

The final model is given by [4b]:

$$\ln \frac{p}{1-p} = -4.445 + 3.055 \times SIZE + 1.925 \times \text{FP}, \tag{10.10}$$

Exhibit 10.3

Logistic regression for categorical and continuous variables

Stepwise Selection Procedure

1 Step 0. Intercept entered:

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	0	0.4082	0.0000	1.0000	.

1a Residual Chi-Square = 16.5512 with 2 DF (p=0.0003)

2 Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > Chi-Square
SIZE	13.5944	0.0002
FP	13.8301	0.0002

3 Step 1. Variable FP entered:

Analysis of Variables Not in the Model

3a

Variable	Score Chi-Square	Pr > Chi-Square
SIZE	5.0283	0.0249

4 Step 2. Variable SIZE entered:

Criteria for Assessing Model Fit

4a

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	35.271	17.789	.
SC	36.449	21.323	.
-2 LOG L	33.271	11.789	21.482 with 2 DF (p=0.0001)
Score	.	.	16.551 with 2 DF (p=0.0003)

4b Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	-4.4450	1.8432	5.8159	0.0159	.
SIZE	3.0552	1.5981	3.6550	0.0559	0.857342
FP	1.9245	0.9116	4.4570	0.0348	1.139820

4c Association of Predicted Probabilities and Observed Responses

Concordant	= 95.8%	Somers' D	= 0.917
Discordant	= 4.2%	Gamma	= 0.917
Tied	= 0.0%	Tau-a	= 0.478
(144 pairs)		c	= 0.958

NOTE: All explanatory variables have been entered into the model.

(continued)

Exhibit 10.3 (continued)

(4d)

Summary of Stepwise Procedure

Step	Variable Entered	Variable Removed	Number In	Score Chi-Square	Wald Chi-Square	Pr > Chi-Square
1	FP		1	13.8301	.	0.0002
2	SIZE		2	5.0283	.	0.0249

(5)

Classification Table

		Predicted EVENT	Predicted NO EVENT	Total
Observed	EVENT	9	3	12
	NO EVENT	1	11	12
	Total	10	14	24

Sensitivity= 75.0% Specificity= 91.7% Correct= 83.3%
False Positive Rate= 10.0% False Negative Rate= 21.4%

(5a) NOTE: An EVENT is an outcome whose ordered response value is 1.

OBS	SUCCESS	SIZE	FP	PHAT	OBS	SUCCESS	SIZE	FP	PHAT
1	1	1	0.58	0.43202	13	2	1	2.28	0.95248
2	1	1	2.80	0.98199	14	2	0	1.06	0.08278
3	1	1	2.77	0.98094	15	2	0	1.08	0.08575
4	1	1	3.50	0.99525	16	2	0	0.07	0.01325
5	1	1	2.67	0.97699	17	2	0	0.16	0.01572
6	1	1	2.97	0.98695	18	2	0	0.70	0.04319
7	1	1	2.18	0.94297	19	2	0	0.75	0.04735
8	1	1	3.24	0.99220	20	2	0	1.61	0.20641
9	1	1	1.49	0.81421	21	2	0	0.34	0.02208
10	1	1	2.19	0.94400	22	2	0	1.15	0.09692
11	1	0	2.70	0.67939	23	2	0	0.44	0.02664
12	1	0	2.57	0.62265	24	2	0	0.86	0.05787

or

$$\frac{p}{1-p} = e^{(-4.445+3.055\times SIZE+1.925\times \mathrm{FP})}$$
$$= e^{-4.445} \times e^{3.055\times SIZE} \times e^{1.925\times \mathrm{FP}}. \qquad (10.11)$$

From Eq. 10.10 it can be seen that the log of the odds of being a most successful FI is positively related to *FP* and *SIZE*. Specifically, for a FI of a given size (i.e., large or small) each unit increase in *FP* increases the log of odds of being most successful by 1.925. And if *FP* is held constant, the log of odds of being most successful increases by 3.055 for a large FI as compared to a small FI.

Equation 10.11 gives the relationship between odds and the independent variables. From the equation we conclude that, everything held constant, the odds of being the most successful FI are increased by a factor of 21.221 (i.e., $e^{3.055}$) for a unit increase in *SIZE*. That is, after adjusting or controlling for the effects of other variables, the odds of being most successful are 21.221 times higher for a large FI than for a small FI. Similarly, everything else being constant, the odds of being a most successful FI are increased by a factor of 6.855 (i.e., $e^{1.925}$) for a unit change in *FP*.

Table 10.6 **Classification Table for Cutoff Value of 0.5**

	Predicted		
Actual	Most Successful	Least Successful	Total
Most successful	11	1	12
Least successful	1	11	12
Total	12	12	24

The correlations between the observed responses and *PHAT* [4c] indicate that the fit of this model is better than the previous model (i.e., the model discussed in Section 10.2; see Exhibit 10.1). Table 10.6 gives the classification matrix, which was obtained by using the reported *PHAT*s [5a] and a cutoff value of .50. From the table it is clear that the overall classification rate of 91.67% is better than the overall classification rate of 87.5% (see Table 10.4) for the previous model. However, the overall classification rate of 83.3% reported in the output [5] is lower than that of the previous model. This is because, as discussed earlier, in order to correct for bias SAS uses the pseudo-jackknifing procedure for classifying observations. Therefore, although the addition of FP is statistically significant, its addition does not help in classifying observations. In fact, the addition of *FP* is detrimental to the bias-adjusted classification rate. Consequently, if classification is the major objective then the previous model, which does not include *FP*, is to be preferred.

10.5 COMPARISON OF LOGISTIC REGRESSION AND DISCRIMINANT ANALYSIS

The data in Table 10.1 were analyzed using discriminant analysis. Exhibit 10.4 gives the partial SPSS output. The overall discriminant function is significant, suggesting that the means of the independent variables for the two groups are significantly different [1]. This conclusion is consistent with that drawn using logistic regression analysis. The nonstandardized coefficient estimates suggest that both independent variables have a positive impact on the success of a FI [2], and this conclusion is also consistent with that obtained from logistic regression analysis. The discriminant analysis procedure correctly classifies 91.67% of the observations [3], which is the same as the biased classification rate of logistic regression given in Table 10.6, but higher than the unbiased classification rate (see [5]; Exhibit 10.3). In sum, there are no appreciable differences in the results of the two techniques for this particular data set. It is possible though that the results could be quite different for other data sets. In such cases, which of the two techniques should be used?

The choice between the two techniques is dependent on the assumptions made by the two techniques. Discriminant analysis assumes that the data come from a multivariate normal distribution, whereas logistic regression analysis makes no such distributional assumptions. As discussed in Chapter 8, violation of the multivariate normality assumption affects the significance tests and the classification rates. Since the multivariate normality assumption will clearly be violated for a mixture of categorical and continuous variables, we suggest that in such cases one should use logistic regression analysis. In the case when there are no categorical variables, logistic regression should be used when the multivariate assumption is violated, and discriminant analysis should be used when

Exhibit 10.4

Discriminant analysis for data in Table 10.1

```
                    Canonical Discriminant Functions
                 Pct of   Cum  Canonical  After  Wilks'
 Fcn Eigenvalue Variance  Pct     Corr      Fcn  Lambda  Chi-square  df  Sig
(1)                                     :    0 .310367     24.570     2  .0000
  1*    2.2220  100.00  100.00    .8304 :

   * Marks the 1 canonical discriminant functions remaining in the analysis.

Unstandardized canonical discriminant function coefficients
(2)
                   Func  1
SIZE             1.8552118
FP                .9162471
(Constant)      -2.3834923

Classification results -

(3)                      No. of    Predicted Group Membership
   Actual Group           Cases          1          2
--------------------     ------    --------   --------

Group        1              12         11          1
                                     91.7%       8.3%

Group        2              12          1         11
                                      8.3%      91.7%

Percent of "grouped" cases correctly classified:  91.67%
```

the multivariate normality assumption is not violated because discriminant analysis is computationally more efficient.

10.6 AN ILLUSTRATIVE EXAMPLE

Consider the case where an investor is interested in developing a model to classify mutual funds that are attractive for investment. Suppose the following measures are available for 138 mutual funds: (1) *SIZE*, size of the mutual fund with 1 representing a large fund and 0 representing a small fund; (2) *SCHARGE*, sales charge in percent; (3) *EXPENRAT*, expense ratio in percent; (4) *TOTRET*, total return in percent; and (5) *YIELD*, five-year yield in percent. It is also known that 59 of the 138 funds have been previously recommended by stockbrokers as most attractive (the other 79 were identified as least attractive). The objective is to develop a model to predict the probability that a given mutual fund would be most attractive given its values for the above measures. The rating of funds by stockbrokers will be the dependent variable. Exhibit 10.5 gives the partial output.

The stepwise procedure selected all the independent variables [2]. The χ^2 value of 135.711 with 132 *df* (i.e., 138 − 6) is not statistically significant, implying that the estimated model, containing the intercept and all the independent variables, fits the

Exhibit 10.5

Logistic regression for mutual fund data

```
Response Variable: RATE
Response Levels: 2
Number of Observations: 138
Link Function: Logit

     Response Profile
Ordered
  Value     RATE      Count
      1        1         59
      2        2         79


Stepwise Selection Procedure

                    Criteria for Assessing Model Fit
(1)                                (1a)                (1b)
                                Intercept
                 Intercept         and
Criterion          Only        Covariates     Chi-Square for Covariates

AIC              190.400        147.711          .
SC               193.327        165.275          .
-2 LOG L         188.400        135.711        52.689 with 5 DF (p=0.0001)
Score              .              .            44.034 with 5 DF (p=0.0001)

NOTE: All explanatory variables have been entered into the model.

(2)                         Summary of Stepwise Procedure
          Variable                Number      Score        Wald          Pr >
Step   Entered      Removed        In     Chi-Square   Chi-Square   Chi-Square
   1   YIELD                       1       21.0379          .         0.0001
   2   TOTRET                      2       11.9103          .         0.0006
   3   SIZE                        3        8.5928          .         0.0034
   4   SCHARGE                     4        4.1344          .         0.0420
   5   EXPENRAT                    5        5.5516          .         0.0185

(3)                 Analysis of Maximum Likelihood Estimates

            Parameter     Standard       Wald          Pr >      Standardized
Variable     Estimate       Error     Chi-Square    Chi-Square     Estimate
INTERCPT      -2.5902       1.2642       4.1981        0.0405            .
SIZE           0.8542       0.4773       3.2020        0.0735       0.236320
SCHARGE       -0.1394       0.0589       5.6088        0.0179      -0.302154
EXPENRAT      -1.4361       0.6793       4.4699        0.0345      -0.321113
TOTRET         0.8090       0.2509      10.3988        0.0013       0.402480
YIELD          0.0553       0.0124      19.9669        0.0001       0.694773

(4) Association of Predicted Probabilities and Observed Responses

 Concordant = 85.5%          Somers' D = 0.711
 Discordant = 14.4%          Gamma     = 0.712
 Tied       =  0.1%          Tau-a     = 0.351
 (4661 pairs)                c         = 0.856
```

(continued)

Exhibit 10.5 (continued)

```
                    Classification Table

(5)                              Predicted
                              EVENT    NO EVENT        Total
                        +----------------------+
              EVENT     ]       45          14 ]         59
Observed                ]                      ]
              NO EVENT  ]       12          67 ]         79
                        +----------------------+
              Total             57          81          138

Sensitivity= 76.3%  Specificity= 84.8%  Correct= 81.2%
False Positive Rate= 21.1%  False Negative Rate= 17.3%
```

data [1a]. The same information is given by the χ^2 value of 44.034 with 5 *df* and is statistically significant ($p = .0001$), suggesting that there is a relationship between the independent variables and the log of odds of a mutual fund being attractive [1b]. All the variables are significant at $p \leq .10$ [3]. The parameter estimates suggest that, as expected, the effects of *SIZE*, *TOTRET*, and *YIELD* on the log of odds of a mutual fund being most attractive are positive, and the effects of *SCHARGE* and *EXPENRAT* are negative. The most influential variable for log of odds is *EXPENRAT* of the mutual fund. The log of odds of a mutual fund being most attractive, after controlling for the effects of all other measures, decreases by 1.4361 for a unit increase in *EXPENRAT*. On the other hand, after controlling for the effects of all other variables, the odds of a mutual fund being most attractive change by a factor of only .238 (i.e., $e^{-1.4361}$) for a unit increase in *EXPENRAT*.

The model appears to have good predictive validity. The classification rate of 81.2% is substantially greater than the naive prediction rate of 57.2% (i.e., 79/138) [5]. Recall that the logistic regression procedure uses the jackknifed estimate of classification rate. The nonjackknifed estimate of classification rate is 82.6%. Using Eqs. 8.18 and 8.20, the Z^* statistic is equal to 7.076, suggesting that the classification rate of 81.2% is statistically significant. All the statistics for the association between probabilities and observed responses are also high, once again suggesting that the model has good predictive validity [4].

10.7 SUMMARY

This chapter provided a discussion of logistic regression, which is an alternative procedure to discriminant analysis. As opposed to discriminant analysis, logistic regression analysis does not make any assumptions regarding the distribution of the independent variables. Therefore, it is preferred to discriminant analysis when the independent variables are a combination of categorical and continuous because in such cases the multivariate normality assumption is clearly violated. Logistic regression analysis is also preferred when all the independent variables are continuous, but are not normally distributed. On the other hand, when the multivariate normality assumption is not violated then discriminant analysis is preferred because it is computationally more efficient than logistic regression analysis.

In the next chapter we discuss *m*ultivariate *an*alysis *o*f *va*riance (MANOVA). MANOVA is a generalization of analysis of variance and is very closely related to discriminant analysis. In fact, much of the output reported by discriminant analysis is also reported by MANOVA.

QUESTIONS

10.1 Both logistic regression and multiple regression belong to the category of generalized linear models. In what ways are the fundamental assumptions associated with the two methods different?

10.2 Consider the following models for k independent variables

$$p = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{1}$$

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k\,, \tag{2}$$

where p is the probability of one of two possible outcomes. While the model in (2) is the logistic regression model, the model in (1) is often referred to as the linear probability model. The difference between the two models lies in the transformation of the dependent variable from a simple probability measure to the log of the odds.

What are the problems associated with the linear probability model that make this transformation attractive (thereby rendering the logistic regression model a vast improvement over the linear probability model)?

10.3 Use the data in the contingency Table Q10.1 to answer the questions that follow.

Table Q10.1

Blood Cholesterol (mg/100 ml)	Heart Disease	
	Present	Absent
< 200	5	30
200 – 225	9	26
226 – 250	14	20
251 – 275	18	16
> 275	23	8

(a) What is the probability that
  (i) Heart disease is present?
  (ii) Blood cholesterol is less than 200 mg/100 ml?
  (iii) Blood cholesterol is greater than 225 mg/100 ml?
  (iv) Heart disease is absent given that blood cholesterol is between 200 and 225 mg/100 ml?
  (v) Heart disease is present given that blood cholesterol is greater than 250 mg/100 ml?
  (vi) Blood cholesterol is greater than 275 mg/100 ml given that heart disease is absent?

(b) What are the odds that
  (i) Heart disease is present (as against its being absent)?
  (ii) Blood cholesterol is less than 225 mg/100 ml (as against its being greater than or equal to 225 mg/100 ml)?
  (iii) Heart disease is present given that blood cholesterol is greater than 275 mg/100 ml?
  (iv) Blood cholesterol is greater than 250 mg/100 ml given that heart disease is present?

10.4 Refer to the admission data in file ADMIS.DAT. Consider the applicants who were admitted and those who were not admitted. Create a new variable that reflects the admission

status of the applicants, and let this variable take values of 0 and 1 respectively for admitted and not-admitted applicants. Recode the GPAs of the applicants to reflect a GPA category as follows:

GPA	GPA Category
< 2.50	1
2.51 to 3.00	2
3.01 to 3.50	3
> 3.50	4

Perform logistic regression on the data using the GPA category as the only independent variable (you will have to use a dummy variable coding to reflect the four GPA categories). Discuss the effect of the GPA category on the admission status and comment on the accuracy of classification.

Include the GMAT score as a second independent variable in the model and interpret the solution. Is there any improvement in the accuracy of classification?

10.5 Table Q10.2 presents data on 114 males between the ages of 40 and 65. These subjects were classified on their blood cholesterol level and subsequently on whether they developed heart disease.

Table Q10.2

Heart Disease	Blood Cholesterol (mg/100 cc)			
	< 200	200–219	220–259	> 259
Present	6	10	30	45
Absent	5	6	5	7

(a) Using hand calculations, compute the various probabilities and the log of odds.
(b) Using the above log of odds compute the logistic regression equation and interpret it (do not use any logistic regression software).
(c) Use a canned software (e.g., SAS or SPSS) to estimate the logistic regression model and compare the results to that obtained above (you will have to use dummy variable coding for the cholesterol levels).
(d) What conclusions can you draw from the classification table? Should the classification table be interpreted? Why or why not?

10.6 Refer to the data in file DEPRES.DAT (refer to file DEPRES.DOC for a description of the data). Using "CASES" as the dependent variable analyze the data by logistic regression. Interpret the solution and compare with the results obtained by discriminant analysis.

10.7 Table Q10.3 presents the data from a survey of 2409 individuals on their preferences for a particular branch of a savings institution. Each of the four variables has two possible responses: 1 indicates a positive answer and 2 indicates a negative answer.

Using B as the dependent variable, analyze the data by logistic regression. Interpret the solution and comment on the appropriateness of using logistic regression as against discriminant analysis.

*Table Q10.3**

Previous Patronage (A)	Recommend (B)	Convenient Location (C)	Familiarity (D) 1.0	Familiarity (D) 2.0
1	1	1	423	187
		2	459	412
	2	1	49	47
		2	68	127
2	1	1	13	84
		2	17	407
	2	1	0	22
		2	3	91

*Variable A is whether the person had previously patronized the branch; Variable B is whether the person strongly recommends the branch; Variable C is the person's opinion whether the branch is conveniently located; and Variable D is whether the person is familiar with the branch. Level 1 of each variable reflects a positive answer and level 2 a negative one.

Source: Dillon, W.R., & Goldstein, M. (1984), *Multivariate Analysis—Methods and Applications*, John Wiley & Sons Inc., New York.

10.8 Table Q10.4 presents data on the drinking habits of 5468 high school students.

Table Q10.4

	Student Drinks	Student Does Not Drink
Both parents drink	398	1401
One parent drinks	421	1814
Neither parent drinks	201	1233

Use logistic regression to analyze and interpret the above data.

10.9 Table Q10.5 is based on records of accidents in 1988 compiled by the Department of Highway Safety and Motor Vehicles in the State of Florida.

Table Q10.5

Safety Equipment in Use	Injury: Fatal	Injury: Nonfatal
None	1601	162527
Seat belt	510	412368

Source: Department of Highway Safety and Motor Vehicles, State of Florida.

(a) Perform a contingency table analysis on the above data.
(b) Use logistic regression to analyze the data.
(c) Compare the results obtained in (a) and (b).

10.10 A sample of elderly people is given psychiatric examinations to determine if symptoms of senility are present. One explanatory variable is the score on a subtest of the Wechsler Adult Intelligence Scale (WAIS). Table Q10.6 shows the data.

*Table Q10.6**

X	Y	X	Y	X	Y	X	Y	X	Y
9	1	7	1	7	0	17	0	13	0
13	1	5	1	16	0	14	0	13	0
6	1	14	1	9	0	19	0	9	0
8	1	13	0	9	0	9	0	15	0
10	1	16	0	11	0	11	0	10	0
4	1	10	0	13	0	14	0	11	0
14	1	12	0	15	0	10	0	12	0
8	1	11	0	13	0	16	0	4	0
11	1	14	0	10	0	10	0	14	0
7	1	15	0	11	0	16	0	20	0
9	1	18	0	6	0	14	0		

*X = WAIS Score; Y = Senility, where 1 = symptoms present and 0 = symptoms absent.

Source: Agresti, A. (1990), *Categorical Data Analysis*, John Wiley & Sons, New York.

(a) Analyze the data using logistic regression. Plot the predicted probabilities of senility against the WAIS values.

(b) Analyze the data using least-squares regression (i.e., fit the linear probability model). Plot the predicted probabilities of senility against the WAIS values.

(c) Compare the plots in (a) and (b) and discuss the appropriateness of using logistic regression versus simple regression.

Appendix

Maximum likelihood estimation is the most popular technique for estimating the parameters of the logistic regression model. In the following section we provide a brief discussion of the maximum likelihood estimation technique.

A10.1 MAXIMUM LIKELIHOOD ESTIMATION

The dependent variable in the logistic regression model is binary, which takes on two values. Let Y be the random binary variable whose value is zero or one. The probability $P(Y = 1)$ is given by

$$P(Y = 1) = p = \frac{e^{\beta \mathbf{X}}}{1 + e^{\beta \mathbf{X}}}, \tag{A10.1}$$

where $\boldsymbol{\beta}$ is the vector of coefficients and $\mathbf{X}$ is the vector of independent variables. This equation can be rewritten as

$$\ln \frac{p}{1 - p} = \boldsymbol{\beta}\mathbf{X}. \tag{A10.2}$$

Equation A10.2 represents the log of odds as a linear function of the independent variables. Unfortunately, the values for the dependent variable (i.e., the log of odds) are not available, so the parameters of Eq. A10.2 cannot be estimated directly. However, the likelihood function provides a solution to this problem.

Each observation can be considered as a Bernoulli trial. That is, it is a binomial with the total number of trials equal to 1. Consequently, for the ith observation

$$P(Y = y_i) = p_i^{y_i}(1 - p_i)^{1-y_i}. \tag{A10.3}$$

Assuming that all the n observations are independent, the likelihood function is given by

$$\begin{aligned} L &= \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{\boldsymbol{\beta}\mathbf{X}}}{1 + e^{\boldsymbol{\beta}\mathbf{X}}}\right)^{y_i} \left(\frac{1}{1 + e^{\boldsymbol{\beta}\mathbf{X}}}\right)^{1-y_i}, \end{aligned} \tag{A10.4}$$

and the log of the likelihood function is given by

$$\ln L = l = \sum_{i=1}^{n} y_i \left(\frac{e^{\boldsymbol{\beta}\mathbf{X}}}{1 + e^{\boldsymbol{\beta}\mathbf{X}}}\right) + \sum_{i=1}^{n} (1 - y_i)\left(\frac{1}{1 + e^{\boldsymbol{\beta}\mathbf{X}}}\right). \tag{A10.5}$$

The estimate for the parameter vector $\boldsymbol{\beta}$ is obtained by maximizing Eq. A10.5. The usual procedure is to take the first-order derivatives of the equation with respect to each independent variable, setting each equation to zero, and then solving the resulting equations. However, the resulting equations do not have an analytical solution. Consequently, $\hat{\boldsymbol{\beta}}$ is obtained by maximizing Eq. A10.5 using efficient iterative techniques such as the Newton-Raphson method (see Haberman 1978 for a discussion of the Newton-Raphson method).

A10.2 ILLUSTRATIVE EXAMPLE

Consider the example given in Section 10.2. The data are given in Table 10.1. Using Eq. A10.5, the likelihood function is given by

$$l = \sum_{i=1}^{n} y_i \left(\frac{e^{\beta_0 + \beta_1 X_1}}{1 + e^{\beta_0 + \beta_1 X_i}}\right) + \sum_{i=1}^{n} (1 - y_i)\left(\frac{1}{1 + e^{\beta_0 + \beta_1 X_i}}\right). \tag{A10.6}$$

Let us obtain the estimates of β_0 and β_1 by maximizing this equation using trial and error. Table A10.1 gives the value of the likelihood function for various values of β_0 and β_1. As can be seen, the likelihood function takes a maximum value of -8.932 when $\beta_0 = -1.7047$ and $\beta_1 = 4.0073$, and these estimates are the same as reported in Exhibit 10.1 [3]. Also, notice that

Table A10.1 **Values of the Maximum Likelihood Function for Different Values of β_0 and β_1**

	β_0			
β_1	−2.00	−1.7047	−1.5	−1.0
2	−13.275	−11.996	−11.333	−10.518
3	−10.096	−9.539	−9.334	−9.469
4.0073	−9.044	−8.932	−8.987	−9.610
5	−9.1857	−9.277	−9.446	−10.272

the value of 17.864 for $-2LogL$ reported in Exhibit 10.1 [2c] is twice the maximum value of $logL$ in Table A10.1 (i.e., $17.864 = -2 \times -8.932$).

Obviously, the trial-and-error method is an inefficient way of identifying values for the parameters that would result in the maximum value for the likelihood function. As mentioned previously, the Newton-Raphson method is an efficient technique and it is the one employed by the logistic regression procedure in SAS.

CHAPTER 11

Multivariate Analysis of Variance

Consider the following scenarios:

- A marketing manager is interested in determining if geographic region (e.g., north, south, east, and west) has an effect on consumers' taste preferences, purchase intentions, and attitudes toward the product.
- A medical researcher is interested in determining whether personality (e.g., Type A or Type B) has an effect on blood pressure, cholesterol, tension, and stress levels.
- A political analyst is interested in determining if party affiliation (Democratic, Republican, Independent) and gender have any effect on voters' views on a number of issues such as abortion, taxes, economy, gun control, and deficit.

For each of these examples we have categorical independent variable(s) with two or more levels, and a set of metric dependent variables. We are interested in determining if the categorical independent variable(s) affect the metric dependent variables. MANOVA (*m*ultivariate *an*alysis *o*f *va*riance) can be used to address each of the preceding problems. In MANOVA the independent variables are categorical and the dependent variables are continuous. MANOVA is a multivariate extension of ANOVA (*an*alysis *o*f *va*riance) with the only difference being that in MANOVA there are multiple dependent variables.

The reader will notice that the objective of MANOVA is very similar to some of the objectives of discriminant analysis. Recall that in discriminant analysis one of the objectives was to determine if the groups are significantly different with respect to a given set of variables. Although the two techniques are similar in this respect, there are some important differences. In this chapter we discuss the similarities and dissimilarities between MANOVA and discriminant analysis. The statistical tests in MANOVA, and consequently in discriminant analysis, are based on a number of assumptions. These assumptions are discussed in Chapter 12.

11.1 GEOMETRY OF MANOVA

A geometric illustration of MANOVA begins by first considering the case of one independent variable at two levels and one dependent variable. The illustration is then extended to the case of two dependent variables, followed by a discussion of p dependent variables. Finally, we discuss the case of more than one independent variable and p dependent variables.

11.1.1 One Independent Variable at Two Levels and One Dependent Variable

As shown in Figure 11.1, the centroid or mean (i.e., $\bar{Y}_1$ and $\bar{Y}_2$) of each group can be represented as a point in the one-dimensional space. If the independent variable has an effect on the dependent variable, then the means of the two groups are different (i.e., they are far apart) and the effect of the independent variable is measured by the difference between the two means (i.e., the distance between the two points). The extent to which the means of the two groups are different (i.e., far apart) can be measured by the euclidean distance between the centroids. However, as discussed in Chapter 3, Mahalanobis distance (*MD*) is the preferred measure of distance between two points.[1] The greater the *MD* between the two centroids, the greater the difference between the two groups with respect to Y and vice versa. Statistical tests are available to determine if the *MD* between the two centroids is large, i.e., significant at a given alpha level. Thus, geometrically, MANOVA is concerned with determining whether the *MD* between group centroids is significantly greater than zero. In the present case, because there are only two groups and one dependent variable, the problem reduces to comparing the means of two groups using a t-test. That is, a two-group independent sample t-test is a special case of MANOVA.

11.1.2 One Independent Variable at Two Levels and Two or More Dependent Variables

First consider the case where we have only two dependent variables. Since the independent variable is at two levels, there are two groups. Let Y and Z be the two dependent variables and $(\bar{Y}_1, \bar{Z}_1)$ and $(\bar{Y}_2, \bar{Z}_2)$, respectively, be the centroids of the two groups. As shown in Figure 11.2, the centroid of each group can be represented as a point or a

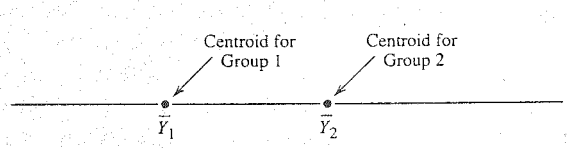


Figure 11.1 One dependent variable and one independent variable at two levels.

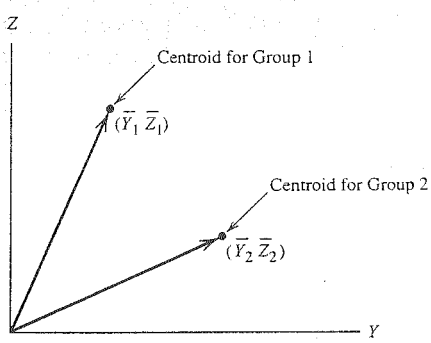


Figure 11.2 Two dependent variables and one independent variable at two levels.

[1]Recall that for uncorrelated variables the Mahalanobis distance reduces to the statistical distance, and if the variances of the variables are equal to one then the statistical distance is equal to the euclidean distance.

vector in the two-dimensional space defined by the dependent variables. Once again, the *MD* between the two points measures the distance between the centroids of the two groups. The larger the distance, the greater the difference between the two groups and vice versa. Once again, geometrically, MANOVA reduces to computing the distance between the centroids of the two groups and determining if the distance is statistically significant. In the case of p variables, the centroids of the two groups can be represented as two points in the p-dimensional space and the problem reduces to determining whether the distance between the two points is different from zero.

11.1.3 More Than One Independent Variable and p Dependent Variables

Consider the example described at the beginning of the chapter in which a political analyst is interested in determining the effect of two independent variables, voters' *Party Affiliation* and *Gender*, on voters' attitude towards a number of issues. In order to illustrate the problem geometrically, let us assume that two dependent variables, Y and Z, are used to measure voters' attitude towards two issues, say tax increase and gun control. Table 11.1 gives the means of the two dependent variables for the different cells. In the table, the first subscript refers to the level for *Gender* and the second subscript refers to the levels for *Party Affiliation*. The dot in the subscript indicates that the mean is computed across all levels of the respective subscript. For example, $\bar{Z}_{11}$ is the mean for male Democrats and $\bar{Z}_{.1}$ is the mean for all Democrats (i.e., male and female). There are three types of effects: (1) main effect of *Gender*; (2) main effect of *Party Affiliation*; and (3) the interaction effect of *Gender* and *Party Affiliation*. Panels I and II of Figure 11.3 give the geometrical representation of the main effects, and Panels III and IV, respectively, give the geometrical representations in the absence and presence of interaction effects.

In Panel I, the main effect of *Gender* is measured by the distance between the two centroids. Similarly, in Panel II the main effect of *Party Affiliation* is measured by the distances between pairs of the three centroids. There will be three distances, each representing the distance between pairs of groups. In Panel III, the solid circles give the centroids for Democrats, the open circles give the centroids for Republicans, and the stars give the centroids for Independents. The distance between the solid circles is a measure of *Gender* effect for Democrats. Similarly, the distances between open circles and stars, respectively, are measures of the *Gender* effect for Republicans and Independents. If the effect of *Gender* is independent of *Party Affiliation* then, as depicted in Panel III, the vectors joining the respective centroids should be parallel. On the other

Table 11.1 **Cell Means**

	Party Affiliation			
Gender	**Democrats**	**Republicans**	**Independents**	**Mean**
Male	$\bar{Z}_{11}$ $\bar{Y}_{11}$	$\bar{Z}_{12}$ $\bar{Y}_{12}$	$\bar{Z}_{13}$ $\bar{Y}_{13}$	$\bar{Z}_{1.}$ $\bar{Y}_{1.}$
Female	$\bar{Z}_{21}$ $\bar{Y}_{21}$	$\bar{Z}_{22}$ $\bar{Y}_{22}$	$\bar{Z}_{23}$ $\bar{Y}_{23}$	$\bar{Z}_{2.}$ $\bar{Y}_{2.}$
Mean	$\bar{Z}_{.1}$ $\bar{Y}_{.1}$	$\bar{Z}_{.2}$ $\bar{Y}_{.2}$	$\bar{Z}_{.3}$ $\bar{Y}_{.3}$	$\bar{Z}_{..}$ $\bar{Y}_{..}$

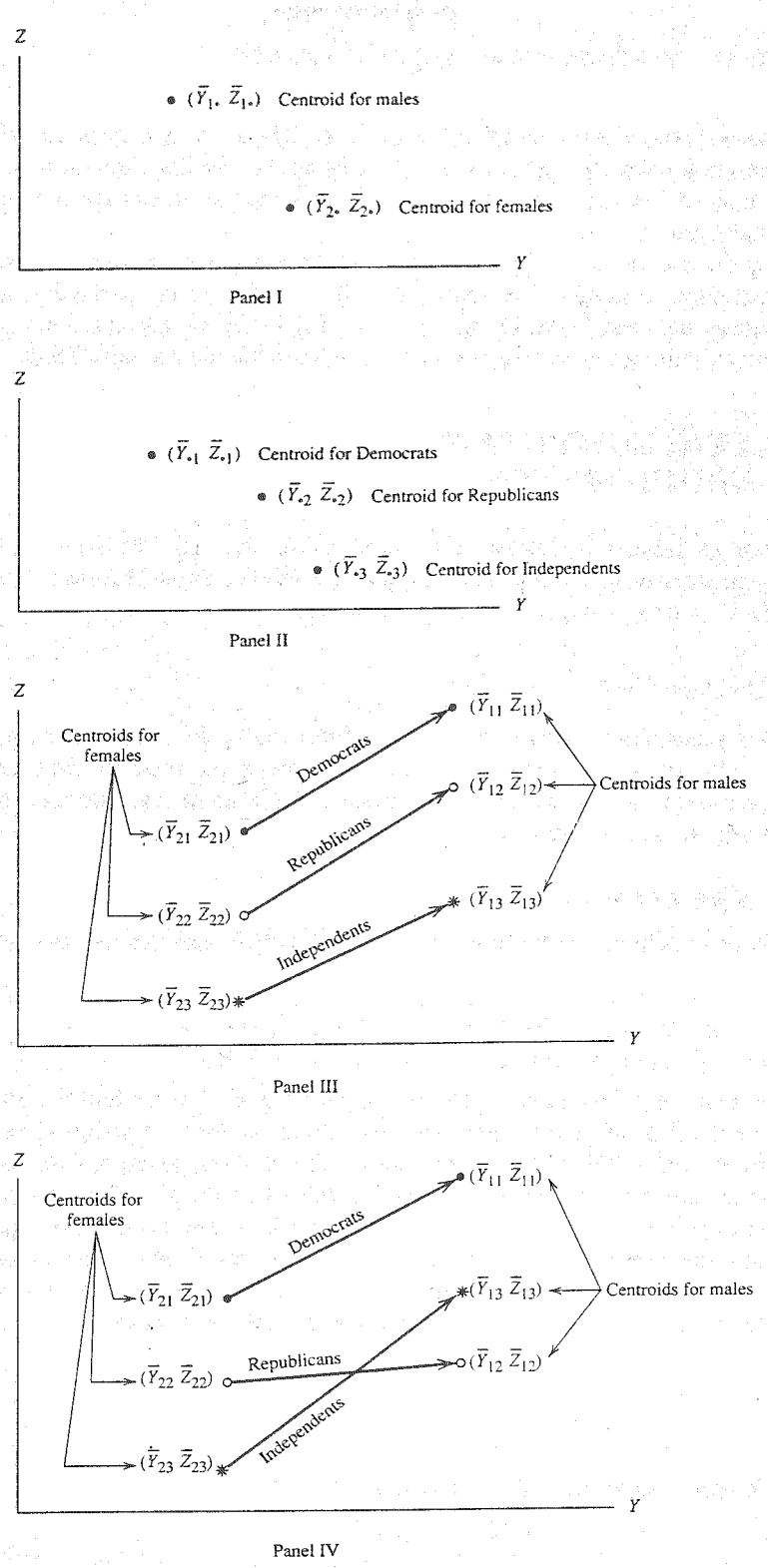


Figure 11.3 More than one independent variable and two dependent variables. Panel I, Main effect of gender. Panel II, Main effect of party affiliation. Panel III, No Gender × party affiliation interaction effect. Panel IV, Gender × party affiliation interaction effect.

hand, if the effect of *Gender* is not independent of *Party Affiliation* then, as depicted in Panel IV, the vectors joining the respective points will not be parallel. The magnitude of the interaction effect between the two variables is indicated by the extent to which the vectors are nonparallel.

The preceding discussion can easily be extended to more than two independent variables and p dependent variables. The centroids will be points in the p-dimensional space. The distances between centroids will give the main effects, and the nonparallelism of the vectors joining the appropriate centroids gives the interaction effects.

11.2 ANALYTIC COMPUTATIONS FOR TWO-GROUP MANOVA

The data set given in Table 8.1, Chapter 8, is employed to illustrate MANOVA. This data set is chosen because it is small and can also be used to show the similarities between MANOVA and two-group discriminant analysis.

11.2.1 Significance Tests

The first step is to determine if the two groups are significantly different with respect to the variables. That is, are the centroids of the two groups significantly different? This question is answered by conducting multivariate and univariate significance tests, which are discussed in the following pages.

Multivariate Significance Tests

The null and alternative hypotheses for multivariate statistical significance testing in MANOVA are:

$$H_o : \begin{pmatrix} \mu_{11} \\ \mu_{21} \end{pmatrix} = \begin{pmatrix} \mu_{12} \\ \mu_{22} \end{pmatrix} \qquad H_a : \begin{pmatrix} \mu_{11} \\ \mu_{21} \end{pmatrix} \neq \begin{pmatrix} \mu_{12} \\ \mu_{22} \end{pmatrix} \tag{11.1}$$

where μ_{ij} is the mean of the ith variable for the jth group. Note that the null hypothesis formally states that the difference between the centroids of the two groups is zero. Table 11.2 shows the various MANOVA computations for the data given in Table 8.1. The formulae used to compute the various statistics in Table 11.2 are given in Chapter 3. MD^2 between the centroids or means of the two groups is 15.155 and is directly proportional to the difference between the two groups. MD^2 can be transformed into various test statistics to determine if it is large enough to claim that the difference between the groups is statistically significant. In the case of two groups, MD^2 and Hotelling's T^2 are related as

$$T^2 = \left(\frac{n_1 \times n_2}{n_1 + n_2}\right) MD^2. \tag{11.2}$$

T^2 can be transformed into an exact F-ratio as follows:

$$F = \frac{(n_1 + n_2 - p - 1)}{(n_1 + n_2 - p)2} T^2, \tag{11.3}$$

which has an F distribution with p and $(n_1 + n_2 - p - 1)$ degrees of freedom. From Table 11.2, the values for T^2 and F-ratio are, respectively, equal to 90.930 and 43.398. The F-ratio is statistically significant at $p < .05$ and the null hypothesis is rejected, i.e., the means of the two groups are significantly different.

Table 11.2 **MANOVA Computations**

$$\bar{X}_1 = (.191 \quad .184) \qquad \bar{X}_2 = (.003 \quad .001) \qquad (\bar{X}_1 - \bar{X}_2) = (.188 \quad .183)$$

$$\mathbf{SSCP}_t = \begin{pmatrix} .265 & .250 \\ .250 & .261 \end{pmatrix} \qquad \mathbf{SSCP}_w = \begin{pmatrix} .053 & .045 \\ .045 & .062 \end{pmatrix} \qquad \mathbf{SSCP}_b = \begin{pmatrix} .212 & .205 \\ .205 & .199 \end{pmatrix}$$

$$\hat{\mathbf{S}}_w = \begin{pmatrix} .00243 & .00203 \\ .00203 & .00280 \end{pmatrix}$$

1. Multivariate Analysis
   (a) Statistical Significance Tests
   - $MD^2 = (.188 \quad .183)'\mathbf{S}_w^{-1}(.188 \quad .183)' = 15.155.$
   - $T^2 = \dfrac{(12 \times 12)15.155}{12 + 12} = 90.930$
   - $F = \dfrac{(12 + 12 - 2 - 1)90.930}{(12 + 12 - 2)2} = 43.398$
   - Eigenvalue of $\mathbf{SSCP}_b\mathbf{SSCP}_w^{-1} = 4.124$
   - Pillai's Trace = .805
   - Hotelling's Trace = 4.124
   - Wilks' Λ = .195
   - Roy's Largest Root = .805
   - F-ratio $= \dfrac{1 - .195}{.195} \times \dfrac{12 + 12 - 2 - 1}{2} = 43.346$

   (b) Effect Size
   - Partial eta square = .805

2. Univariate Analysis

	Statistic	*EBITASS*	*ROTC*
(a)	Statistical Analysis		
	MD^2	$\dfrac{(.191 - .003)^2}{.00243}$	$\dfrac{(.184 - .001)^2}{.00280}$
		$= 14.545$	11.960
	$T^2 = F$	$\dfrac{12 \times 12}{24} \times 14.545$	$\dfrac{12 \times 12}{24} \times 11.960$
		$= 87.270$	$= 71.760$
	t	9.342	8.471
(b)	Effect Size		
	Partial eta square	$\dfrac{87.270}{87.270 + 22}$	$\dfrac{71.760}{71.760 + 22}$
		$= .799$	$= .765$

From the above discussion it is clear that in the case of two groups the objective of MANOVA is to obtain the difference (i.e., distance) between groups using a suitable measure (e.g., MD^2) and assessing whether the difference is statistically significant. Besides MD^2, other measures of the difference between the groups are available. In Chapter 8 it was seen that for a univariate case SS_b/SS_w or $SS_b \times SS_w^{-1}$ is one of the

measures for the difference between two groups and it is related to the t-value, T^2, MD^2, and the F-ratio. For multiple dependent variables the multivariate analog for the differences between groups is a function of the eigenvalue(s) of the $\mathbf{SSCP}_b \times \mathbf{SSCP}_w^{-1}$ matrix.[2] Some of the measures formed using the eigenvalues are

$$\text{Pillai's Trace} = \sum_{i=1}^{K} \frac{\lambda_i}{1+\lambda_i}$$

$$\text{Hotelling's Trace} = \sum_{i=1}^{K} \lambda_i$$

$$\text{Wilks'}\Lambda = \prod_{i=1}^{K} \frac{1}{1+\lambda_i}$$

$$\text{Roy's Largest Root} = \frac{\lambda_{max}}{1+\lambda_{max}}, \tag{11.4}$$

where λ_i is the ith eigenvalue and K is the number of eigenvalues. Notice that all the measures differ with respect to how a single index is computed from the eigenvalues. Olson (1974) found that the test statistic based on Pillai's trace was the most robust and had adequate power to detect true differences under different conditions, and therefore we recommend its use to test multivariate significance. It can be shown that for two groups all of the above measures are equivalent and can be transformed into T^2 or an exact F-ratio.[3] For example, from Table 9.4, Chapter 9, the relationship between Wilks' Λ and F-ratio is

$$F = \left(\frac{1-\Lambda}{\Lambda}\right)\left(\frac{n_1+n_2-p-1}{p}\right). \tag{11.5}$$

Table 11.2 gives the values for all the above statistics and the F-ratio. The F-ratio for all the statistics is significant at $p < .05$. That is, the two groups are significantly different with respect to the dependent variables.

Univariate Significance Tests

Having determined that the means of the two groups are significantly different, the next obvious question is: Which variables are responsible for the differences between the two groups? One suggested procedure is to compare the means of each variable for the two groups. That is, conduct a series of t-tests for comparing the means of two groups. Table 11.2 also gives MD^2, T^2, t-value, and the F-ratio for each variable.[4] It can be seen that both *EBITASS* and *ROTC* significantly contribute to the differences between the two groups.

11.2.2 Effect Size

The statistical significance tests determine whether the differences in the means of the groups are statistically significant. Once again, for large sample sizes even small dif-

[2]The number of eigenvalues is equal to $min(G-1, p)$ where G and p, respectively, are the number of groups and number of dependent variables.

[3]As shown later in the chapter, the various measures are not the same for more than two groups and can only be approximately transformed into the F-ratio.

[4]The t-value is equal to $\sqrt{T^2}$ and for two groups the F-ratio is equal to T^2.

ferences are statistically significant. Consequently, one would like to measure the differences between the groups and then decide if they are large enough to be practically meaningful. That is, one would also like to assess the practical significance of the differences between the groups. Effect sizes can be used for such purposes. The effect size of any given independent variable or factor is the extent to which it affects the dependent variable(s). Univariate effect sizes are for the respective dependent variables, whereas multivariate effect sizes are for all the dependent variables combined. Discussion of univariate and multivariate effect sizes follows.

Univariate Effect Size

A number of related measures of effect size can be used. One common measure is MD^2 which is related to T^2 and the F-ratio. A second and more popular measure of effect size is the partial eta square, which is equal to SS_b/SS_t. The advantage of using partial eta square (PES) is that it ranges between zero and one, and it gives the proportion of the total variance that is accounted for by the differences between the two groups. In Chapter 8, it was seen that for the univariate case

$$\Lambda = \frac{SS_b}{SS_w}.$$

Or,

$$PES = 1 - \Lambda = 1 - \frac{SS_w}{SS_t} = \frac{SS_b}{SS_t}, \tag{11.6}$$

which is equal to PES. Since Λ can be transformed into an F-ratio, PES is also equal to

$$\frac{F \times df_b}{F \times df_b + df_w} \tag{11.7}$$

where df_b and df_w are, respectively, between-groups and within-group degrees of freedom. Using Eq. 11.7 and information from Table 11.2, the PESs for *EBITASS* and *ROTC*, respectively, are equal to .799 and .765. High values for the PESs suggest that a substantial proportion of the variance in the dependent variables is accounted for by the differences between the groups.

Multivariate Effect Size

The multivariate effect size is given by the difference between the centroids of the two groups. As discussed earlier, MD^2 measures the distance between the two groups and hence it can be used as a measure for the multivariate effect size. The larger the distance, the greater the effect size. The most popular measure of effect size, however, is once again PES and it gives the amount of variance in all the dependent variables that is accounted for by group differences. PES can be computed using Eq. 11.6 or Eq. 11.7. Using Eq. 11.6, the value of PES is equal to .805. This high value for PES suggests that a large proportion of the variance in *EBITASS* and *ROTC* is accounted for by the differences between the two groups. That is, the differences between the groups with respect to the dependent variables are meaningful.

11.2.3 Power

The power of a test is its ability to correctly reject the null hypothesis when it is false. That is, it is the probability of making a correct decision. The power of a test is directly

proportional to sample size and effect size, and inversely related to p-value. Power of the test can be obtained from power tables using effect size, p-value, and the sample size. The use of power tables is not illustrated as it requires a number of power tables that are not available in standard textbooks. Furthermore, the power of tests can be requested as a part of the MANOVA output in SPSS, or one can use SOLO Power Analysis (1992), which is a software package for computing the power of a number of statistical tests. The interested reader is referred to Cohen (1977) for further details on computation and use of power tables.

11.2.4 Similarities between MANOVA and Discriminant Analysis

One of the objectives of discriminant analysis is to identify a linear combination (called the discriminant function) of the variables that would give the maximum separation between the two groups. Next, a statistical test is performed to determine if the groups are significantly different with respect to the linear combination (discriminant scores). A significant difference between the groups with respect to the linear combination is equivalent to testing that the two groups are different with respect to the variables forming the linear combination.

In MANOVA, we test whether the centroids of the two groups are significantly different. Although a linear combination, which provides the maximum separation between the two groups, is not computed in MANOVA, the multivariate significance tests discussed earlier *implicitly* test whether the mean scores of the two groups obtained from such a linear combination are significantly different. Note that the null and the alternative hypotheses given in Eq. 11.1 are the same as those given in Section 8.3.2 of Chapter 8. Also the F-ratios for the univariate analysis reported in Table 11.2 are, within rounding errors, the same as those reported in Exhibit 8.1, Chapter 8 [4].

From the preceding discussion it is clear that in the case of one independent variable there is no difference between MANOVA and discriminant analysis. In the case of more than one independent variable, however, MANOVA provides additional insights into the effects of independent variables on dependent variables that are not provided by discriminant analysis. Further discussion of the additional insights provided by MANOVA are provided in Sections 11.4 and 11.5.

In the following sections we discuss the resulting output from the MANOVA procedure in SPSS. We will first discuss the output for the data given in Table 8.1, which gives the financial ratios for most-admired and least-admired firms. MANOVA that assesses the differences between two groups is sometimes referred to as two-group MANOVA. Next we discuss the output for multiple groups. MANOVA for assessing the differences between three or more groups is referred to as multiple-group MANOVA. Finally, we discuss the use of MANOVA to assess the effect of two or more independent variables.

11.3 TWO-GROUP MANOVA

Table 11.3 gives the SPSS commands for the data set given in Table 8.1. The PRINT subcommand specifies the options for the desired output. The CELLINFO option requests group means and standard deviations; the HOMOGENEITY option requests printing of Box's M statistic for testing the equality of covariance matrices; the ERROR option requests that the SSCP matrices be printed; the SIGNIF option requests printing of multivariate and univariate significance tests, the hypothesized sum of squares, effect sizes, and the eigenvalues of the $\mathrm{SSCP}_b \times \mathrm{SSCP}_w^{-1}$ matrix. The POWER subcommand requests that power for the F-test and the t-test be reported for the specified

Table 11.3 **SPSS Commands**

```
MANOVA EBITASS, ROTC BY EXCELL(1,2)
 /PRINT=CELLINFO(MEANS)
   HOMOGENEITY(BOXM)
   ERROR(SSCP)
   SIGNIF(MULTIV UNIV HYPOTH EFSIZE EIGEN)
 /POWER F(.05) T(.05)
 /DESIGN EXCELL
FINISH
```

p-values. The DESIGN subcommand specifies the effects (i.e., main and interaction effects) that the researcher is interested in testing. Since there is only one factor, only one main effect is specified. Exhibit 11.1 gives the output. Discussion of the output is brief because many of the reported statistics have been discussed in the previous section. The reader should compare the reported statistics in the output to the computed statistics in the previous section.

11.3.1 Cell Means and Homogeneity of Variances

The means and standard deviation for the dependent variables are reported for each cell or group [1]. One of the assumptions of MANOVA is that the covariance matrices for the two groups are the same. Note that this assumption is the same as that made in discriminant analysis. The test statistic used is Box's M, which can be approximately transformed to an F-ratio. The F-ratio is significant at $p < .05$, suggesting that the covariance matrices for the two groups are different [2]. Note that this test is also reported by the discriminant analysis procedure and the value reported here is the same as that reported in Exhibit 8.1, Chapter 8 [13]. As discussed in Chapter 12, the effect of violation of this assumption is not appreciable because the two groups are equal in size.

11.3.2 Multivariate Significance Tests and Power

In the present case there is only one main effect that pertains to the effect of EXCELL (firm's excellence) on firm's performance as measured by the two variables, *EBITASS* and *ROTC*. A significant main effect implies that the two groups of firms (i.e., most- and least-admired) are significantly different with respect to these variables. All the test statistics indicate a statistically significant effect for EXCELL at $p < .05$ [4a]. Note that values of Wilks' Λ, eigenvalue, and canonical correlation [4a, 4c] are the same as those reported in Exhibit 8.1, Chapter 8 [8]. From Eq. 11.7, the *PES* is equal to

$$PES = \frac{43.301 \times 2}{(43.301 \times 2) + 21} = .805,$$

and is the same as the effect size reported in the output [4b], and in Table 11.2. The power for the multivariate tests is large, suggesting that the probability of rejecting the null hypothesis when it is false is very high [4b]. The output also reports the value of the noncentrality parameter [4b]. The noncentrality parameter is closely related to the effect size and the corresponding test statistic. For any given test statistic, the value of the noncentrality parameter is zero if the null hypothesis is true. For example, the noncentrality parameter for T^2 is zero if the null hypothesis is true (i.e., the effect size is zero). As the effect size increases, the probability of rejecting the null also increases and so does the value of the noncentrality parameter. In a two-group case the value of

Exhibit 11.1

MANOVA for most-admired and least-admired firms

(1)

	CELL NUMBER	
Variable	1	2
EXCELL	1	2

Cell Means and Standard Deviations

Variable .. EBITASS

FACTOR	CODE	Mean	Std. Dev.	N
EXCELL	1	.191	.053	12
EXCELL	2	.003	.045	12
For entire sample		.097	.107	24

Variable .. ROTC

FACTOR	CODE	Mean	Std. Dev.	N
EXCELL	1	.183	.030	12
EXCELL	2	.001	.069	12
For entire sample		.092	.107	24

(2) Multivariate test for Homogeneity of Dispersion matrices
Boxs M = 21.50395 F WITH (3,87120) DF = 6.46365, P = .000 (Approx.)
Chi-Square with 3 DF = 19.38614, P = .000 (Approx.)

(3a) WITHIN CELLS Sum-of-Squares and Cross-Products

	EBITASS	ROTC
EBITASS	.053	
ROTC	.045	.062

EFFECT .. EXCELL
(3b) Adjusted Hypothesis Sum-of-Squares and Cross-Products

	EBITASS	ROTC
EBITASS	.212	
ROTC	.206	.199

(4a) Multivariate Tests of Significance (S = 1, M = 0, N = 9 1/2)

Test Name	Value	Exact F	Hypoth. DF	Error DF	Sig. of F
Pillais	.80484	43.30142	2.00	21.00	.000
Hotellings	4.12394	43.30142	2.00	21.00	.000
Wilks	.19516	43.30142	2.00	21.00	.000
Roys	.80484				

Note.. F statistics are exact.

(4b) Multivariate Effect Size and Observed Power at .0500 Level

TEST NAME	Effect Size	Noncent.	Power
(All)	.805	86.603	1.00

(4c) Eigenvalues and Canonical Correlations

Root No.	Eigenvalue	Pct.	Cum. Pct.	Canon Cor.
1	4.124	100.000	100.000	.897

(5) Univariate F-tests with (1,22) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig. of F
EBITASS	.21206	.05338	.21206	.00243	87.40757	.000
ROTC	.19929	.06169	.19929	.00280	71.06986	.000

Variable	ETA Square	Noncent.	Power
EBITASS	.79892	87.40757	1.00000
ROTC	.76362	71.06986	1.00000

the noncentrality parameter for all the multivariate test statistics is approximately equal to the T^2 statistic.[5]

11.3.3 Univariate Significance Tests and Power

The hypothesized and error sums of squares for univariate tests reported in the output are, respectively, the between-groups and within-group sums of squares [5]. The sums of squares are taken from the diagonals of the respective SSCP matrices [3a, 3b]. The univariate F-ratios for both variables are significant at $p < .05$, implying that both variables contribute to the differences between the groups [5]. Values for *PES* are the same as reported in Table 11.2 and are quite high, once again suggesting that a high proportion of the variance in the dependent variables is accounted for by the differences between the groups [5]. Note that the univariate F-ratios are exactly the same as those reported in Exhibit 8.1, Chapter 8 [4]. The power for the univariate tests pertaining to each independent variable is very high. In the univariate case the value of the noncentrality parameter is equal to the respective F-ratio.

11.3.4 Multivariate and Univariate Significance Tests

A multivariate test was first performed to determine if the centroids or the mean vectors of the two groups are significantly different, then a univariate test was done to determine which variables contribute to the difference between the two groups. One might wonder why a multivariate test was necessary when it was followed by a univariate test. There are two important reasons for first conducting a multivariate test. First, if all the univariate tests are independent, then the overall Type I error will be much higher than the chosen alpha. For example, if five independent univariate tests, each using an alpha level of .05, are performed then the probability that at least one of them is statistically significant due to chance alone will be equal to .226 [i.e., $1 - (1 - .05)^5$]. That is, the overall Type I error is not .05 but is .226. The actual Type I error will be larger if the tests are not independent. Second, as shown below, it is possible that the multivariate test is significant even though none of the univariate tests are significant.

Consider the data set given in Table 11.4 and plotted in Figure 11.4. Exhibit 11.2 gives the partial SPSS output. The univariate tests indicate that none of the means are

Table 11.4 **Hypothetical Data to Illustrate the Presence of Multivariate Significance in the Absence of Univariate Significance**

	Group 1		Group 2	
	X_1	X_2	X_1	X_2
	1	3	4	5
	2	5	5	5
	4	7	5	6
	6	11	8	7
	6	12	8	9
Mean	3.80	6.00	7.60	6.40

[5]The formula for computing the noncentrality parameter for a two-group case is equal to $[(n-3)/(n-2)]T^2$.

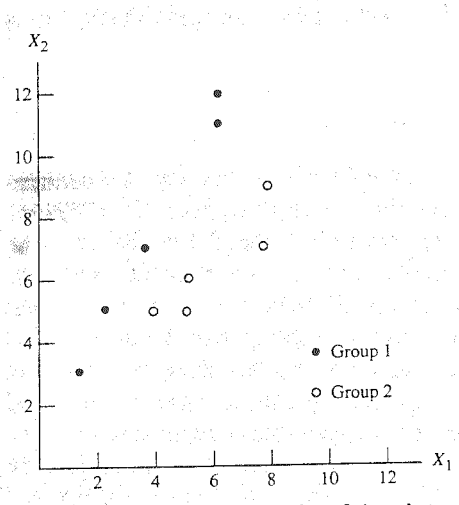


Figure 11.4 Presence of multivariate significance in the absence of univariate significance.

Exhibit 11.2

Multivariate significance, but no univariate significance

```
(1) WITHIN CELLS Sum-of-Squares and Cross-Products
                        Y1          Y2
Y1                  34.800
Y2                  45.600      70.400

(2) Multivariate Tests of Significance (S = 1, M = 0, N = 2 1/2)
Test Name           Value    Exact F Hypoth. DF   Error DF  Sig. of F
Pillais            .80993   14.91429       2.00       7.00       .003
Hotellings        4.26123   14.91429       2.00       7.00       .003
Wilks              .19007   14.91429       2.00       7.00       .003
Roys               .80993
Note.. F statistics are exact.

(3) Univariate F-tests with (1,8) D. F.

Variable    Hypoth. SS   Error SS Hypoth. MS   Error MS          F  Sig. of F
Y1           12.10000   34.80000   12.10000    4.35000    2.78161       .134
Y2            3.60000   70.40000    3.60000    8.80000     .40909       .540
```

significantly different for the two groups at $p < .05$ [3]. However, the multivariate test is significant at $p < .05$ [2]! Examination of Figure 11.4 gives a clue as to why this is the case. Note that there is not much separation between the two groups with respect to each variable; however, in the two-dimensional space the two groups are separated. Further insights regarding the presence of multivariate significance and absence of univariate significance can be gained by examining the pooled within-group SSCP_w matrix, which is equal to [1]

$$\text{SSCP}_w = \begin{pmatrix} 34.800 & 45.600 \\ 45.600 & 70.400 \end{pmatrix}.$$

The error term, MS_w, for the multivariate test in MANOVA is given by the determinant of the SSCP_w matrix. For the above matrix $MS_w = |\text{SSCP}_w| = 370.560$. Now, if the

variables were not correlated and the difference between the means of the two groups were the same, then $\mathbf{SSCP}_w$ would be equal to

$$\begin{pmatrix} 34.800 & 0 \\ 0 & 70.400 \end{pmatrix}$$

and $MS_w = 2449.92$, which is almost 6.6 times larger than if the variables were correlated. That is, the computation of the MS_w for multivariate tests takes into account the correlation among the variables. In other words, the multivariate tests take into account the correlation among the variables, whereas univariate tests ignore this information in the data.

11.4 MULTIPLE-GROUP MANOVA

Suppose a medical researcher hypothesizes that a treatment consisting of the simultaneous administration of two drugs is more effective than a treatment consisting of the administration of only one of the drugs. A study is designed in which 20 subjects are randomly divided into four groups of five subjects each. Subjects in the first group are given a placebo, subjects in the second group are given a combination of the two drugs, subjects in the third group are given only one of the two drugs, and subjects in the fourth group are given the other drug. The effectiveness of the drugs (i.e., treatment effectiveness) is measured by two response variables, Y_1 and Y_2. Table 11.5 gives the data and the group means. Note that this study manipulates one factor, labeled *DRUG*, and it has four levels. A one-factor study with more than two levels is often referred to as multiple-group MANOVA. Table 11.6 gives the SPSS commands and Exhibit 11.3 gives the partial output.

Table 11.5 **Data for Drug Effectiveness Study**

	Treatments							
	1		2		3		4	
	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
	1	2	8	9	2	4	4	5
	2	1	9	8	3	2	3	3
	3	2	7	9	3	3	3	4
	2	3	8	9	3	5	5	6
	2	2	8	10	4	6	5	7
Means	2	2	8	9	3	4	4	5

Table 11.6 **SPSS Commands for Drug Study**

```
MANOVA Y1 Y2 BY DRUG(1,4)
 /PRINT=CELLINFO(MEANS)
        HOMOGENEITY(BOXM)
        ERROR(SSCP,COV)
        SIGNIF(MULTIV,UNIV)
 /DESIGN=DRUG
```

Exhibit 11.3

MANOVA for drug study

```
(1) Multivariate test for Homogeneity of Dispersion matrices
    Boxs M =                          13.09980
    F WITH (9,2933) DF =               1.12256, P =    .343 (Approx.)
    Chi-Square with 9 DF =            10.14325, P =    .339 (Approx.)

    EFFECT .. DRUG
(2a) Multivariate Tests of Significance (S = 2, M = 0, N = 6 1/2)

    Test Name          Value  Approx. F Hypoth. DF   Error DF  Sig. of F

    Pillais          1.02715    5.63102       6.00      32.00       .000
    Hotellings      11.41361   26.63176       6.00      28.00       .000
    Wilks            .07253   13.56607       6.00      30.00       .000
    Roys             .91865
    Note.. F statistic for WILKS' Lambda is exact.

(2b) Univariate F-tests with (3,16) D. F.

    Variable   Hypoth. SS   Error SS Hypoth. MS   Error MS          F  Sig. of F
    Y1         103.75000   10.00000   34.58333     .62500   55.33333       .000
    Y2         130.00000   24.00000   43.33333    1.50000   28.88889       .000
```

11.4.1 Multivariate and Univariate Effects

As Box's M statistic is not significant at $p < .05$, one fails to reject the null hypothesis suggesting that the covariance matrices of the groups are not different [1]. The multivariate effect of *DRUG* is significant at $p < .05$ [2a]. That is, the mean vectors of the four groups are significantly different. The univariate tests indicate that the four groups significantly differ with respect to both the dependent variables [2b]. Based on these results, the researcher can conclude that treatment groups are different with respect to their effectiveness. But, which pairs of groups or combinations of groups are different? For example, if the researcher wants to determine whether the effectiveness of the two drugs is different or the same then he/she would test whether groups 3 and 4 are different with respect to their effectiveness. Or, if the researcher is interested in determining whether the simultaneous administration of the drugs is different from the administration of only one drug then the approach would be to test for differences between the effectiveness of group 2 and the average effectiveness of groups 3 and 4. Testing for differences between specific groups or combinations of groups is referred to as comparison or contrast testing. Notice that testing which groups are different with respect to a given set of variables was not an explicit objective of discriminant analysis. In this respect, therefore, the objectives of discriminant analysis and MANOVA are different.

11.4.2 Orthogonal Contrasts

Statistical significance testing of comparisons can be assessed by first forming contrasts and then testing for their significance. A contrast is a linear combination of the group means of a given factor. It should be noted that it is a good statistical practice to perform contrast analysis determined or stated *a priori*, rather than test all possible contrasts in search of significant effects. For instance, let us assume that the researcher is interested in answering the following questions pertaining to the study:

1. Is the effectiveness of the placebo (i.e., the first treatment or control group) different from the average effectiveness of the drugs given to the other three groups? A statistically significant difference would suggest that the drugs are effective either individually or when administered simultaneously.
2. Is the effectiveness of the two drugs administered to the second treatment group significantly different from the average effectiveness of the drugs administered to treatment groups 3 and 4? A significant difference would suggest that the effectiveness of simultaneously administering both drugs is different from the effectiveness of administering only one drug at a time.
3. Is the effectiveness of the drug given to the third treatment group significantly different from the effectiveness of the drug given to the fourth treatment group? A significant difference would suggest that the two drugs differ in their effectiveness.

Each of these questions or hypotheses can be answered by forming a contrast and testing for its significance. In univariate significance tests, each contrast is tested separately for each dependent variable, whereas in multivariate significance tests each contrast is tested simultaneously for all the dependent variables. In order to simplify the discussion, we first discuss univariate significance tests, then multivariate significance tests. However, note that univariate contrasts should only be interpreted if the corresponding multivariate contrast is significant.

Univariate Significance Tests for the Contrasts

Consider the following linear combination:

$$C_{ij} = c_{i1}\mu_{1j} + c_{i2}\mu_{2j} + \cdots + c_{ik}\mu_{kj}$$

where C_{ij} is the ith contrast for the jth variable, c_{ik} are the coefficients of the contrast, and μ_{kj} is the mean of the kth group for the jth variable. Contrasts are said to be orthogonal if

$$\sum_{k=1}^{G} c_{ik} = 0 \qquad \text{for all } i \tag{11.8}$$

and

$$\sum_{k=1}^{G} \frac{c_{ik}c_{lk}}{n_k} = 0 \qquad \text{for all } i \neq l \tag{11.9}$$

where i and l are any two contrasts. For equal sample size the preceding equation reduces to

$$\sum_{k=1}^{G} c_{ik}c_{lk} = 0 \qquad \text{for all } i \neq l. \tag{11.10}$$

From Eqs. 11.8 to 11.10, two contrasts are orthogonal if the sum of the coefficients of each contrast is equal to zero, and the sum of the product of the corresponding coefficients of the two contrasts is also equal to zero. If these conditions do not hold, then the contrasts are correlated. In general, orthogonal contrasts are desired; however, the researcher is not constrained to only orthogonal contrasts. Correlated contrasts are discussed later. The total number of contrasts for any given factor or effect is equal to its degrees of freedom. However, there can be infinite sets of contrasts with each set consisting of the maximum number of allowable contrasts. For the present study, there

will be a maximum of three contrasts in each set; however, there can be infinite sets of contrasts, with each set consisting of three contrasts. Following is an example of the set of contrasts (for the dependent variable Y_1) that would address the research questions posed earlier:

$$C_{11} = \mu_{11} - \frac{1}{3}\mu_{21} - \frac{1}{3}\mu_{31} - \frac{1}{3}\mu_{41} = \mu_{11} - \frac{\mu_{21} + \mu_{31} + \mu_{41}}{3} \tag{11.11}$$

$$C_{21} = \mu_{21} - \frac{1}{2}\mu_{31} - \frac{1}{2}\mu_{41} = \mu_{21} - \frac{\mu_{31} + \mu_{41}}{2} \tag{11.12}$$

$$C_{31} = \mu_{31} - \mu_{41}. \tag{11.13}$$

The coefficients of this set of contrasts are presented as Set 1 in Table 11.7. It can be readily checked that the three contrasts in Set 1 are orthogonal as the sum of the coefficients of each contrast is equal to zero and the sum of the product of the corresponding coefficients of each pair of contrasts is also equal to zero. The table also presents three other sets of orthogonal contrasts. The specific set of contrasts that the researcher would test will depend on the questions that need to be addressed by the study.

The null and the alternative hypotheses for testing the significance of univariate contrasts are

$$H_o : C_{ij} = 0 \qquad H_a : C_{ij} \neq 0.$$

For example, the null and alternative hypotheses for C_{11} (i.e., the first contrast for Y_1) given by Eq. 11.11 are:

$$H_o : C_{11} = 0 \qquad H_a : C_{11} \neq 0,$$

which can be rewritten as

$$H_o : \mu_{11} = \frac{\mu_{21} + \mu_{31} + \mu_{41}}{3} \qquad H_a : \mu_{11} \neq \frac{\mu_{21} + \mu_{31} + \mu_{41}}{3}.$$

From these hypotheses, it is clear that the contrast essentially tests whether or not two means are significantly different, where each mean could be a weighted average of two or more means.

Table 11.7 **Coefficients for the Contrasts**

	Groups			
Contrast	**1**	**2**	**3**	**4**
Set 1	1	−1/3	−1/3	−1/3
	0	1	−1/2	−1/2
	0	0	1	−1
Set 2	−1/3	−1/3	−1/3	1
	−1/2	−1/2	1	0
	−1	−1	0	0
Set 3	1	−1	0	0
	0	0	1	−1
	1/2	1/2	−1/2	−1/2
Set 4	1	0	0	−1
	0	1	−1	0
	1/2	−1/2	−1/2	1/2

It can be shown that the standard error of contrast C_{ij} is equal to

$$\sqrt{MSE_j \sum_{k=1}^{G} \frac{c_{ik}^2}{n_k}} \tag{11.14}$$

where MSE_j is the mean square error for the jth variable and its estimate is given by its pooled within-group variance. The resulting t-value is

$$t = \frac{C_{ij}}{\sqrt{MSE_j \sum_{k=1}^{G} c_{ik}^2/n_k}}, \tag{11.15}$$

or

$$T^2 = t^2 = \frac{C_{ij}^2}{MSE_j \sum_{k=1}^{G} c_{ik}^2/n_k},$$

which can be rewritten as

$$T^2 = t^2 = \left(\sum_{k=1}^{G} \frac{c_{ik}^2}{n_k}\right)^{-1} C_{ij} MSE_j^{-1} C_{ij}. \tag{11.16}$$

In the univariate case, T^2 is equal to the F-ratio, which has an F distribution with 1 and $n - G$ degrees of freedom. Equation 11.6 can also be written as

$$F = \frac{C_{ij}^2 / \left(\sum_{k=1}^{G} c_{ik}^2/n_k\right)}{MSE_j}. \tag{11.17}$$

Note that in Eq. 11.16 the term $C_{ij} MSE_j^{-1} C_{ij}$ is equal to MD^2 and consequently T^2, and the F-ratio is proportional to MD^2 or the statistical distance between the two means. That is, once again the problem reduces to obtaining the distance between two means and determining if this distance is statistically significant. In Eq. 11.17, the numerator is the hypothesis mean square and, since the hypothesis to be tested has one degree of freedom, it is also the hypothesis sum of squares, and the denominator is the error mean square.

Multivariate Significance Test for the Contrasts

Multivariate contrasts are used to simultaneously test for the effects of all the dependent variables. A multivariate contrast is given by

$$\mathbf{C}_i = c_{i1}\boldsymbol{\mu}_1 + c_{i2}\boldsymbol{\mu}_2 + \cdots + c_{ij}\boldsymbol{\mu}_k$$

where $\boldsymbol{\mu}_k$ is a vector of means for the kth group and $\mathbf{C}_i$ is the ith contrast vector.

The null and alternative hypotheses for the multivariate significance test for the ith contrast are

$$H_o : \mathbf{C}_i = \mathbf{0} \qquad H_a : \mathbf{C}_i \neq \mathbf{0}.$$

The test statistic, T^2, is the multivariate analog of Eq. 11.6 and is given by

$$T^2 = \left(\sum_{k=1}^{G} \frac{c_{ik}^2}{n_k}\right)^{-1} \mathbf{C}_i' \hat{\boldsymbol{\Sigma}}_w^{-1} \mathbf{C}_i, \tag{11.18}$$

where $\hat{\Sigma}_w^{-1}$ is the inverse of the pooled within-group covariance matrix. T^2 can be transformed into an F-ratio using

$$F = \left(\frac{df_e - p + 1}{df_e \times p}\right)T^2 \tag{11.19}$$

and has an F distribution with p and $df_e + p - 1$ degrees of freedom, where df_e is the error degrees of freedom.

Estimation and Significance Testing of Contrasts Using SPSS

The MANOVA program can be used to estimate and conduct significance tests for a number of different types of contrasts. Table 11.8 gives the SPSS commands for forming and testing the set of contrasts given by Eqs. 11.11 to 11.13. The CONTRAST subcommand specifies that contrasts for the *DRUG* factor are to be formed and tested. Following the equals sign, the desired set of contrasts is specified. The set of contrasts given by Eqs. 11.11 to 11.13 are called Helmert contrasts. MANOVA automatically generates the coefficients for Helmert and a number of other commonly used contrasts (refer to the SPSS manual for other commonly used contrasts). The keyword HELMERT requests the forming and testing of Helmert contrasts. In Helmert contrasts, the first contrast tests for the statistical significance of the mean of the first group with the average of the means of the remaining groups, the second contrast tests for the significance of the mean of the second group with the average of the means of the remaining groups, and so on. SPSS gives the user the option of specifying contrast coefficients using the SPECIAL keyword, illustrated later in this chapter. The PARAMETER option requests the printing of the contrast estimates and their significance tests. In the DESIGN subcommand, DRUG(i) refers to the ith contrast. For example, DRUG(1) is the first contrast (i.e., C_1), DRUG(2) is the second contrast (i.e., C_2), and so on. Exhibit 11.4 gives the partial output.

UNIVARIATE SIGNIFICANCE TESTS. Consider the contrast given by Eq. 11.11, which is the first contrast and is represented in the output as Drug(1). Substituting the means reported in Table 11.5, we get

$$C_{11} = 2 - \frac{8 + 3 + 4}{3} = -3.000$$

and

$$C_{12} = 2 - \frac{9 + 4 + 5}{3} = -4.000,$$

which are the same as those reported in Exhibit 11.4 [4a, 5a]. From the output the MS_e for Y_1 and Y_2, respectively, are equal to 0.625 and 1.500 [3d]. Using Eq. 11.14, the

Table 11.8 **SPSS Commands for Helmert Contrasts**

```
MANOVA Y1 Y2 BY DRUG(1,4)
 /CONTRAST(DRUG)=HELMERT
 /PRINT=SIGNIF(MULTIV,UNIV)
        PARAMETERS(ESTIM)
 /DESIGN=DRUG(1) DRUG(2) DRUG(3)
```

Exhibit 11.4

Helmert contrasts for drug study

1a EFFECT .. DRUG(3)
Multivariate Tests of Significance (S = 1, M = 0, N = 6 1/2)

Test Name	Value	Exact F	Hypoth. DF	Error DF	Sig. of F
Pillais	.20747	1.96335	2.00	15.00	.175
Hotellings	.26178	1.96335	2.00	15.00	.175
Wilks	.79253	1.96335	2.00	15.00	.175
Roys	.20747				

Note.. F statistics are exact.

1b Univariate F-tests with (1,16) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig. of F
Y1	2.50000	10.00000	2.50000	.62500	4.00000	.063
Y2	2.50000	24.00000	2.50000	1.50000	1.66667	.215

2a EFFECT .. DRUG(2)
Multivariate Tests of Significance (S = 1, M = 0, N = 6 1/2)

Test Name	Value	Exact F	Hypoth. DF	Error DF	Sig. of F
Pillais	.87605	53.01047	2.00	15.00	.000
Hotellings	7.06806	53.01047	2.00	15.00	.000
Wilks	.12395	53.01047	2.00	15.00	.000
Roys	.87605				

Note.. F statistics are exact.

2b Univariate F-tests with (1,16) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig. of F
Y1	67.50000	10.00000	67.50000	.62500	108.00000	.000
Y2	67.50000	24.00000	67.50000	1.50000	45.00000	.000

3a EFFECT .. DRUG(1)
Multivariate Tests of Significance (S = 1, M = 0, N = 6 1/2)

Test Name	Value	Exact F	Hypoth. DF	Error DF	Sig. of F
Pillais	.80330	30.62827	2.00	15.00	.000
Hotellings	4.08377	30.62827	2.00	15.00	.000
Wilks	.19670	30.62827	2.00	15.00	.000
Roys	.80330				

Note.. F statistics are exact.

3b Univariate F-tests with (1,16) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS (3c)	Error MS (3d)	F (3e)	Sig. of F
Y1	33.75000	10.00000	33.75000	.62500	54.00000	.000
Y2	60.00000	24.00000	60.00000	1.50000	40.00000	.000

(continued)

Exhibit 11.4 (continued)

```
(4) Estimates for Y1
    --- Individual univariate .9500 confidence intervals

                        (4a)       (4b)       (4c)
  Parameter            Coeff.   Std. Err.   t-Value      Sig. t  Lower -95%  CL-Upper

  DRUG(1)
        2          -3.0000000     .40825   -7.34847      .00000   -3.86545   -2.13455
  DRUG(2)
        3           4.50000000    .43301   10.39230      .00000    3.58205    5.41795
  DRUG(3)
        4          -1.0000000     .50000   -2.00000      .06277   -2.05995     .05995

(5) Estimates for Y2
    --- Individual univariate .9500 confidence intervals

                        (5a)       (5b)       (5c)
  Parameter            Coeff.   Std. Err.   t-Value      Sig. t  Lower -95%  CL-Upper

  DRUG(1)
        2          -4.0000000     .63246   -6.32456      .00001   -5.34075   -2.65925
  DRUG(2)
        3           4.50000000    .67082    6.70820      .00001    3.07792    5.92208
  DRUG(3)
        4          -1.0000000     .77460   -1.29099      .21505   -2.64207     .64207
```

standard error for C_{11} is

$$\sqrt{0.625\left(\frac{1}{5}+\frac{1}{9}\times\frac{1}{5}+\frac{1}{9}\times\frac{1}{5}+\frac{1}{9}\times\frac{1}{5}\right)} = .408,$$

and the standard error for C_{12} is

$$\sqrt{1.500\left(\frac{1}{5}+\frac{1}{9}\times\frac{1}{5}+\frac{1}{9}\times\frac{1}{5}+\frac{1}{9}\times\frac{1}{5}\right)} = .632,$$

which is also the same as reported in the output [4b, 5b]. The t-values reported in the output for C_{11} and C_{12}, respectively, are -7.352 $(-3/.408)$ and -6.329 $(-4/.632)$ [4c, 5c]. Contrasts C_{11} and C_{12} are statistically significant at $p < .05$, from which the researcher can conclude that the drugs are effective when administered either individually or when administered simultaneously.

Univariate significance tests for the contrasts are also reported in another section of the output. The reported test statistics are computed using Eq. 11.17. Once again, consider contrast C_{11}. The numerator of Eq. 11.17 is equal to

$$\frac{-3^2}{\left(\frac{1}{5}+\frac{1}{9}\times\frac{1}{5}+\frac{1}{9}\times\frac{1}{5}+\frac{1}{9}\times\frac{1}{5}\right)} = 33.750$$

and is equal to the reported hypothesis mean square [3c], and the denominator is equal to 0.625, which is the error mean square [3d]. The F-ratio for the contrast is therefore equal to 54.000 (i.e., 33.750/.625) [3e]. The only difference between the significance tests reported in [4] and [3b] is that in the former part of the output, the contrast estimates are also reported.

The univariate contrasts for DRUG(3) (cf. Eq. 11.13) are not significant at $p = .05$, which leads to the conclusion that the effectiveness of the two drugs is the same [1b, 4, 5]. The significant univariate contrasts for DRUG(2) (cf. Eq. 11.12) suggest that simultaneously administering both drugs is more effective than administering only one drug at a time [2b, 4, 5]. The final conclusion is that the two drugs are equally effective, but the effectiveness of a treatment consisting of administering both drugs simultaneously is much greater than administering each drug separately.

MULTIVARIATE SIGNIFICANCE TESTS. The multivariate estimate for the contrast, DRUG(1), is given by

$$\begin{aligned} \mathbf{C}_1' &= (2\ 2) - \frac{1}{3}(8\ 9) - \frac{1}{3}(3\ 4) - \frac{1}{3}(4\ 5) \\ &= (-3\ \ -4), \end{aligned}$$

which can be easily obtained from the corresponding univariate estimates for the contrasts [4a, 5a]. The T^2 for the above contrast is (see Eq. 11.18)

$$\begin{aligned} T^2 &= \left(\frac{1}{5} + \frac{1}{9} \times \frac{1}{5} + \frac{1}{9} \times \frac{1}{5} + \frac{1}{9} \times \frac{1}{5}\right)^{-1} (-3\ \ -4)\hat{\boldsymbol{\Sigma}}_w^{-1}(-3\ \ -4)' \\ &= 3.750(-3\ \ -4)\begin{pmatrix} .625 & .438 \\ .438 & 1.500 \end{pmatrix}^{-1} (-3\ \ -4)' = 65.318, \end{aligned}$$

and the corresponding F-ratio reported in the output is (see Eq. 11.19):

$$\left(\frac{16 - 2 + 1}{16 \times 2}\right) 65.318 = 30.618,$$

with 2 and 15 degrees of freedom, and is significant at $p < .05$ [3a]. All the multivariate significance tests indicate that at an alpha level of .05, contrasts $\mathbf{C}_1$ and $\mathbf{C}_2$ are significant [3a, 2a] and contrast $\mathbf{C}_3$ is not significant [1a]. The overall conclusion reached is the same as discussed in the previous section. Once again, note that univariate analysis of contrasts should only be done if the corresponding multivariate significance tests are significant.

Correlated Contrasts

Suppose the researcher is interested in comparing the mean effectiveness of each treatment group with that of the placebo or control group (i.e., group 1). Table 11.9 gives the coefficients for the set of contrasts that would achieve that objective. Note that the contrasts are not orthogonal because the sum of the product of the corresponding coefficients for any pair of contrasts is not equal to zero. Table 11.10 gives the SPSS commands for testing the significance of the contrasts.

Table 11.9 **Coefficients for Correlated Contrasts**

	Coefficients			
Contrast	**1**	**2**	**3**	**4**
$\mathbf{C}_1$: DRUG(1)	−1	1	0	0
$\mathbf{C}_2$: DRUG(2)	−1	0	1	0
$\mathbf{C}_3$: DRUG(3)	−1	0	0	1

Table 11.10 **SPSS Commands for Correlated Contrasts**

```
MANOVA Y1 Y2 BY DRUG(1,4)
 /CONTRAST=SPECIAL(1 1 1 1
                  -1 1 0 0
                  -1 0 1 0
                  -1 0 0 1)
 /PRINT=SIGNIF(MULTIV,UNIV)
 /METHOD=SSTYPE(UNIQUE) DESIGN(SOLUTION)
 /DESIGN=DRUG(3) DRUG(2) DRUG(1)
```

The CONTRAST command specifies the coefficients for the contrasts, which are followed by the SPECIAL keyword. The first line gives the coefficients for the constant, which must be all ones. (The constant is common for all contrasts and is not interpreted.) Lines 2 through 4 give the coefficients for the three contrasts given in Table 11.9. That is, the second line represents contrast DRUG(1), the third line represents contrast DRUG(2), and the fourth line represents contrast DRUG(3). The METHOD command specifies the procedure or method to be used for computing or extracting the sum of squares—unique and sequential being the two available methods. For orthogonal contrasts, both procedures give identical results. However, for correlated contrasts the results and the subsequent conclusions depend on the method used for extracting the sums of squares. In order to illustrate the interpretational complexity of correlated contrasts, we will compare the results obtained from unique and sequential methods. Table 11.11 summarizes the significance results from the resulting output, which is not reproduced. From the multivariate significance tests for the unique method, contrasts DRUG(1) (i.e., C_1) and DRUG(3) (i.e., C_3) are significant at $p = .05$ and contrast DRUG(2) (i.e., C_2) is not significant at $p = .05$. That is, the mean effectiveness of groups 2 and 4 is significantly different from group 1.

Now suppose that the sequential method, instead of the unique method, is specified for extracting the sum of squares. The command for requesting the sequential method is: METHOD = SSTYPE(SEQUENTIAL). Exhibit 11.5 gives the partial output and Table 11.11 also summarizes the significance test results. SPSS prints a warning message indicating that use of the sequential method for nonorthogonal designs may

Table 11.11 **Summary of Significant Tests**

Contrast	Effects Controlled or Partialled	Multivariate Significance	Univariate Significance Y_1	Univariate Significance Y_2
Unique Method				
DRUG(1)	DRUG(2), DRUG(3)	.000	.000	.000
DRUG(2)	DRUG(3), DRUG(1)	.055	.063	.020
DRUG(3)	DRUG(1), DRUG(2)	.000	.001	.001
Sequential Method				
DRUG(1)	DRUG(2), DRUG(3)	.000	.000	.000
DRUG(2)	DRUG(3)	.002	.000	.040
DRUG(3)		.682	.426	1.000

Exhibit 11.5

SPSS output for correlated contrasts using the sequential method

```
(1) 0>Warning # 12189
    >You are using SEQUENTIAL Sums of Squares with a potentially
    >nonorthogonal partition of an effect.  Either the design is
    >unbalanced, you have specified a nonorthogonal contrast for
    >the partitioned factor (default DEVIATION, SIMPLE, or
    >REPEATED), or you have specified a SPECIAL contrast for the
    >partitioned factor.  If you must interpret SEQUENTIAL F-tests
    >for nonorthogonally partitioned effects, see the solution
    >matrix to determine the actual hypotheses tested.  Default
    >UNIQUE Sums of Squares are directly interpretable for
    >nonorthogonal partitions.

    Solution Matrix for Between-Subjects Design
(2)  1-DRUG
    FACTOR                                          PARAMETER

                 Constant  Drug(3)   Drug(2)   Drug(1)
      1               1        2         3         4

      1            1.118    -.645     -.913    -1.581
      2            1.118    -.645     -.913     1.581
      3            1.118    -.645     1.826      .000
      4            1.118    1.936      .000      .000
```

not be appropriate [1]. The results are very different from those obtained for the unique method. From the multivariate tests, it is now concluded that contrasts DRUG(1) and DRUG(2) are significant at $p = .05$ and contrast DRUG(3) is not significant at $p = .05$. That is, the treatment effectiveness of group 4 is not different from group 1, whereas the unique method concluded that it was different. Also, it is now concluded that DRUG(2) is significant at $p = .05$; that is, group 3 is significantly different from group 1, whereas the unique method concluded that it was not different. What is the reason for these drastic differences in the results? The obvious answer is that there is correlation among the contrasts. As we now discuss, the two methods differ with respect to how the sums of squares are extracted.

In the unique method the sums of squares are computed *after* the effects of all other contrasts are removed or partialled out, irrespective of the order of the contrasts specified in the DESIGN subcommand. For example, contrast DRUG(1) is tested after the effects of the other contrasts, DRUG(2) and DRUG(3), are removed. In the sequential method, however, the partialling of the effect of other contrasts depends on the order in which the contrasts are specified in the DESIGN subcommand. The sum of squares for each contrast is extracted after the effect of the contrasts specified to its left have been partialled out. For example, for the DESIGN = DRUG(3) DRUG(2) DRUG(1) statement the sum of squares for DRUG(1) is computed after the effects of DRUG(2) and DRUG(3) have been partialled out, and therefore the respective sums of squares reflect the effect of DRUG(1) after the effects of all other contrasts have been taken into consideration. On the other hand, the effect of DRUG(3) is computed without partialling out the effects of other contrasts, and therefore the computed sum of squares includes not only the effect of DRUG(1), but also the effects of other contrasts that are correlated with this contrast.

The preceding analysis implies that the hypotheses tested may not correspond to the hypotheses specified by the contrast statement, as the computed sum of squares also

includes the effect of other contrasts. SPSS gives the option of printing a solution matrix that contains information about the actual hypotheses tested. The solution matrix can be obtained by specifying DESIGN(SOLUTION) option in the PRINT subcommand. Exhibit 11.5 [2] gives the solution matrix. The columns of the solution matrix represent the contrasts and correspond to the contrasts specified in the DESIGN subcommand. From the solution matrix it is clear that the contrasts tested are not the same as those intended by the researcher. For example, DRUG(3) actually tests the difference between group 4 and the average of groups 1, 2, and 3, and not the difference between groups 4 and 1 as intended by the researcher. Similarly, DRUG(2) tests the difference between group 3 and the average of groups 1 and 2 and not groups 3 and 1.

The question then becomes: Which of the two methods for extracting the sum of squares should be used? The sequential method has the desirable property that the sum of squares of each effect and the error sum of squares add up to the total sum of squares. However, the sequential method has the undesirable property that the actual contrasts tested may not be the same as those specified by the researcher. The unique method partials out the effect of all other contrasts and therefore the contrasts tested are the same as those intended. But, it has the undesirable property that the sum of squares of each effect and the error sum of squares do not add up to the total sum of squares. However, the emphasis should be on testing the correct contrasts and not whether the sum of squares of all the different contrasts add up to the total sum of squares. Therefore, the unique method should be preferred over the sequential method. Fortunately, the unique method is the default method for extracting the sum of squares in SPSS.

11.5 MANOVA FOR TWO INDEPENDENT VARIABLES OR FACTORS

Suppose the advertising department has prepared three ads for introducing a new product and it is interested in identifying the best ad. The first ad uses a humorous appeal, the second uses an emotional appeal, and the third uses a comparative approach. It is further believed that respondent's gender could have an effect on his/her preference for the type of ad. An experiment is conducted in which 12 males and 12 females are

Table 11.12 **Data for the Ad Study**[a]

		Type of Ad			
Gender		Humorous	Emotional	Comparative	Means
Male	Y_1	8 8 10 10 (9.000)	5 5 7 7 (6.000)	2 2 4 4 (3.000)	6.000
	Y_2	6 7 9 10 (8.000)	3 4 6 7 (5.000)	1 2 3 2 (2.000)	5.000
Female	Y_1	2 2 4 4 (3.000)	4 4 2 2 (3.000)	10 10 8 8 (9.000)	5.000
	Y_2	1 2 2 3 (2.000)	2 6 3 1 (3.000)	10 9 6 7 (8.000)	4.333
Means	Y_1	6.000	4.500	6.000	5.000
	Y_2	5.000	4.000	5.000	4.667

[a] Numbers in parentheses are cell means.

Table 11.13 **SPSS Commands for the Ad Study**

```
MANOVA Y1 TO Y2 BY GENDER(1,2) AD(1,3)
 /PRINT=CELLINFO(MEANS SSCP)
 HOMOGENEITY(BOXM)
 ERROR(SSCP)
 SIGNIF(MULTIV UNIV HYPOTH EIGEN EFSIZE DIMENR)
 /DISCRIM=RAW
 /POWER F(.05) T(.05)
 /DESIGN GENDER AD GENDER BY AD
```

exposed to one of the ads. The 12 males are divided randomly into three groups of four subjects each. Each group is exposed to a different ad and the respondents are asked to evaluate the ad with respect to how informative (Y_1) and believable (Y_2) the ad is on an 11-point scale with 1 indicating low informativeness and low believability, and 11 indicating high informativeness and high believability. The procedure is repeated for the female sample. Table 11.12 gives the data and Table 11.13 gives the **SPSS** commands. The DISCRIM subcommand requests discriminant analysis for each effect and the printing of the raw (i.e., unstandardized) coefficients. The DESIGN statement gives the effects that the researcher is interested in. Since there are two factors, there are two main effects (i.e., *GENDER* and *AD*) and one two-way interaction (i.e., *GENDER BY AD*). Multivariate significance for each of these effects is tested using the previously described multivariate tests. Exhibit 11.6 gives the partial output.

11.5.1 Significance Tests for the $GENDER \times AD$ Interaction

The first effect tested is the $GENDER \times AD$ interaction. MANOVA labels the between-groups **SSCP** matrix for a given effect as the hypothesis **SSCP** matrix. The eigenvalues of the $\mathbf{SSCP}_h \times \mathbf{SSCP}_w^{-1}$ matrix are used for computing the test statistics given by Eq. 11.4 [1c]. For example,

$$\Lambda = \left(\frac{1}{1 + 6.594}\right)\left(\frac{1}{1 + .061}\right) = 0.124,$$

which is the same as reported in the output [1a]. The F-ratios for all the tests indicate that the $GENDER \times AD$ interaction is statistically significant at an alpha of .05 [1a]. The effect sizes are quite large and have high power [1b].

Recall that in Section 11.2.4 we discussed the similarity between discriminant analysis and MANOVA. The dimension reduction analysis pertains to the results that would be obtained if a discriminant analysis is done for the interaction part of the MANOVA problem. The number of functions that one can have depends on the rank of the $\mathbf{SSCP}_h \times \mathbf{SSCP}_w^{-1}$ matrix, which is equal to the degrees of freedom for the respective effect. Note that only the first discriminant function (i.e., the first eigenvalue, or λ) is statistically significant [1d], and accounts for 99% of the interaction effect [1c]. The univariate tests for each measure of the $GENDER \times AD$ interaction effect are statistically significant and have high effect sizes and powers [1e, 1f].

Interpretation of the GENDER × AD Interaction

Results of discriminant analysis can be used to gain further insight into the nature of the interaction effect. Coefficients of the retained discriminant function(s) can be used

Exhibit 11.6

MANOVA for ad study

EFFECT .. GENDER BY AD

1a Multivariate Tests of Significance (S = 2, M = -1/2, N = 7 1/2)

Test Name	Value	Approx. F	Hypoth. DF	Error DF	Sig. of F
Pillais	.92596	7.75922	4.00	36.00	.000
Hotellings	6.65546	26.62185	4.00	32.00	.000
Wilks	.12409	15.62986	4.00	34.00	.000
Roys	.86832				

Note.. F statistic for WILKS' Lambda is exact.

1b Multivariate Effect Size and Observed Power at .0500 Level

TEST NAME	Effect Size	Noncent.	P
Pillais	.463	31.037	.99
Hotellings	.769	106.487	1.00
Wilks	.648	62.519	1.00

1c Eigenvalues and Canonical Correlations

Root No.	Eigenvalue	Pct.	Cum. Pct.	Canon Cor.
1	6.594	99.081	99.081	.932
2	.061	.919	100.000	.240

1d Dimension Reduction Analysis

Roots	Wilks L.	F	Hypoth. DF	Error DF	Sig. of F
1 TO 2	.12409	15.62986	4.00	34.00	.000
2 TO 2	.94236	1.10103	1.00	18.00	.308

1e Univariate F-tests with (2,18) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig. of F
Y1	156.00000	24.00000	78.00000	1.33333	58.50000	.000
Y2	149.33333	48.00000	74.66667	2.66667	28.00000	.000

1f

Variable	ETA Square	Noncent.	Power
Y1	.86667	117.00000	1.00000
Y2	.75676	56.00000	1.00000

1g Raw discriminant function coefficients

	Function No.
Variable	1
Y1	.984
Y2	-.114

(continued)

Exhibit 11.6 (continued)

2 EFFECT .. AD

```
Multivariate Tests of Significance (S = 2, M = -1/2, N = 7 1/2)

Test Name          Value  Approx. F Hypoth. DF   Error DF  Sig. of F
Pillais          .37696    2.09032       4.00      36.00       .102
Hotellings       .60504    2.42017       4.00      32.00       .069
Wilks            .62304    2.26867       4.00      34.00       .082
Roys             .37696
Note.. F statistic for WILKS' Lambda is exact.
```

3 EFFECT .. GENDER

```
Multivariate Tests of Significance (S = 1, M = 0, N = 7 1/2)

Test Name          Value    Exact F Hypoth. DF   Error DF  Sig. of F
Pillais          .23226    2.57143       2.00      17.00       .106
Hotellings       .30252    2.57143       2.00      17.00       .106
Wilks            .76774    2.57143       2.00      17.00       .106
Roys             .23226
Note.. F statistics are exact.
```

Table 11.14 **Cell Means for Multivariate Gender × Ad Interaction**

	Type of Ad		
Gender	**Humorous**	**Emotional**	**Comparative**
Male	7.944	5.334	2.724
Female	2.724	2.610	7.944

The discriminant function is (see Exhibit 11.6 [1g])

$$= .984\,Y_1 - .114\,Y_2$$

Average discriminant score for cell 1 (i.e., males, humorous ad)

$$= .984 \times 9 - .114 \times 8$$
$$= 7.944.$$

to form discriminant scores to represent the multivariate *GENDER* × *AD* interaction effect [1g]. In the present case only one discriminant function is retained. Table 11.14 gives average discriminant scores for the various cells of the *GENDER* × *AD* interaction effect and a sample computation, and Figure 11.5 gives the plot of the scores. From the figure it is clear that ad preference is a function of respondents' gender. Specifically, males prefer the humorous ad whereas females prefer the comparative ad. Figure 11.5 also shows the univariate plot for the interaction effects. The nature of the interaction effect is the same for both the dependent measures. That is, the males prefer the humorous ad while the females prefer the comparative ad.

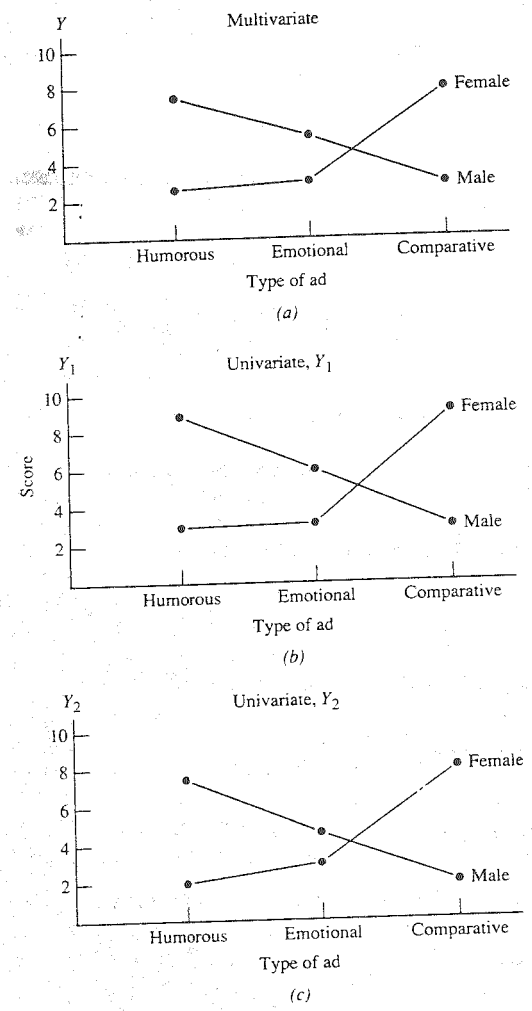


Figure 11.5 Gender × Ad Interaction. (*a*) Multivariate (*b*) Univariate, Y_1 (*c*) Univariate, Y_2

Significance Tests for Main Effects

Multivariate statistics for the two main effects (i.e., *GENDER* and *AD*) indicate that none of them are statistically significant [2, 3]. The univariate effects are not shown as they are normally interpreted only if the corresponding multivariate tests are significant. Also, no further interpretation of the main effects is provided as they are not significant.

11.6 SUMMARY

In this chapter we discussed multivariate analysis of variance (MANOVA), a technique that determines the effect of categorical independent variables on a number of continuous dependent variables. As such, it is a direct generalization of analysis of variance (ANOVA). It was seen that there is a close relationship between MANOVA and discriminant analysis, especially for

two-group or multiple-group MANOVA. In two-group MANOVA there is only one independent variable and it is at two levels. For multiple-group MANOVA the independent variable has more than two levels. In the next chapter we discuss the assumptions made in MANOVA and discriminant analysis, the effects of violating these assumptions on the results, and the available procedures for testing the assumptions.

QUESTIONS

11.1 For the following two data sets:

Group 1		Group 2	
Y_1	Y_2	Y_1	Y_2
1	2	5	11
3	5	7	15
4	9	8	17
5	9	10	20

Group 1		Group 2	
Y_1	Y_2	Y_1	Y_2
1	5	6	10
2	4.7	7	9.7
5	1	10	6
4	3.2	9	8.2
5	1	10	6
3	4.1	8	4.1

(a) State the null and the alternative hypotheses in words.
(b) Replicate the calculations given in Table 11.2 of the text.
(c) What conclusions can you draw from your calculations?
(d) Will there be a difference in the results of univariate and multivariate significance tests? Explain. (*Hint:* Compute the correlation between Y_1 and Y_2.)
(e) If a two-group discriminant analysis is done, what would be the eigenvalue?
(f) Analyze the data using two-group discriminant analysis and compare the results.

11.2 Table 7.19 of the text gives the nutrient data for various food items. Assuming that there are three clusters (the cluster memberships are given in Exhibit 7.5) conduct a MANOVA using the food nutrients as the dependent variables, and interpret the results. Compare your results to those obtained by multiple-group discriminant analysis.

11.3 The management of a national department store chain developed a discriminant model to classify various departments (e.g., men's wear, appliances, electronics, etc.) into high-, medium-, and low-performance departments. File STORE.DAT gives the discriminant scores for four consecutive quarters for the departments classified into each category (*Note:* Higher scores imply better performance). Analyze the data using a repeated-measures MANOVA with the following SPSS commands:

```
MANOVA QTR1, QTR2, QTR3, QTR4 BY PERFORM(1,3)
 /WSFACTOR=QTR(4)
 /WSDESIGN=QTR
 /DESIGN=PERFORM
```

(a) What conclusions can you draw from the analysis?
(b) For each performance category plot the average score using the quarters on the X-axis. Based on a visual analysis, is there a trend? Is the trend different across the groups (*Note:* Use the interaction between QTR and PERFORM)? What conclusions can you draw from the trend?
(c) Decompose the trend of each performance category into linear, quadratic, and cubic trends. The following SPSS commands can be used for this purpose:

```
SELECT IF (PERFORM EQ 1)
MANOVA QTR1, QTR2, QTR3, QTR4
 /WSFACTOR=QTR(4)
 /WSDESIGN=QTR
 /CONTRAST=POLYNOMIAL
```

What conclusions can you draw from the analysis?

11.4 Analyze the data in file FIN.DAT using industry as the grouping variable. Interpret the results, and compare the results to those obtained by multiple-group discriminant analysis.

11.5 Analyze the data given in files DEPRES.DAT and PHONE.DAT using MANOVA (use CASES as the grouping variable for data given in DEPRES.DAT, and the number of phones owned as the grouping variable for PHONE.DAT) and compare your results to the results obtained by discriminant analysis.

11.6 Consider the following data:

Y_1	Y_2	Y_3	Y_4	Group
7	3	1	2	1
8	2	3	3	1
9	1	5	5	1
9	5	3	4	2
10	4	5	5	2
11	8	7	6	2

Run a MANOVA and interpret the results. What is the problem with the data and how would you rectify it?

11.7 For each of the following contrasts indicate whether they are orthogonal or nonorthogonal and discuss the effects that are being tested.

μ_1	μ_2	μ_3	μ_4
Contrast 1			
1	0	0	−1
0	1	0	−1
0	0	1	−1
Contrast 2			
0.5	0.5	−1	0
−0.5	0.5	0	0
0.33	0.33	0.33	−1
Contrast 3			
1	0	−1	0
0	1	−1	0
0	0	−1	1
Contrast 4			
−3	−1	1	3
1	−1	−1	1
1	3	−3	1

11.8 (a) Analyze the data in the table given below using MANOVA and interpret the solution.

	Factor B			
	1		2	
	Y_1	Y_2	Y_1	Y_2
Factor A				
1	2	8	6	4
	5	7	4	3
2	6	12	10	11
	7	9	12	8
3	8	15	10	11
	14	13	12	12

(b) Using appropriate contrasts, determine if group 2 of factor A is significantly different from group 1 of factor A, and if group 3 of factor A is significantly different from group 1 of factor A. Are these contrasts orthogonal? Why or why not?

(c) Using appropriate contrasts, determine if groups 2 and 3 of factor A are significantly different from each other. Are these contrasts orthogonal? Why or why not?

CHAPTER 12

Assumptions

As is the case with most of the multivariate techniques, properties of the statistical tests in MANOVA and discriminant analysis are derived or obtained by making a number of assumptions. In most empirical studies some or all of the assumptions made will be violated to varying degrees, which could affect the properties of the statistical tests. In this chapter we discuss these assumptions, and the effects of their violation on the statistical tests. Also, we discuss suggested procedures to test whether the data meet the assumptions, appropriate data transformations that can be used to make the data conform to the assumptions, and steps that can be taken to mitigate the effect of violation of the assumptions on statistical tests. The assumptions are:

1. The data come from a multivariate normal distribution.
2. The covariance matrices for all the groups are equal.
3. The observations are independent. In other words, each observation is independent of other observations.

Violation of the above assumptions can affect the significance and power of the statistical tests. Before we discuss the effects of each of these assumptions and the available tests for checking if an assumption is violated, we provide a brief discussion of significance and power of test statistics.

12.1 SIGNIFICANCE AND POWER OF TEST STATISTICS

Two types of errors are commonly made while testing the null and alternative hypotheses: Type I and Type II errors. Type I error, usually labeled as the significance or alpha level of a given test statistic, is the probability of *falsely rejecting the null hypothesis due to chance*. The researcher typically selects the desired or nominal alpha level (i.e., the value for Type I error) to test the hypotheses. A nominal alpha level of, say, .05 means that if the study were replicated several times, one would expect to falsely reject the null hypothesis in about 5% of the studies due to chance alone. However, if some of the assumptions are violated then the actual number of times the null hypothesis will be falsely rejected could be more or less than the nominal alpha level. For example, it is quite possible that violating the multivariate normality assumption could result in an actual alpha level of .20 even though the nominal or desired alpha level selected was .05.

Type II error, usually represented by β, is the probability of failing to reject the null hypothesis when in fact it is false. The power of a test is given by $1 - \beta$ and it is the *probability of correctly rejecting the null hypothesis when it is false.* If the power is low, then the probability of finding statistically significant results decreases. Obviously, the researcher would like to have a small alpha level and a high power. Therefore, it is important to know how the alpha level and power of the test statistic are affected by violation of the assumptions.

12.2 NORMALITY ASSUMPTIONS

Almost all parametric statistical techniques assume that the data come from a multivariate normal distribution. Research has found that for univariate (e.g., ANOVA) and multivariate techniques (e.g., MANOVA and discriminant analysis), violation of the normality assumption does not have an appreciable effect on the Type I error (Glass, Peckham, and Sanders 1972; Everitt 1979; Hopkins and Clay 1963; Mardia 1971; and Olson 1974). As discussed in Chapter 8, violation of the normality assumption does have an effect on the classification rates. Also, as discussed in the following, violation of the normality assumption does affect the power of the test statistic.

A univariate normal distribution has zero skewness and a kurtosis of three. Sometimes the kurtosis is normalized by subtracting three so that its value is zero for the normal distribution. Henceforth we will use the term *kurtosis* to refer to its normalized value. That is, a univariate normal distribution has zero skewness and zero kurtosis. A negatively skewed distribution has a skewness of less than zero and a positively skewed distribution has a skewness of greater than zero. Similarly, a multivariate distribution is said to be skewed if its multivariate measure of skewness is not equal to zero. Research has shown that the power of a test is not affected by violation of the normality assumption if the nonnormality is solely due to skewness.

A distribution is said to be leptokurtic (i.e., peaked) if its kurtosis is positive and platykurtic (i.e., flat) if its kurtosis is negative. Kurtosis does seem to have an effect on the power of a test statistic; however, the effect is more severe for platykurtic distributions than for leptokurtic distributions. Olson (1974) found that for MANOVA the power of the test decreased substantially for platykurtic distributions. Furthermore, the severity of the effect increases as the assumption is violated for more than one cell or group. Since normality affects the power of the test, it is advisable to determine if the normality assumption has been violated. In the following section we discuss tests for assessing univariate normality, and in Section 12.4 we discuss tests for assessing multivariate normality.

12.3 TESTING UNIVARIATE NORMALITY

Tests of univariate normality are discussed for several reasons. First, tests of multivariate normality are more complex and difficult, and understanding them is facilitated by an understanding of univariate tests. Second, although it is possible that the multivariate distribution may not be normal even though all the marginal distributions are normal, such cases are rare. As stated by Gnandesikan (1977), it is only in rare cases that multivariate nonnormality will not be detected by univariate normality tests. Finally, if the data do not come from a multivariate normal distribution then one would like to further

investigate which variable's distribution is not normal. Such an investigation is necessary if one wants to transform the data for achieving normality. To test for univariate normality, one can employ graphical or analytical tests, as described in the following pages.

12.3.1 Graphical Tests

A number of graphical tests such as the stem-and-leaf plot, the box-and-whiskers plot, and the Q-Q plot have been proposed. Of these, the Q-Q plot is the most popular and is the plot discussed in this chapter. For discussion of other tests the interested reader is referred to Tukey (1977). Use of the Q-Q plot is illustrated with hypothetical data simulated from a normal distribution having a mean of 10 and a standard deviation of 2. The Q-Q plot is obtained as follows:

1. Order the observations in ascending order such that $X_1 < X_2 \cdots < X_n$, where n is the number of observations. The ordered observations are given in Column 2 of Table 12.1. Given that the value of each observation is unique, as is usually the case for continuous variables, then exactly j observations will be less than or equal to X_j. Each ordered observation therefore represents a sample quantile.
2. The proportion of observations that are less than X_j is estimated by $(j - .5)/n$. The quantity .5 is subtracted for continuity correction. Column 3 of Table 12.1 gives the proportion of observations that would be less than X_j. For each j, these proportions are assumed to be percentiles or probability levels for the cumulative standard normal distribution and the corresponding Z-values give the expected or theoretical quantiles, for a normal distribution. The Z-values can be obtained from the cumulative normal distribution table or from the PROBIT function in SAS.

Table 12.1 **Hypothetical Data Simulated from Normal Distribution**

Observation Number (j) (1)	Ordered Value (X_j) (2)	Probability Level or Percentile (3)	Z-value (4)
1	4.813	0.033	−1.834
2	6.937	0.100	−1.282
3	7.027	0.167	−0.967
4	7.804	0.233	−0.728
5	8.560	0.300	−0.524
6	8.727	0.367	−0.341
7	9.754	0.433	−0.168
8	9.996	0.500	0.000
9	10.053	0.567	0.168
10	10.139	0.633	0.341
11	10.202	0.700	0.524
12	10.240	0.767	0.728
13	11.918	0.833	0.967
14	11.943	0.900	1.282
15	12.027	0.967	1.834

3. Column 4 of Table 12.1 gives the Z values. The plot between the ordered observations or values (i.e., X_j) and theoretical quantiles (e.g., Z) is called the Q-Q plot and is shown in Figure 12.1. A linear plot indicates that the distribution is normal and a nonlinear plot indicates that the distribution is nonnormal.

The plot in Figure 12.1 is approximating linear, suggesting that the data in Table 12.1 are normally distributed. The plot is not completely linear because of the small sample size. For a large sample size the plot would have been linear, as the simulated data do come from a normal distribution. To illustrate the plot for a nonnormal distribution, the data in Table 12.1 were transformed as $Y = e^X$. The plot of the transformed data is given in Figure 12.2. It is clear that the Q-Q plot in Figure 12.2 deviates substantially from linearity.

Obviously, the test based on the Q-Q plot is subjective, for the researcher has to visually establish whether the plot is linear or not. For this purpose one could use "training plots" given in Daniel and Wood (1980) to assess the linearity of the plot. Alternatively, and more preferably, the linearity of the Q-Q plot can be assessed by computing the correlation coefficient between the sample (i.e., X_j) and theoretical quantiles, and comparing it with the critical value given in Table T.5 in the Statistical Tables following Chapter 14. Values in Table T.5 give the percent points of the cumulative sampling distribution of the correlation between sample values and theoretical quantiles obtained empirically by Filliben (1975). The correlation coefficients for the plots in Figures 12.1 and 12.2, respectively, are .967 and .814. It can be seen that the correlation of .967 is well above the critical value of .937 for alpha level of .05 and $n = 15$, suggesting that the respective Q-Q plot is linear and, therefore, the data set in Table 12.1 does come from a normal distribution. However, the transformed data do not come from a normal distribution as the correlation of .814 for the plot in Figure 12.2 is not greater than the critical value of .937.

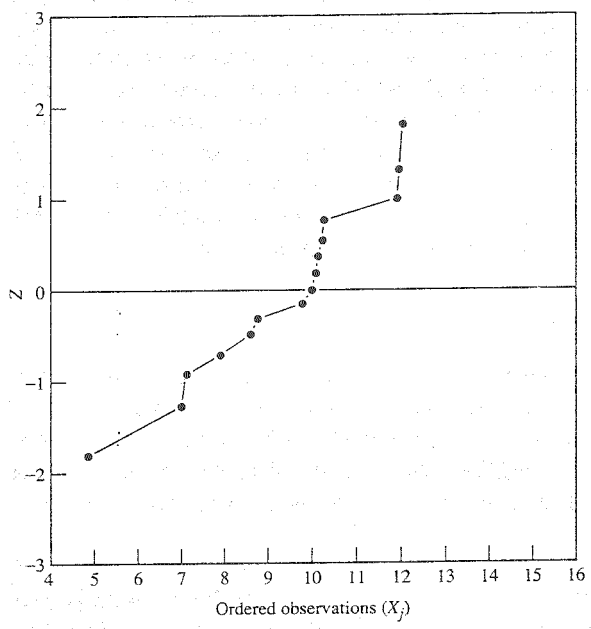


Figure 12.1 Q-Q Plot for data in Table 12.1.

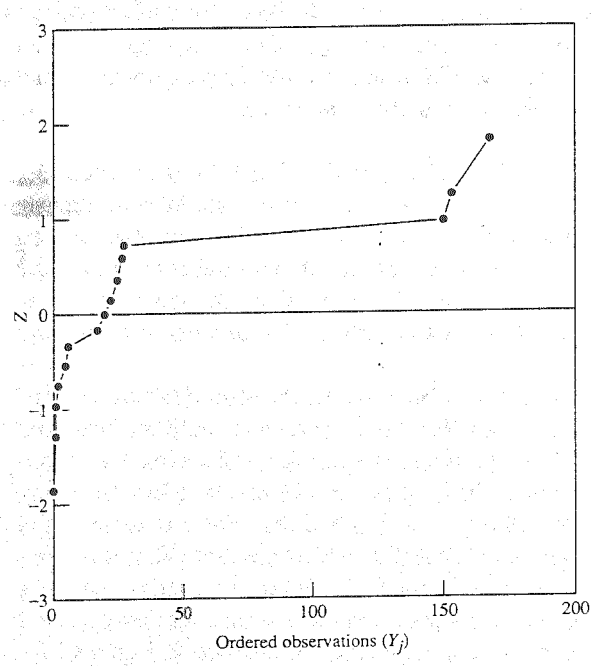


Figure 12.2 Q-Q Plot for transformed data.

12.3.2 Analytical Procedures for Assessing Univariate Normality

Some of the analytical procedures or tests for assessing normality are the chi-square goodness of fit, the Kolmogorov-Smirnov test, and the Shapiro-Wilk test. Simulation studies conducted by Wilk, Shapiro, and Chen (1968) concluded that the Shapiro-Wilk test was the most powerful test in assessing univariate normality. In case the data do not come from a normal distribution, further assessment can be done by examining the skewness and kurtosis of the distribution. The EXAMINE procedure in **SPSS** can be used to obtain the preceding statistics. The EXAMINE procedure also can be used to obtain the Q-Q plot discussed in the previous section. In the following section we use the data set given in Table 12.2, which gives the financial ratios for most- and least-admired firms, to discuss the resulting output from the EXAMINE procedure. This data set has been used in Chapter 8 to illustrate two-group discriminant analysis.

12.3.3 Assessing Univariate Normality Using SPSS

The **SPSS** commands for the EXAMINE procedure to assess the normality assumption for *EBITASS* are given in Table 12.3. The EXAMINE command requests information for *EBITASS* to evaluate its distribution. The plot option specifies the type of plot desired. NPPLOT gives the Q-Q plot described earlier. The DESCRIPTIVE option in the STATISTICS subcommand requests printing of a number of descriptive statistics. Exhibit 12.1 gives the output.

SPSS refers to the Q-Q plot as the normal plot. It can be seen that the plot is similar to that given in Figure 12.1, and its linearity suggests that *EBITASS* is normally distributed [2a]. The *detrended normal plot* gives the plot of the residuals after removing the linearity effect [2b]. For a normal distribution, the detrended plot should be random

Table 12.2 **Financial Data for Most-Admired and Least-Admired Firms**

Obs	*ROTC*	*EBITASS*	Squared Mahalanobis Distance (MD^2)
1	0.182	0.158	1.289
2	0.206	0.210	1.150
3	0.188	0.207	1.098
4	0.236	0.280	3.648
5	0.193	0.197	0.901
6	0.173	0.227	3.058
7	0.196	0.148	3.101
8	0.212	0.254	2.856
9	0.147	0.079	4.802
10	0.128	0.149	0.390
11	0.150	0.200	2.331
12	0.191	0.187	0.879
13	−0.031	−0.012	1.415
14	0.053	0.036	0.641
15	0.036	0.038	0.305
16	−0.074	−0.063	2.440
17	−0.119	−0.054	6.335
18	−0.005	0.000	0.850
19	0.039	0.005	1.787
20	0.122	0.091	1.173
21	−0.072	−0.036	2.918
22	0.064	0.045	0.643
23	−0.024	−0.026	1.318
24	0.026	0.016	0.672

Table 12.3 **SPSS Commands**

```
EXAMINE VARIABLES=EBITASS
 /PLOT=NPPLOT
 /STATISTICS=DESCRIPTIVE
```

and centered around zero. This appears to be the case, suggesting once again that the distribution of *EBITASS* is normal. Both the Shapiro-Wilk and the Kolmogorov-Smirnov test statistics are not significant [3]. It is obvious that all the procedures suggest that the distribution of *EBITASS* is normal.

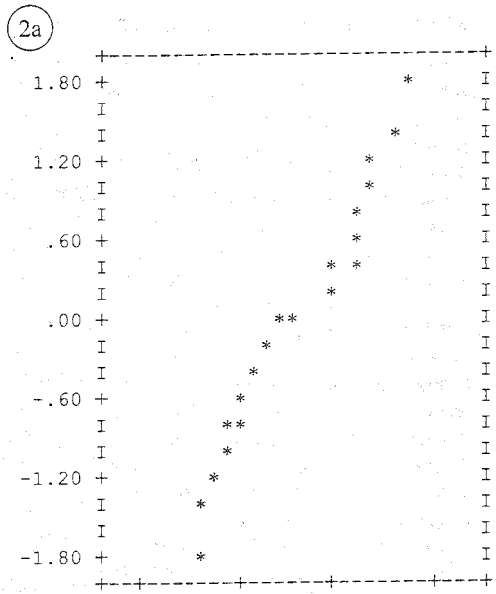
If the distribution of *EBITASS* were not normal, one would examine skewness and kurtosis to obtain further insights into the nature of nonnormality. Skewness and kurtosis, along with their standard errors, are printed for *EBITASS* [1]. For large sample sizes (e.g., 25 or more) the standard errors can be used to compute the Z-values, which are .176 (.0833/.472) for skewness and 1.562 (−1.433/.918) for kurtosis [1]. Since both of these are less than the critical value of 1.96 for an alpha level of .05, it is concluded that the distribution of *EBITASS* is normal. For smaller samples the critical values obtained via simulation by D'Agostino and Tietjen (1971; 1973) can be used. These critical values are reproduced in Table T.6 (Statistical Tables).

Exhibit 12.1

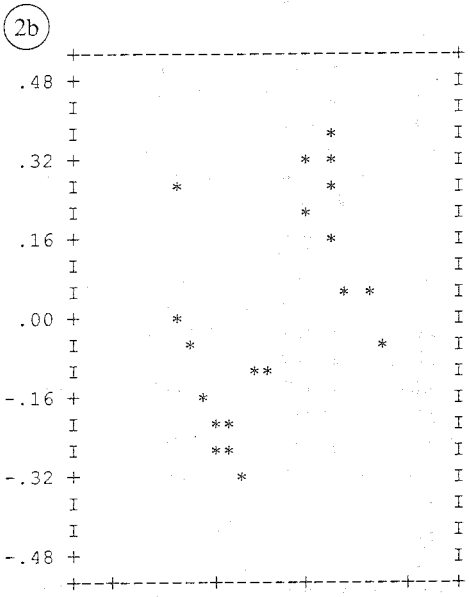
Univariate normality tests for data in Table 12.2

```
(1) EBITASS
    Valid cases:     24.0   Missing cases:     .0   Percent missing:     .0

Mean        .0973  Std Err     .0219  Min       -.0630  Skewness      .0833
Median      .0850  Variance    .0115  Max        .2800  S E Skew      .4723
5% Trim     .0962  Std Dev     .1074  Range      .3430  Kurtosis    -1.4332
                                      IQR        .1980  S E Kurt      .9178
```

(2a)



(2b)



(3)

	Statistic	df	Significance
Shapiro-Wilks	.9301	24	.1010
K-S (Lilliefors)	.1453	24	> .2000

12.4 TESTING FOR MULTIVARIATE NORMALITY

There are very few tests for examining multivariate normality. The graphical test is similar to the Q-Q plot discussed for the univariate case. The analytical tests simply assess the multivariate measures of skewness and kurtosis. Unfortunately, not many programs have the option for computing these statistics.[1] Furthermore, the distribution of these test statistics is not known, leading to their limited use for assessing multivariate normality. Consequently, only the graphical procedure is described below. The data set given in Table 12.2 is used to illustrate the graphical test.

The first step is to compute squared Mahalanobis distance ($M\,D^2$) of each observation from the sample centroid. This distance is also reported in Table 12.2. It has been shown that when the parent population is normal and the sample size is sufficiently large

[1]EQS, a covariance structure analysis program in BMDP, computes these statistics.

(e.g., 25 or more) these distances behave like a chi-square random variable (Johnson and Wichern 1988). This property can be used to obtain a chi-square plot as follows (see Gnandesikan 1977):

1. First, order the MD^2 from lowest to the highest such that $MD_1^2 < MD_2^2 < \cdots < MD_n^2$ where n is the number of observations. The ordered distances are presented in Column 2 of Table 12.4.
2. For each MD^2, compute the $(j - .5)/n$ percentile where j is the observation number. These percentiles are reported in Column 3 of Table 12.4.
3. The χ^2 values for the percentiles are obtained from the χ^2 distribution with p degrees of freedom where p is the number of variables. The χ^2 values can be obtained from the χ^2 tables or using the CINV function in SAS. The χ^2 values are given in Column 4 of Table 12.4.
4. MD^2 and χ^2 are then plotted. The χ^2 plot is shown in Figure 12.3 and is similar to the Q-Q plot. The plot should be linear and any deviation from linearity indicates nonnormality.

Table 12.4 **Ordered Squared Mahalanobis Distance and Chi-Square Value**

Observation Number (j) (1)	Ordered Squared Mahalanobis Distance (MD^2) (2)	Percentiles (3)	Chi-Square Value (4)
1	0.305	0.021	0.042
2	0.390	0.063	0.129
3	0.641	0.104	0.220
4	0.643	0.146	0.315
5	0.672	0.188	0.415
6	0.850	0.229	0.521
7	0.879	0.271	0.632
8	0.901	0.313	0.749
9	1.098	0.354	0.874
10	1.150	0.396	1.008
11	1.173	0.438	1.151
12	1.289	0.479	1.305
13	1.318	0.521	1.471
14	1.415	0.563	1.653
15	1.787	0.604	1.854
16	2.331	0.646	2.076
17	2.440	0.688	2.326
18	2.856	0.729	2.613
19	2.918	0.771	2.947
20	3.058	0.813	3.348
21	3.101	0.854	3.851
22	3.648	0.896	4.524
23	4.802	0.938	5.545
24	6.335	0.979	7.742

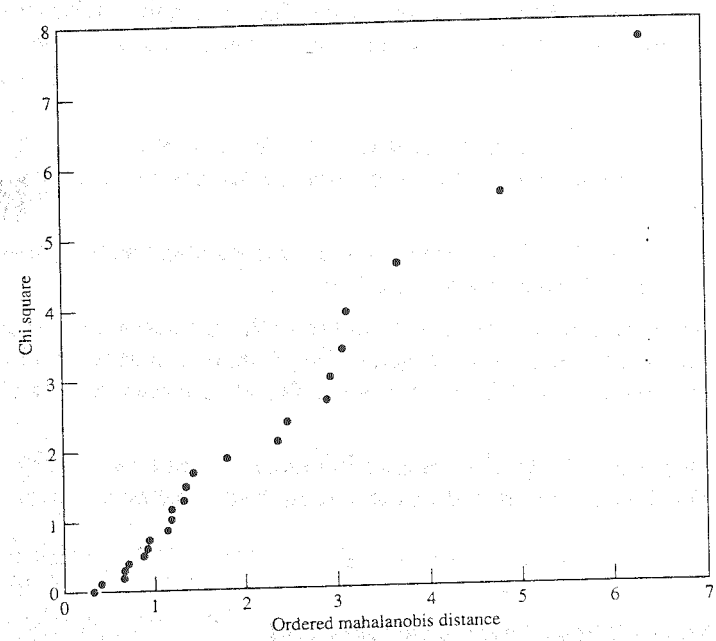


Figure 12.3 Chi-square plot for total sample.

The plot in Figure 12.3 appears to be linear, from which we conclude that the assumption of multivariate normality is a reasonable one. As discussed earlier, one could compute the correlation coefficient of the plot and compare it with the critical values given in Table T.5 of the Statistical Tables. Although these critical values were obtained for univariate distributions, we feel that they provide a reasonable benchmark. The correlation coefficient for the plot in Figure 12.3 is 0.990, which is greater than the critical value of 0.957 for alpha = .05 and $n = 24$, from which it can be concluded that the data do come from a multivariate normal distribution. Unfortunately, none of the statistical packages has a procedure to obtain the χ^2 plots. However, PROC IML in SAS can be used to obtain these plots. The Appendix to this chapter gives the PROC IML program for obtaining the χ^2 plot.

12.4.1 Transformations

If the data do not come from a normal distribution, one can transform the data such that the distribution of the transformed variable is normal. In the multivariate case, each variable whose marginal distribution is not normal is transformed to make its distribution normal. The type of transformation depends on the type of nonnormality with respect to skewness and kurtosis. However, in general, the square-root transformation works best for data based on counts, the logit transformation for proportions, and the Fisher's Z transformation for correlation coefficients. Table 12.5 gives various transformations that have been suggested to achieve normality. In situations where none of these transformations are appropriate, one can use analytical procedures to identify the type of power transformation necessary to achieve normality. These procedures are discussed in Johnson and Wichern (1988).

Table 12.5 **Transformations To Achieve Normality**

Type of Scale	Transformation
Counts	Square-root transformation
Proportions (p)	$\text{logit(p)} = 0.5log\left(\frac{p}{1-p}\right)$
Correlations (r)	$\text{Fisher's } Z = 0.5log\left(\frac{1+r}{1-r}\right)$

12.5 EFFECT OF VIOLATING THE EQUALITY OF COVARIANCE MATRICES ASSUMPTION

In a univariate case (e.g., ANOVA) the covariance matrix is a scalar and the assumption is met if the variance of the dependent variable is the same for all the cells. However, in the case of MANOVA and discriminant analysis, the equality of covariance matrices assumption is met only if the covariance matrices of all the cells are equal. Two matrices are said to be equal *if and only if* all the corresponding elements of the matrices are equal. For example, in the case of three dependent variables there will be six elements in the covariance matrix: three variances, σ_1^2, σ_2^2, and σ_3^2, and three covariances, σ_{12}, σ_{13}, and σ_{23}. All the corresponding six elements of the matrices would have to be equal to satisfy the equality of covariance matrices assumption. Therefore, there is a greater chance that the equality of covariance matrices assumption will be violated in MANOVA and discriminant analysis than in ANOVA.

Violation of the equality of covariance matrices assumption affects Type I and Type II errors. However, simulation studies have found that the effect is much more for a Type I error than for a Type II error. Consequently, most of our discussion pertains to how the *significance* level is affected (the Type I error). Research has shown that for equal cell sizes the significance level is not appreciably affected by unequal covariance matrices (Holloway and Dunn 1967; Hakstian, Roed, and Linn 1979; and Olson 1974). Therefore, every effort should be made to have equal cell sizes. However, for unequal cell sizes the significance level can be severely affected even for moderate differences in the covariance matrices. In the case of two groups, the following specific findings were obtained from simulation studies (Holloway and Dunn 1967; Hakstian, Roed, and Linn 1979). The test is liberal if the smaller group has more variability, that is, the actual alpha level of the test is more than the nominal alpha level. On the other hand, if the variability of the larger group is more than that of the smaller group then the test is conservative; that is, the actual alpha level will be less than the nominal alpha level. The implications of these findings are as follows:

1. If the test statistic is conservative due to unequal covariance matrices, then there is no need to be concerned about significant results because the results would still be significant after transforming the variables to achieve equality of covariance matrices, and consequently the conclusions of the study will not change. On the other hand, there is a need for concern about insignificant results. In this case,

transformation of variables to obtain equal covariance matrices could result in significant findings that will change the conclusions of the study.

2. If the test is liberal due to unequal covariance matrices, then there is no need to be concerned about insignificant results as they will still be insignificant after the necessary transformations. But, one does need to be concerned about significant results. In this case the researcher may not be sure whether the significance is due to actual differences or due to chance because of the effect of the inequality of covariance matrices.

As discussed in Chapter 8, violation of the equality of covariance matrices does affect the classification rates of discriminant analysis.

Table 12.6 **Data for Purchase Intention Study**

Segment (Group) 1		Segment (Group) 2	
Y_1	Y_2	Y_1	Y_2
1.180	5.209	12.017	5.209
2.480	6.563	14.774	6.563
5.107	5.054	20.349	5.054
4.273	6.698	18.579	6.698
5.240	4.638	20.630	4.638
3.913	5.694	17.815	5.694
2.025	5.858	13.810	5.858
3.469	7.012	16.873	7.012
4.232	6.517	18.492	6.517
4.660	4.159	19.401	4.159
2.193	7.487	14.166	7.487
3.288	4.935	16.489	4.935
4.656	6.765	19.392	6.765
5.442	5.770	21.060	5.770
4.024	5.305	18.052	5.305
4.686	5.024		
4.465	5.711		
3.157	5.385		
4.382	6.945		
3.216	7.557		
2.172	6.094		
5.607	4.271		
2.596	6.232		
2.332	5.419		
4.492	8.324		
4.363	4.575		
3.032	3.403		
5.100	7.931		
5.040	7.052		
4.853	6.879		

12.5.1 Tests for Checking Equality of Covariance Matrices

Almost all the test statistics for assessing equality of covariance matrices are sensitive to nonnormality. Therefore, the data must first be checked for normality. If the data do not come from a multivariate normal distribution then appropriate transformations to achieve normality should be done before testing for equality of covariance matrices. The most widely used test statistic is Box's M and it is available in the MANOVA and the discriminant analysis procedure in SPSS. The use of Box's M statistic is illustrated here.

Consider a study that compares awareness and attention levels of a commercial for two segments or groups of sizes 30 and 15. Table 12.6 gives the data. The first step is to test for multivariate normality. Figure 12.4 gives the χ^2 plot obtained by using the PROC IML program given in the Appendix to this chapter, and it appears to be linear. The correlation coefficient for the plot is .984, which is greater than the critical value of .974 for an alpha level of .05 for $n = 45$. That is, the data meet the multivariate normality assumption.

Exhibit 12.2 gives the partial MANOVA output. The Box's M statistic is significant, indicating that the covariance matrices are not equal [3]. The generalized variances, given by the determinant of the covariance matrix, for groups 1 and 2, respectively, are 1.992 and 6.392 [2b, 2c]. Since the variability of the smaller group is more, it would be expected that the test is liberal. That is, the significance of the multivariate tests could be due to chance [4]. Therefore, the analysis should be repeated after using the appropriate transformations so that the covariance matrices of the two groups are equal. For this purpose univariate tests are done to determine which variable has a variance

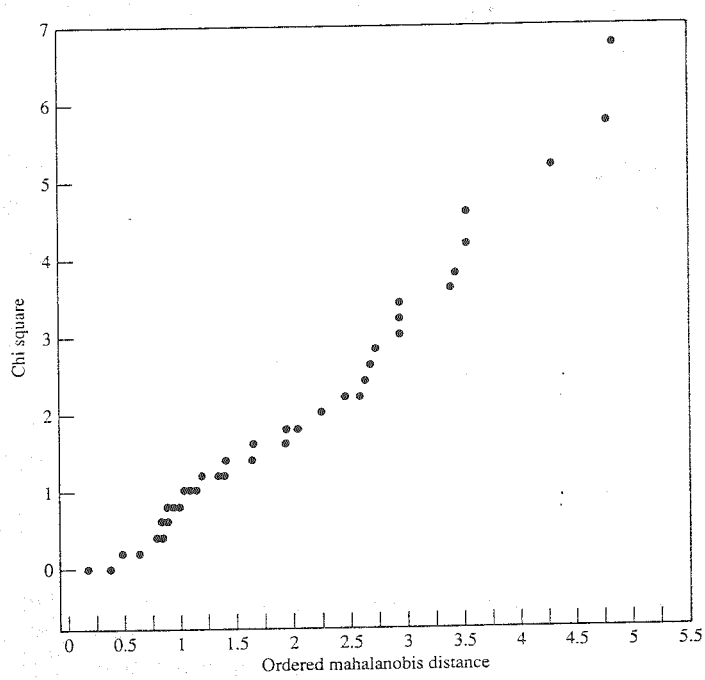


Figure 12.4 Chi-square plot for ad awareness data.

Exhibit 12.2

Partial MANOVA Output for Checking Equality of Covariance Matrices Assumption

```
Cell Means and Standard Deviations

(1) Variable .. Y1                                  Variable .. Y2
  FACTOR    CODE      Mean Std.Dev.  N      FACTOR    CODE    Mean  Std.Dev.  N

 EXCELL      1       3.856    1.183 30    EXCELL       1    5.949     1.193 30
 EXCELL      2      17.460    2.725 15    EXCELL       2    5.844      .966 15
For entire sample   8.391    6.734 45   For entire sample 5.914     1.112 45

(2a) Univariate Homogeneity of Variance Tests
Variable .. Y1

      Cochrans C(22,2) =                          .84135, P =  .000 (approx.)
      Bartlett-Box F(1,3952) =                  13.96908, P =  .000

Variable .. Y2

      Cochrans C(22,2) =                          .60372, P =  .331 (approx.)
      Bartlett-Box F(1,3952) =                    .77112, P =  .380

(2b) Cell Number .. 1

Determinant of Variance-Covariance matrix =          1.99163
LOG(Determinant) =                                    .68895

(2c) Cell Number .. 2

Determinant of Variance-Covariance matrix =          6.39151
LOG(Determinant) =                                   1.85497

(3) Multivariate test for Homogeneity of Dispersion matrices
Boxs M =                            15.56744
F WITH (3,18522) DF =                4.87854, P =   .002 (Approx.)
Chi-Square with 3 DF =              14.63813, P =   .002 (Approx.)

(4) Multivariate Tests of Significance (S = 1, M = 0, N = 20 )

Test Name          Value    Exact F Hypoth. DF   Error DF  Sig. of F
Pillais           .92842  272.37374       2.00      42.00       .000
Hotellings      12.97018  272.37374       2.00      42.00       .000
Wilks            .07158  272.37374       2.00      42.00       .000
Roys             .92842
Note: F statistics are exact.
```

that is different for the two groups. Cochran's C and Bartlett-Box's F-test indicate that only the variance of Y_1 is different for the two groups [2a]. One of the transformations to stabilize the variance of a given variable is the square-root transformation. The square-root transformation works best when the mean-to-variance ratio is equal for the

Exhibit 12.3

Partial MANOVA Output for Checking Equality of Covariance Matrices Assumption for Transformed Data

(1) Multivariate test for Homogeneity of Dispersion matrices

```
Boxs M =                          1.55529
F WITH (3,18522) DF =              .48740, P =   .691 (Approx.)
Chi-Square with 3 DF =            1.46244, P =   .691 (Approx.)
```

(2) Multivariate Tests of Significance (S = 1, M = 0, N = 20)

Test Name	Value	Exact F	Hypoth. DF	Error DF	Sig. of F
Pillais	.91502	226.11447	2.00	42.00	.000
Hotellings	10.76736	226.11447	2.00	42.00	.000
Wilks	.08498	226.11447	2.00	42.00	.000
Roys	.91502				

Note: F statistics are exact.

groups.[2] The mean-to-variance ratios of Y_1 for groups 1 and 2, respectively, are 2.755 ($3.856/1.183^2$) and 2.351 ($17.460/2.725^2$) [1]. Since these are approximately equal, Y_1 will be transformed by taking its square root. Exhibit 12.3 gives the partial MANOVA output.

The Box's M is not significant, suggesting that the covariance matrices are equal [1]. The multivariate tests are still significant and the conclusion does not change. However, notice that, as expected, the corresponding F-values are lower than those in Exhibit 12.2 [2].

12.6 INDEPENDENCE OF OBSERVATIONS

It is unfortunate that most textbooks do not discuss the independence assumption, as this assumption has a substantial effect on significance level and the power of a test. Two observations are said to be independent if the outcome of one observation is not dependent on another observation. In behavioral research this implies that the response of one subject should be independent of the responses of other subjects. This may not be true though. For example, dependent observations will normally result when data are collected in a group setting because it is quite possible that nonverbal and other modes of communication among respondents could affect their responses. Although group setting is not the only source, it is the most common source for dependent observations. Kenny and Judd (1986) discuss other situations that could give rise to dependent observations.

[2]Other transformations that are commonly used include the arcsine transformation, which works best for proportions. The log transformation is also used commonly. Both the log and square-root transformations will not work on negative values of the data. A constant can be added to all data values to resolve the problem.

Research has indicated that for correlated observations the actual alpha level could be as much as ten times the nominal alpha level (Scariano and Davenport 1986), and the effect worsens as the sample size increases. It is interesting to note that this is one situation where a larger sample may not be desirable!

Unfortunately, no sophisticated tests are available to check if the independence assumption is violated. Therefore, while collecting the data the researcher should take extreme caution in making sure that the observations are independent. Glass and Hopkins (1984) make the following statement regarding the conditions under which the independence assumption is most likely to be violated: "whenever the treatment is individually administered, observations are independent. But where treatments involve interaction among persons, such as 'discussion' method or group counseling, the observations may influence each other" (p. 353). If it is found that the independence assumption does not hold then one can use a more stringent alpha level. For example, if the actual alpha level is ten times the nominal alpha level then instead of using an alpha level of .05 one can use an alpha level of .005.

12.7 SUMMARY

The statistical tests in MANOVA and discriminant analysis assume that the data come from a multivariate normal distribution and the equality of covariance matrices for the groups. Violation of these assumptions affects the significance and power of the tests. This chapter describes the effect of the violation of these assumptions on the statistical significance and power of the tests, and the available graphical and analytical techniques to test the assumptions. Also discussed are data transformations that one can employ so that the data meet these assumptions.

The next chapter discusses canonical correlation analysis. As mentioned in Chapter 1, most of the dependence techniques discussed so far are special cases of canonical correlation analysis.

QUESTIONS

12.1 File TAB12-1.DAT presents data on three variables (X_1–X_3) for two groups. Do the following:
- (a) Extract the first 10 observations from each group.
  - (i) Check the normality of each variable for the total sample and for each group.
  - (ii) Check the multivariate normality for the total sample and for each group.
  - (iii) Check for equality of covariance matrices for the two groups.
- (b) Extract the first 30 observations from each group and repeat the assumption checks given above.
- (c) Repeat the assumption checks given above for the complete data.
- (d) Comment on the effect of sample size on the results of the assumption checks.

12.2 File TAB12-2.DAT presents data on three variables.
- (a) Check the normality of each variable.
- (b) Check the multivariate normality for the three variables.
- (c) For the variables in part (a) that do not have a normal distribution, what is an appropriate transformation that can be applied to make the distribution of the transformed variables normal?
- (d) Check the multivariate normality of the three variables after suitable transformations have been applied to ensure univariate normality.

FOR EACH OF THE DATA SETS INDICATED BELOW DO THE FOLLOWING:

1. Check for univariate and multivariate normality of the data.
2. What transformations, if any, are required to ensure univariate/multivariate normality of the data?

12.3 Data in file FOODP.DAT.

12.4 Data in file AUDIO.DAT.

12.5 Data in file NUT.DAT.

12.6 Data in file SOFTD.DAT.

12.7 Data in file SCORE.DAT.

12.8 Data in Table Q8.2. Also check for equality of covariance matrices for the two groups (least likely to buy and most likely to buy).

12.9 Data referred to in Question 8.8. Also check for equality of covariance matrices for the two groups (users and nonusers of mass transportation).

12.10 Data in file PHONE.DAT. Also check for equality of covariance matrices for the three groups (families owning one, two, or three or more telephones).

12.11 Data in file ADMIS.DAT.

Appendix

PROC IML PROGRAM FOR OBTAINING CHI-SQUARE PLOT

```
TITLE CHI-SQUARE PLOT FOR MULTIVARIATE NORMALITY TEST;
OPTIONS NOCENTER;
DATA EXCELL;
 INPUT MKTBOOK ROTC ROE REASS EBITASS EXCELL;
KEEP EBITASS ROTC;
CARDS;
insert data here
PROC IML;
USE EXCELL;
READ ALL INTO X;
N=NROW(X); * N CONTAINS THE NUMBER OF OBSERVATIONS;
ONE=J(N,1,1); * NX1 VECTOR CONTAINING ONES;
DF=N-1;
MEAN=(ONE`*X)/N; * MATRIX OF MEANS;
XM=X-ONE*MEAN;  * XM CONTAINS THE MEAN-CORRECTED DATA;
SSCPM=XM`*XM;
SIGMA=SSCPM/DF;
SIGMAINV=INV(SIGMA);
MD=XM*SIGMAINV*XM`;
MD=VECDIAG(MD);
PRINT MD;
CREATE MAHAL FROM MD;
```

```
APPEND FROM MD;
QUIT;
PROC SORT;
 BY COL1;
PROC IML;
USE MAHAL;
READ ALL INTO DIST;
N=NROW(DIST);
ID=1:N;
ID=ID`;
HALF=J(N,1,.5);
PLEVEL=(ID-HALF)/N;
USE EXCELL;
READ ALL INTO X;
NC=NCOL(X);
CHISQ=CINV(PLEVEL,NC);
NEW=DIST || CHISQ;
MD={'MAHALANOBIS DISTANCE'};
CHISQ={'CHI SQUARE'};
QQ={'CHI-SQUARE PLOT'};
PRINT NEW;
CALL PGRAF(NEW,,MD,CHISQ,QQ);
CREATE TOTAL FROM NEW;
APPEND FROM NEW;
QUIT;
PROC CORR;
 VAR COL1 COL2;
```

CHAPTER 13

Canonical Correlation

Consider each of the following scenarios:

- The health department is interested in determining if there is a relationship between housing quality—measured by a number of variables such as type of housing, heating and cooling conditions, availability of running water, and kitchen and toilet facilities—and incidences of minor and serious illness, and the number of disability days.
- A medical researcher is interested in determining if individuals' lifestyles and eating habits have an effect on their health measured by a number of health-related variables such as hypertension, weight, anxiety, and tension levels.
- The marketing manager of a consumer goods firm is interested in determining if there is a relationship between types of products purchased and consumers' lifestyles and personalities.

Each of these scenarios attempts to determine if there is a relationship between two sets of variables. Canonical correlation is the appropriate technique for identifying relationships between two sets of variables. If, based on some theory, it is known that one set of variables is the predictor or independent set and another set of variables is the criterion or dependent set then the objective of canonical correlation analysis is to determine if the predictor set of variables affects the criterion set of variables. However, it is not necessary to designate the two sets of variables as the dependent and independent sets. In such cases the objective is simply to ascertain the relationship between the two sets of variables. The next section provides a geometrical view of the canonical correlation procedure.

13.1 GEOMETRY OF CANONICAL CORRELATION

Consider a hypothetical data set consisting of two predictor variables (X_1 and X_2) and two criterion variables (Y_1 and Y_2).[1] The data set given in Table 13.1 can be represented in a four-dimensional space. Since it is not possible to depict a four-dimensional space, the geometrical representation of the data is shown separately for the X and Y variables.

[1]For the rest of this chapter we will assume that based on some underlying theory one set of variables is identified as the predictor set and another set of variables is identified as the criterion set. The predictor and the criterion set of variables, respectively, will be referred to as the X and Y variables.

Table 13.1 **Hypothetical Data**

	Mean Corrected Data				New Variables	
Observation	X_1	X_2	Y_1	Y_2	V_1	W_1
1	1.051	−0.435	0.083	0.538	0.262	0.959
2	−0.419	−1.335	−1.347	−0.723	−1.513	−0.645
3	1.201	0.445	1.093	−0.112	0.989	1.260
4	0.661	0.415	0.673	−0.353	0.512	0.723
5	−1.819	−0.945	−0.817	−1.323	−1.220	−1.956
6	−0.899	0.375	−0.297	−0.433	−0.427	−0.820
7	3.001	1.495	1.723	2.418	2.446	3.215
8	−0.069	−2.625	−2.287	−1.063	−2.513	−0.524
9	−0.919	0.385	−0.547	0.808	−0.238	−0.838
10	−0.369	−0.265	−0.447	−0.543	−0.606	−0.410
11	−0.009	−0.515	0.943	−0.633	0.670	−0.098
12	0.841	1.915	1.743	1.198	2.047	1.161
13	0.781	1.845	1.043	2.048	1.680	1.089
14	0.631	−0.495	0.413	−0.543	0.202	0.535
15	−1.679	−0.615	−1.567	−0.643	−1.692	−1.760
16	−0.229	−0.525	−0.777	−0.252	−0.817	−0.317
17	−0.709	−0.975	0.523	−0.713	0.248	−0.868
18	−0.519	0.055	−0.357	0.078	−0.309	−0.502
19	0.051	0.715	0.133	0.328	0.237	0.174
20	0.221	0.245	0.403	0.238	0.460	0.260
21	−1.399	−0.645	−0.817	−1.133	−1.155	−1.490
22	0.651	0.385	1.063	−0.633	0.782	0.708
23	−0.469	−0.125	−0.557	−0.393	−0.658	−0.484
24	0.421	1.215	−0.017	1.838	0.612	0.625
Mean	0.000	0.000	0.000	0.000	0.000	0.000
SD	1.052	1.033	1.018	1.011	1.109	1.140

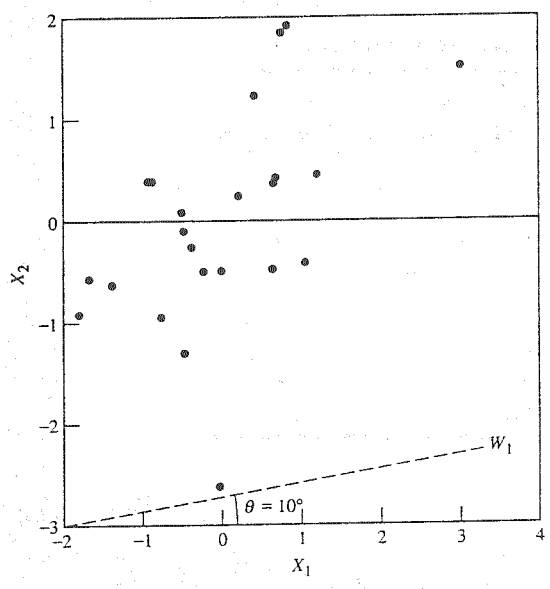
Panels I and II of Figure 13.1, respectively, give plots of the X and Y variables. Now suppose that in Panel I we identify a new axis, W_1, that makes an angle of, say, $\theta_1 = 10^\circ$ with X_1. The projection of the points onto this new axis gives a new variable that is a linear combination of the X variables. As discussed in Section 2.7 of Chapter 2, the values of the new variable can be computed from the following equation:

$$\begin{aligned} W_1 &= \cos 10^\circ \times X_1 + \sin 10^\circ \times X_2 \\ &= .985X_1 + .174X_2. \end{aligned} \tag{13.1}$$

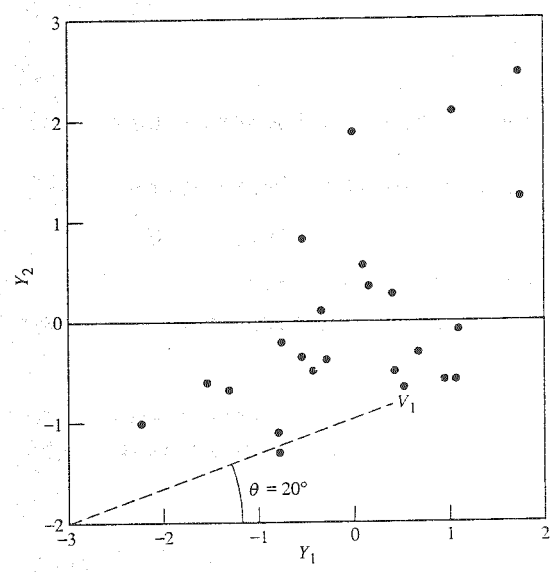
Table 13.1 gives the values of the new variable W_1. Similarly, in Panel II of Figure 13.1 we identify a new axis, V_1, that makes an angle of, say, $\theta_2 = 20^\circ$ with Y_1. The projection of the points onto Y_1 gives a new variable that is a linear combination of the Y variables. Values of this new variable can be computed from the following equation:

$$\begin{aligned} V_1 &= \cos 20^\circ Y_1 + \sin 20^\circ Y_2 \\ &= .940 \times Y_1 + .342 \times Y_2. \end{aligned} \tag{13.2}$$

Table 13.1 also gives the values of this new variable. The simple correlation between the two new variables (i.e., W_1 and V_1) is equal to .831.



Panel I



Panel II

Figure 13.1 Plot of predictor and criterion variables. Panel I, Plot of predictor variables. Panel II, Plot of criterion variables.

Table 13.2 gives the correlation between the two new variables for a few combinations of θ_1 and θ_2. It can be seen that the correlation between the two new variables is greater for some sets of new axes than for other sets of new axes. In fact, the correlation between W_1 and V_1 is highest when the angle between W_1 and X_1 (i.e., θ_1) is 57.6 degrees and the angle between V_1 and Y_1 (i.e., θ_2) is 47.2 degrees. Panels I and II, respectively, of Figure 13.2 show the two new axes, W_1 and V_1. The projections of the

Table 13.2 **Correlation between Various New Variables**

Angle between W_1 and X_1 (θ_1)	Angle between V_1 and Y_1 (θ_2)	Correlation
10	20	.830
20	10	.846
10	30	.843
40	40	.946
57.6	47.2	.961
30	10	.872
20	40	.894
40	20	.919
60	70	.937

points onto these two axes, which give the new variables, can be computed by using the following equations:

$$\begin{aligned} W_1 &= \cos 57.6^\circ \times X_1 + \sin 57.6^\circ \times X_2 \\ &= .536X_1 + .844X_2 \end{aligned} \tag{13.3}$$

$$\begin{aligned} V_1 &= \cos 47.2^\circ \times Y_1 + \sin 47.2^\circ \times Y_2 \\ &= .679Y_1 + .734Y_2. \end{aligned} \tag{13.4}$$

Table 13.3 gives the resulting new variables, and the correlation between the two variables is equal to .961.

Having identified W_1 and V_1, it is possible to identify another set of axes (i.e., W_2 and V_2) such that:

1. The correlation between the new variables, W_2 and V_2, is maximum.
2. The second set of new variables, W_2 and V_2, is uncorrelated with the previous set of new variables, W_1 and V_1.

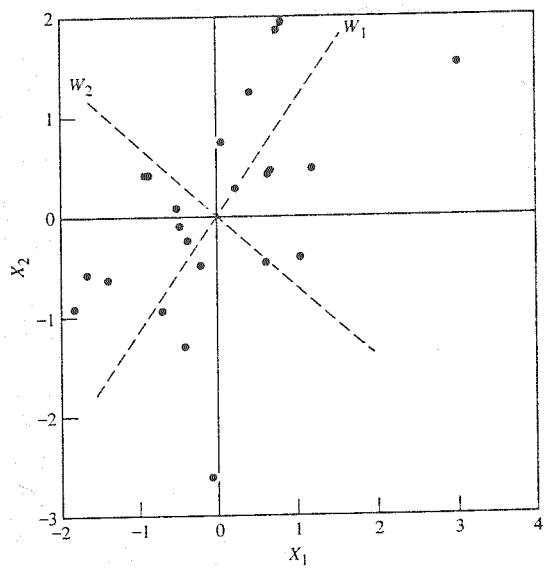
Figure 13.2 also shows the second set of new axes, W_2 and V_2, whose angles with X_1 and Y_1 are, respectively, equal to 138.33° and 135.30°. The respective equations for forming the new variables are:

$$\begin{aligned} W_2 &= \cos 138.33^\circ \times X_1 + \sin 138.33^\circ \times X_2 \\ &= -.747X_1 + .665X_2 \end{aligned} \tag{13.5}$$

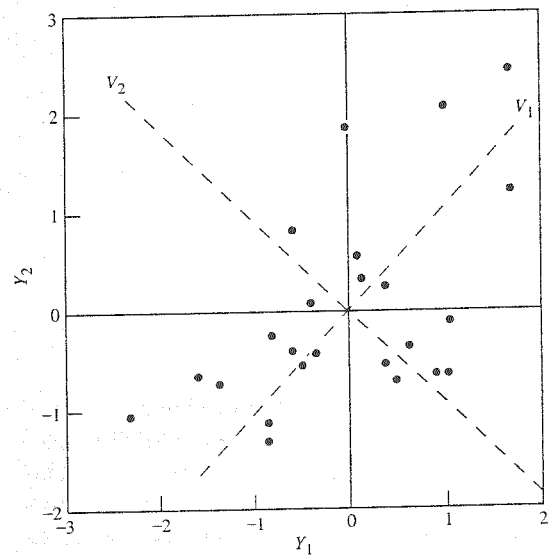
$$\begin{aligned} V_2 &= \cos 135.30^\circ \times Y_1 + \sin 135.30^\circ \times Y_2 \\ &= -.711Y_1 + .703Y_2. \end{aligned} \tag{13.6}$$

The procedure is continued until no further set of new variables can be identified. In the present case it is not possible to identify other sets of axes that meet the preceding criteria because in a given dimensional space the number of independent axes can only equal the dimensionality of the space. That is, the number of independent sets of axes, and therefore the corresponding sets of variables, can only be equal to the minimum of p and q, where p and q are, respectively, the number of X and Y variables.

In canonical correlation terminology, Eqs. 13.3 and 13.4 are the first set of canonical equations, which give the first set of new variables, W_1 and V_1. These new variables are called canonical variates. Equations 13.5 and 13.6 are the second set of canonical equa-



Panel I



Panel II

Figure 13.2 New axes for Y and X variables. Panel I, New axes for X variables. Panel II, New axes for Y variables.

tions giving the second set of canonical variates, W_2 and V_2. The correlation between each pair of canonical variates is called the canonical correlation.

To summarize, the objective of canonical correlation is to identify pairs of new axes (i.e., W_i and V_i), with each pair resulting in two new variables—where one variable is a linear combination of the X variables and the other variable is a linear combination of the Y variables—such that: (1) the correlation between W_i and V_i is maximum, and (2) each set of new variables is uncorrelated with other sets of new variables.

Table 13.3 **Variables W_1 and V_1**

	New Variables	
Observation	V_1	W_1
1	0.450	0.196
2	−1.447	−1.351
3	0.663	1.020
4	0.201	0.705
5	−1.524	−1.772
6	−0.519	−0.165
7	2.943	2.871
8	−2.337	−2.253
9	0.217	−0.167
10	−0.702	−0.421
11	0.180	−0.439
12	2.064	2.068
13	2.209	1.977
14	−0.115	−0.080
15	−1.539	−1.419
16	−0.715	−0.566
17	−0.164	−1.203
18	−0.187	−0.231
19	0.330	0.631
20	0.448	0.326
21	−1.386	−1.294
22	0.262	0.674
23	−0.667	−0.357
24	1.332	1.252
Mean	0.000	0.000
SD	1.263	1.262

Notice that the objective of canonical correlation analysis is very similar to that of conducting a principal components analysis on each set of variables. The difference is with respect to the criterion used to identify the new axes. In principal components analysis, the first new axis results in a new variable that accounts for the maximum variance in the data. In canonical correlation, on the other hand, a new axis is identified for each set of variables such that the correlation between the two resulting new variables is maximum. It is quite possible that only a few canonical variates are needed to adequately represent the association between the two sets of variables. In this sense canonical correlation analysis is also a data reduction technique. Rather than examining numerous correlations between the two sets of variables to discern the association between them, each set is first reduced to a few linear combinations and only the correlations between the few linear combinations are interpreted. For example, rather than interpreting $p \times q$ correlations between the X and the Y variables, only a few canonical correlations need to be interpreted. The number of canonical correlations that needs to be interpreted will be $m < min(p, q)$, as not all of the canonical correlations will be statistically and/or practically significant. Therefore, an additional objective of canonical correlation is to determine the minimum number of canonical correlations needed to adequately represent the association between the two sets of variables.

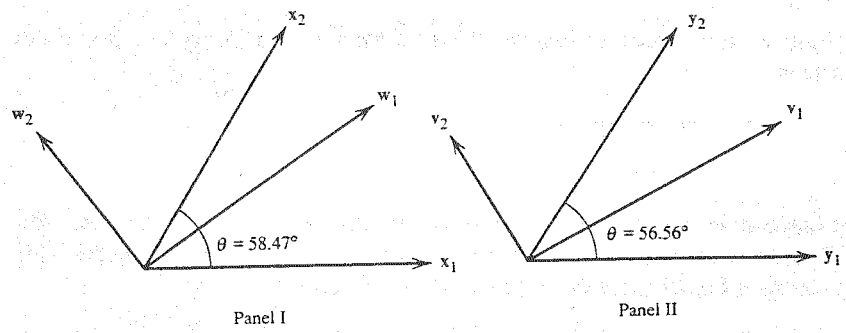


Figure 13.3 Geometrical illustration in subject space.

13.1.1 Geometrical Illustration in the Observation Space

Further insights into the objectives of canonical correlation can be obtained by using the observation space. As discussed in Chapter 3, data can also be plotted in the observation space where the observations are the dimensions and each variable is represented as a point or a vector in the given dimensional space. For the data given in Table 13.1, each variable can be represented by a vector in the 24-dimensional observation space. Vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{y}_1$, and $\mathbf{y}_2$ will lie in a four-dimensional space embedded in the 24-dimensional space. Furthermore, $\mathbf{x}_1$ and $\mathbf{x}_2$ will lie in a two-dimensional space embedded in the 24-dimensional space and $\mathbf{y}_1$ and $\mathbf{y}_2$ also will lie in a two-dimensional space embedded in the 24-dimensional space. Once again, because we cannot represent a four-dimensional space, Panel I of Figure 13.3 depicts $\mathbf{x}_1$ and $\mathbf{x}_2$, and Panel II depicts $\mathbf{y}_1$ and $\mathbf{y}_2$. In Panel I, the cosine of the angle between $\mathbf{x}_1$ and $\mathbf{x}_2$ gives the correlation between the two predictor variables and in Panel II the cosine of the angle between $\mathbf{y}_1$ and $\mathbf{y}_2$ gives the correlation between the two criterion variables. The objective of canonical correlation analysis is to identify $\mathbf{w}_1$, which lies in the same two-dimensional space as $\mathbf{x}_1$ and $\mathbf{x}_2$, and $\mathbf{v}_1$, which lies in the same two-dimensional space as $\mathbf{y}_1$ and $\mathbf{y}_2$, such that the angle between $\mathbf{w}_1$ and $\mathbf{v}_1$ is minimum. That is, the cosine of the angle between $\mathbf{w}_1$ and $\mathbf{v}_1$, which gives the correlation between the two linear combinations, is maximum. Next, another set of vectors, $\mathbf{w}_2$ and $\mathbf{v}_2$, is identified such that the angle between these two vectors is minimum. Figure 13.3 also shows the other set of vectors, $\mathbf{w}_2$ and $\mathbf{v}_2$. This procedure is continued until no additional sets of vectors can be identified.

13.2 ANALYTICAL APPROACH TO CANONICAL CORRELATION

Consider the following two equations:

$$W_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \tag{13.7}$$

$$V_1 = b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q. \tag{13.8}$$

Equation 13.7 gives the new variable W_1, which is a linear combination of the X variables, and Eq. 13.8 gives the new variable V_1, which is a linear combination of the Y variables. Let C_1 be the correlation between W_1 and V_1. The objective of canonical correlation is to estimate $a_{11}, a_{12}, \ldots, a_{1p}$ and $b_{11}, b_{12}, \ldots, b_{1q}$ such that C_1 is maximum. As mentioned earlier, Eqs. 13.7 and 13.8 are the canonical equations, W_1 and V_1 are the canonical variates, and C_1 is the canonical correlation.

Once W_1 and V_1 have been estimated, the next step is to identify another set of canonical variates

$$W_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p$$
$$V_2 = b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2q}Y_q$$

such that the correlation, C_2, between them is maximum, and W_2 and V_2 are uncorrelated with W_1 and V_1. That is, the two sets of canonical variates are uncorrelated. This procedure is continued until the mth set of canonical variates,

$$W_m = a_{m1}X_1 + a_{m2}X_2 + \cdots + a_{mp}X_p$$
$$V_m = b_{m1}Y_1 + b_{m2}Y_2 + \cdots + b_{mq}Y_q,$$

is identified such that C_m is maximum.

To summarize, the objective of canonical correlation is to identify the m sets of canonical variates, $(W_1, V_1), (W_2, V_2), \ldots, (W_m, V_m)$, such that corresponding canonical correlations, $C_1, C_2, \ldots, C_m$, are maximum, and

$$Cor(V_j, V_k) = 0 \qquad \text{for all } j \neq k$$
$$Cor(W_j, W_k) = 0 \qquad \text{for all } j \neq k$$
$$Cor(W_j, V_k) = 0 \qquad \text{for all } j \neq k.$$

This is clearly a maximization problem subject to certain constraints, the details of which are given in the Appendix.

13.3 CANONICAL CORRELATION USING SAS

The data in Table 13.1 are used to discuss the canonical correlation output obtained by the PROC CANCORR procedure in SAS. Table 13.4 gives the SAS commands. The variables following the VAR command are one set of variables and the variables following the WITH command are the other set of variables. Labels to the two sets of variables can be provided by the VNAME and the WNAME options. The label following VNAME is for the variables given in the VAR command and the label following WNAME is for variables in the WITH command. For the present data set, variables in the VAR command are labeled as Y variables and variables in the WITH command are labeled as the X variables. Exhibit 13.1 gives the SAS output. The circled numbers in the exhibit correspond to the bracketed text numbers.

Table 13.4 **SAS Commands for the Data in Table 13.1**

```
OPTIONS NOCENTER;
TITLE CANONICAL CORRELATION ON DATA IN TABLE 13-1;
DATA TABLE1;
INPUT X1 X2 Y1 Y2;
insert data here
PROC CANCORR ALL VNAME='Y variables' WNAME='X Variables';
VAR Y1 Y2;
WITH X1 X2;
```

Exhibit 13.1

Canonical correlation analysis on data in Table 13.1

Correlations Among the Original Variables

1

1a Correlations Among the Y VARIABLES

	Y1	Y2
Y1	1.0000	0.5511
Y2	0.5511	1.0000

1b Correlations Among the X VARIABLES

	X1	X2
X1	1.0000	0.5233
X2	0.5233	1.0000

1c Correlations Between the Y VARIABLES and the X VARIABLES

	X1	X2
Y1	0.7101	0.7551
Y2	0.6605	0.8094

Canonical Correlation Analysis

2

		Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
2a	1	0.961496	0.959656	0.015748	0.924475
2b	2	0.111249	.	0.205934	0.012376

Eigenvalues of INV(E)*H = CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	12.2407	12.2282	0.9990	0.9990
2	0.0125	.	0.0010	1.0000

3

Test of H0: The canonical correlations in the current row and all that follow are zero

		Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
3a	1	0.07458981	26.6151	4	40	0.0001
3b	2	0.98762376	0.2632	1	21	0.6133

Multivariate Statistics and F Approximations

4 S=2 M=-0.5 N=9

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.07458981	26.6151	4	40	0.0001
Pillai's Trace	0.93685172	9.2527	4	42	0.0001
Hotelling-Lawley Trace	12.25326382	58.2030	4	38	0.0001
Roy's Greatest Root	12.24073249	128.5277	2	21	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.
NOTE: F Statistic for Wilks' Lambda is exact.

5

Raw Canonical Coefficients for the Y VARIABLES

	V1	V2
Y1	0.5399100025	-1.045558224
Y2	0.5793206702	1.0347152249

Raw Canonical Coefficients for the X VARIABLES

	W1	W2
X1	0.4246027871	-1.031439812
X2	0.6689936658	0.9183104987

(continued)

Exhibit 13.1 (continued)

6 6a

Standardized Canonical Coefficients for the Y VARIABLES

	V1	V2
Y1	0.5499	-1.0649
Y2	0.5855	1.0457

6b

Standardized Canonical Coefficients for the X VARIABLES

	W1	W2
X1	0.4467	-1.0851
X2	0.6910	0.9485

Canonical Structure

7a

Correlations Between the Y VARIABLES and Their Canonical Variables

	V1	V2
Y1	0.8725	-0.4885
Y2	0.8885	0.4588

7b

Correlations Between the X VARIABLES and Their Canonical Variables

	W1	W2
X1	0.8083	-0.5888
X2	0.9247	0.3807

8 Correlations Between the Y VARIABLES and the Canonical Variables of the X VARIABLES

	W1	W2
Y1	0.8390	-0.0543
Y2	0.8543	0.0510

Correlations Between the X VARIABLES and the Canonical Variables of the Y VARIABLES

	V1	V2
X1	0.7771	-0.0655
X2	0.8891	0.0423

Canonical Redundancy Analysis

9a Standardized Variance of the Y VARIABLES

	Explained by Their Own Canonical Variables			Explained by The Opposite Canonical Variables	
	Proportion	Cumulative Proportion	Canonical R-Squared	Proportion	Cumulative Proportion
1	0.7754	0.7754	0.9245	0.7168	0.7168
2	0.2246	1.0000	0.0124	0.0028	0.7196

9b Standardized Variance of the X VARIABLES

	Explained by Their Own Canonical Variables			Explained by The Opposite Canonical Variables	
	Proportion	Cumulative Proportion	Canonical R-Squared	Proportion	Cumulative Proportion
1	0.7542	0.7542	0.9245	0.6972	0.6972
2	0.2458	1.0000	0.0124	0.0030	0.7003

10a Squared Multiple Correlations Between the Y VARIABLES and the First 'M' Canonical Variables of the X VARIABLES

M	1	2
Y1	0.7038	0.7068
Y2	0.7299	0.7325

10b Squared Multiple Correlations Between the X VARIABLES and the First 'M' Canonical Variables of the Y VARIABLES

M	1	2
X1	0.6039	0.6082
X2	0.7905	0.7923

13.3.1 Initial Statistics

Correlations among the Y variables [1a], among the X variables [1b], and between the X and Y variables are reported [1c]. Correlations between the X and the Y variables indicate the degree to which the variables in the two sets are correlated. In the present case, it can be seen that there is a strong and positive association between the two sets of variables. Notice that since there are a total of only 4 (i.e., 2×2) correlations, it is easy to discern the type of association between the two sets of variables. However, discerning the association between the two sets of variables by examining this correlation matrix may not be feasible for a large number of variables. For example, 120 correlation coefficients need to be interpreted if there are ten X variables and twelve Y variables. Consequently, this part of the output will be difficult to interpret for data sets with a large number of variables.

13.3.2 Canonical Variates and the Canonical Correlation

There are a maximum of two pairs of canonical variates (i.e., $m = min(2, 2)$) resulting in two canonical correlations. The first set of canonical variates is given by the following canonical equations [5]:

$$W_1 = 0.425X_1 + 0.669X_2 \tag{13.9}$$

$$V_1 = 0.540Y_1 + 0.579Y_2. \tag{13.10}$$

The coefficients of these equations are called raw canonical coefficients because they can be used to form the canonical variates from the raw data (i.e., unstandardized data). As in discriminant analysis, the coefficients of canonical equations are not unique. Only the ratios of the coefficients are unique. The coefficients of Eqs. 13.9 and 13.10 are scaled such that the resulting canonical variates are standardized to have a mean of zero and a variance of one. The equivalence of the coefficients in Eqs. 13.9 and 13.10 and the coefficients in Eqs. 13.3 and 13.4 can be established by normalizing the coefficients of Eqs. 13.9 and 13.10 such that the squares of the coefficients sum to one. The coefficients of Eq. 13.9 are normalized by dividing them by 0.793 (i.e., $\sqrt{.425^2 + .669^2}$) and that of Eq. 13.10 by 0.792 (i.e., $\sqrt{.540^2 + .579^2}$). That is,

$$W_1 = \frac{0.425}{0.793}X_1 + \frac{0.669}{0.793}X_2 = 0.536X_1 + 0.844X_2$$

and

$$V_1 = \frac{0.540}{0.792}Y_1 + \frac{0.579}{0.792}Y_2 = 0.682Y_1 + 0.731Y_2.$$

The coefficients of the above equations, within rounding errors, are the same as those of Eqs. 13.3 and 13.4. The canonical correlation between the canonical variates given by the preceding equations is equal to .961 [2a] and is the same as the maximum correlation in Table 13.2. This sample estimate of the canonical correlation is biased. The adjusted canonical correlation reported in the output is an approximate unbiased estimate of the canonical correlation [2a]. The square of the canonical correlation gives the amount of variance accounted for in V_1 by W_1 [2a].[2] The approximate standard error can be

[2]If the X variables had been designated as the dependent set of variables and the Y variables had been designated as the independent set of variables then the square of the canonical correlation would represent the amount of variance accounted for in W_1 by V_1.

used to assess the statistical significance of the canonical correlations; however, better tests for this purpose are reported later in the output, as we discuss in the next section. Consequently, this and the remaining statistics in this part of the output [2], which are not very useful for interpreting the results, are not discussed.

SAS also reports the standardized coefficients for the canonical variates [6a]. These coefficients can be used for forming canonical variates from standardized data, and, as before, the coefficients are scaled such that the resulting variates are standardized to have a variance of one.

The value of the second canonical correlation is 0.111 [2b], and it is the correlation between the canonical variates resulting from the following equations [5a, 5b]:

$$W_2 = -1.0314X_1 + 0.918X_2$$
$$V_2 = -1.0456Y_1 + 1.0347Y_2.$$

Once again, the coefficients of the above equations have been scaled such that the resulting canonical variates have a mean of zero and a variance of one. The reader can easily verify that the normalized coefficients of the preceding equations are the same as the coefficients of Eqs. 13.5 and 13.6.

13.3.3 Statistical Significance Tests for the Canonical Correlations

Before interpreting the canonical variates and the canonical correlations, one needs to determine if the canonical correlations are statistically significant. The null and alternative hypotheses for assessing the statistical significance of the canonical correlations are

$$H_o : C_1 = C_2 = \cdots = C_m = 0$$
$$H_a : C_1 \neq C_2 \neq \cdots \neq C_m \neq 0.$$

The null hypothesis, which states that all the canonical correlations are equal to zero, implies that the correlation matrix containing the correlations among the X and Y variables is equal to zero, i.e., $\mathbf{R}_{XY} = \mathbf{0}$ where $\mathbf{R}_{XY}$ is the correlation matrix containing the correlations between the X and the Y variables. A number of test statistics can be used for testing the above hypotheses. We will briefly discuss the test statistic based on Wilks' Λ. The Wilks' Λ is given by

$$\Lambda = \prod_{i=1}^{m}(1 - C_i^2), \tag{13.11}$$

which from the output is equal to [2a, 2b]:

$$\Lambda = (1 - .924)(1 - .0124)$$
$$= 0.0751.$$

Notice that this value, within rounding error, is the same as the reported value of Wilks' Λ [4], and it is also equal to the reported value of the likelihood ratio [3a]. The statistical significance of the Wilks' Λ or the likelihood ratio can be tested by computing the following test statistic:

$$B = -\left[n - 1 - \frac{1}{2}(p + q + 1)\right]\ln\Lambda, \tag{13.12}$$

which has an approximate χ^2 distribution with $p \times q$ degrees of freedom. In the present case,

$$B = -\left[24 - 1 - \frac{1}{2}(2 + 2 + 1)\right] \ln(.0751)$$
$$= 53.073$$

which has a χ^2 distribution with 4 degrees of freedom. The value of 53.073 is statistically significant at $p = .05$ and the null hypothesis is rejected. That is, all the canonical correlations are not equal to zero.

The preceding statistical test is an overall test because it tests for the statistical significance of all the canonical correlations. Rejection of the null hypothesis implies that at least the first canonical correlation is statistically significant. It is quite possible that the remaining $m - 1$ canonical correlations may not be statistically significant. The statistical significance of the second canonical correlation can be tested by computing Wilks' Λ after removing the effect of the first pair of canonical variates. In general, the significance of the rth canonical correlation can be assessed by computing the Wilks' Λ_r from the modified equation

$$\Lambda_r = \prod_{i=r}^{m}(1 - C_i^2). \tag{13.13}$$

The value of the corresponding test statistic is equal to

$$B_r = -\left[n - 1 - \frac{1}{2}(p + q + 1)\right] \ln \Lambda_r, \tag{13.14}$$

and has an approximate χ^2 with $(p - r)(q - r)$ degrees of freedom.

For example, for the second canonical correlation [2b]

$$\Lambda_2 = (1 - .0124)$$
$$= 0.988,$$

which is the same as the value for the likelihood ratio reported in the output [3b]. The corresponding test statistic is

$$B_2 = \left[24 - 1 - \frac{1}{2}(2 + 2 + 1)\right] \ln(.988)$$
$$= 0.247$$

and is approximately χ^2 distributed with 1 degree of freedom. The value of 0.247 is not greater than the critical χ^2 value at an alpha level of .05 and it is concluded that the second canonical correlation is not significant. Note that this procedure, which tests the statistical significance of each canonical correlation, is exactly the same as the statistical significance tests for each discriminant function in discriminant analysis.

The SAS output does not report the χ^2 tests for testing the statistical significance of the canonical correlations. Instead it reports the approximate F-test for assessing the significance of the Wilks' Λ and the likelihood ratios [3]. The χ^2 test was discussed because it is one of the most widely used test statistics, and because it shows the similarity between the statistical significance tests of the canonical correlations and the discriminant functions. The conclusions drawn from the F-test and the χ^2 tests are the same. For example, according to the F-test only the first canonical correlation is statistically significant at an alpha level of .05 [3a, 3b].

The preceding tests assess only the statistical significance of the canonical correlations. However, for large sample sizes even small values of canonical correlations are statistically significant. Therefore, the practical significance of the canonical correlations also needs to be assessed. The practical significance of a canonical correlation pertains to how much variance in one set of variables is accounted for by another set of variables. Assessing the practical significance of the canonical correlations is discussed in Section 13.3.5.

13.3.4 Interpretation of the Canonical Variates

Having assessed the statistical significance of the canonical correlations, the next step is to interpret the canonical variates. Typically, only those canonical variates are interpreted whose canonical correlations are statistically significant. Since the canonical variates are linear composites of the original variables, one should attempt to determine what the linear combinations of the significant canonical correlations represent. The problem is similar to that of interpreting the principal components in principal components analysis, the latent factors in factor analysis, and the discriminant functions in discriminant analysis. Standardized coefficients can be used for this purpose. The standardized coefficients are similar to the standardized regression coefficients in multiple regression. The standardized coefficient of a given variable indicates the extent to which the variable contributes to the formation of the canonical variate. In the case of the Y variables, both Y_1 and Y_2 have about an equal amount of influence in the formation of V_1 [6a], and for the X variables X_2 has a greater influence than X_1 in the formation of W_1 [6b].

Research has shown that standardized coefficients can be quite unstable for small sample sizes and in the presence of multicollinearity in the data. Consequently, many researchers also use simple correlations between the variables and the canonical variates for interpreting the canonical variates. These correlations are referred to as *loadings* or *structural correlations*. Use of loadings to interpret the canonical variates is similar to the use of loadings in interpreting the latent factors, the principal components, and the discriminant functions, respectively, in factor analysis, principal components analysis, and discriminant analysis. The loadings for the X variables suggest that both X_1 and X_2 are about equally influential in forming W_1 [7b], and the loadings for the Y variables suggest that Y_1 and Y_2 also are about equally important in forming V_1 [7a]. Since this is a hypothetical data set, we cannot assign any meaningful labels or names to the canonical variates.

The effect of the X variables on the Y variables is assessed by the signs of the standardized coefficients or the loadings. Since all the coefficients and loadings are positive, we conclude that the X variables have a positive impact on the Y variables. In some instances, due to the instability problem previously discussed, it is quite possible that the signs of the standardized coefficients and loadings for the corresponding variables may not agree. In such cases the researcher should perform external validation to assess the amount of instability in the coefficients. External validation procedures are discussed in Section 13.5.

13.3.5 Practical Significance of the Canonical Correlation

As mentioned earlier, for large sample sizes even small canonical correlations could be statistically significant. In addition, it is possible that a large canonical correlation may not imply a strong correlation between the X and the Y variables. This is because the

canonical correlation maximizes the correlation between *linear composites* of the Y and X variables, and not the amount of variance accounted for in one set of variables by the other set of variables. Stewart and Love (1968) suggest a redundancy measure (RM) to determine how much of the variance in one set of variables is accounted for by the other set of variables. Redundancy measures can be computed for each canonical correlation. Let $RM_{V_i|W_i}$ be the amount of variance in the Y variables that is accounted for by the X variables for the ith canonical correlation, C_i. As illustrated in the following, computing the RM is a two-step procedure.

First, the average amount of variance in the Y variables that is extracted or accounted for by V_i is computed, and it is equal to the average of the squared loadings of the Y variables on V_i. That is,

$$AV(Y|V_i) = \frac{\sum_{j=1}^{q} LY_{ij}^2}{q} \tag{13.15}$$

where $AV(Y|V_i)$ is the average variance in Y variables that is accounted for by the canonical variate, V_i, and LY_{ij} is the loading of the jth Y variable on the ith canonical variate. Because C_i^2 gives the shared variance between V_i and W_i, the redundancy measure is equal to the product of the average variance and the shared variance. That is,

$$RM_{V_i|W_i} = AV(Y|V_i) \times C_i^2. \tag{13.16}$$

To illustrate the computation of the redundancy measure, let us compute $RM_{V_1|W_1}$. From Eq. 13.15 [7a],

$$AV_{Y|V_1} = \frac{.8725^2 + .8885^2}{2} = .775$$

and it is the same as the reported value in the output [9a]. From Eq. 13.16

$$RM_{V_1|W_1} = .775 \times .961^2 = .716$$

and is also the same as that reported in the output [9a]. A redundancy measure of .716 suggests that for the first canonical correlation about 71.6% of the variance in the Y variables is accounted for by the X variables. This value is quite large and we conclude that the first canonical correlation has a high practical significance. Once again, one is faced with the issue of how high is "high" as there are no established guidelines for this purpose.

The total variance explained in one set of variables by the other set of variables is called the total redundancy. It has been shown that the total redundancy for the Y variables, which is the total variance explained in the Y variables by the X variables, is equal to

$$\begin{aligned} RM_{Y|X} &= \sum_{i=1}^{m} RM_{V_i|W_i} \\ &= \frac{\sum_{i=1}^{q} R_{Y_i}^2}{q}, \end{aligned} \tag{13.17}$$

where $RM_{Y|X}$ is the total redundancy of the Y variables and $R_{Y_i}^2$ is the squared multiple correlation that would be obtained by regressing the ith Y variable on the X variables. That is, the total redundancy of a given set of variables is the average of the squared multiple R^2s that would be obtained from a multiple regression analysis for each of the Y variables as the dependent variables and the X variables as the independent variables.

From the output the total redundancy for the Y variables is equal to $.7168 + .0028 = .7196$ [9a]. That is, the X variables account for 71.96% of the variance in the Y variables. However, most of this variance is accounted for by the first canonical variate. The SAS output also reports the R^2s [10a, 10b]. The last column of the matrix given in [10a] gives the R^2 that would result from regressing each of the Y variables on the X variables and the last column of [10b] gives the R^2 that would result from regressing each of the X variables on the Y variables. It can be seen that the total redundancy for the Y variables is equal to the average of the R^2s. That is [10a]:

$$\frac{.7068 + .7325}{2} = .7196.$$

13.4 ILLUSTRATIVE EXAMPLE

In this section we illustrate the use of canonical correlation analysis to determine if there is a relationship between people's demographic/socioeconomic characteristics and the sources from which they seek information regarding the nutritional content of various food items. The demographic variables are: EDUC, respondent's education level; CHILD, number of children; INCOME, household income; AGE, age of respondent; and EDSPOUSE, spouse's education. The sources of information are: books, newspapers, magazines, packages, and TV. It is hypothesized that the demographic characteristics have an effect on the sources of information. In other words, the sources of information are the criterion variables and the demographic characteristics are the predictor variables. Exhibit 13.2 gives the partial output.

From the correlation matrix, one observes that there is a positive correlation among the criterion variables and a positive correlation among the predictor variables [1]. Also, the correlations between the criterion and the predictor variables are positive [2], suggesting that the relationships between the criterion and the predictor variables are positive. That is, sources from which nutritional information is obtained are positively correlated with demographic/socioeconomic characteristics.

The value of the first canonical correlation is 0.698 [3], and the likelihood ratio test indicates that it is statistically significant at an alpha level of .05 [4]. The remaining canonical correlations are not statistically significant.[3] Hence, the correlations between the two sets of variables can be accounted for by just one pair of canonical variates. The redundancy measure of .2836 for the first canonical variate suggests that about 28% of the variance in the criterion variables is accounted for by the predictor variables [8].

The standardized canonical coefficients of the first canonical variate for the criterion variables (i.e., V_1) suggest that the variables, *Books* and *Package*, are more influential in forming the first canonical variate [5a]. The loadings, however, present a slightly different picture [6a]. All the loadings are positive and large, but the signs of the loadings for *Paper* and *TV* variables do not agree with the signs of their canonical coefficients. As mentioned earlier, this could happen as a result of multicollinearity in the data, or as a result of a small sample size. It should be noted that the loadings represent a bivariate relationship between a single variable and the canonical variate. The loadings essentially ignore the presence of other variables. Canonical coefficients, on the other hand, give the contribution of each variable in the presence of all the other variables.

[3]For presentation simplicity, the output for the nonsignificant canonical correlations have been deleted from the output.

Exhibit 13.2

Canonical correlation analysis for nutrition information study

1 Correlations Among the SOURCE OF NUTRITION INFO

	BOOKS	PAPER	MAGAZINE	PACKAGE	TV
BOOKS	1.0000	0.7514	0.7332	0.7751	0.5816
PAPER	0.7514	1.0000	0.8993	0.7485	0.7021
MAGAZINE	0.7332	0.8993	1.0000	0.7217	0.6758
PACKAGE	0.7751	0.7485	0.7217	1.0000	0.5075
TV	0.5816	0.7021	0.6758	0.5075	1.0000

Correlations Among the DEMOGRAPHIC CHARACTERISTICS

	EDUC	CHILD	INCOME	AGE	EDSPOUSE
EDUC	1.0000	0.4723	0.5388	0.3776	0.5264
CHILD	0.4723	1.0000	0.3503	0.4878	0.5169
INCOME	0.5388	0.3503	1.0000	0.3930	0.4086
AGE	0.3776	0.4878	0.3930	1.0000	0.5345
EDSPOUSE	0.5264	0.5169	0.4086	0.5345	1.0000

2 Correlations Between the SOURCE OF NUTRITION INFO
and the DEMOGRAPHICS CHARACTERISTICS

	EDUC	CHILD	INCOME	AGE	EDSPOUSE
BOOKS	0.5239	0.3177	0.5950	0.3836	0.4876
PAPER	0.4138	0.2480	0.4358	0.2098	0.3598
MAGAZINE	0.4248	0.2415	0.4311	0.2266	0.3611
PACKAGE	0.5171	0.3768	0.5143	0.3767	0.4758
TV	0.2579	0.1057	0.2890	0.1223	0.2324

3

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation	Eigenvalues of INV(E)*H = CanRsq/(1-CanRsq) Eigenvalue	Difference	Proportion	Cumulative
1	0.698132	0.682154	0.038640	0.487389	0.9508	0.9129	0.9334	0.9334
2	0.191172	.	0.072623	0.036547	0.0379	0.0130	0.0372	0.9706
3	0.155890	.	0.073546	0.024302	0.0249	0.0208	0.0245	0.9951
4	0.064335	.	0.075066	0.004139	0.0042	0.0033	0.0041	0.9991
5	0.029631	.	0.075312	0.000878	0.0009	.	0.0009	1.0000

Test of H0: The canonical correlations in the
current row and all that follow are zero

4

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.47945910	5.4431	25	621.8794	0.0001
2	0.93532711	0.7106	16	513.8861	0.7842
3	0.97080689	0.5599	9	411.4522	0.8298
4	0.99498665	0.2139	4	340	0.9307
5	0.99912199	0.1503	1	171	0.6988

Multivariate Statistics and F Approximations

S=5 M=-0.5 N=82.5

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.47945910	5.4431	25	621.8794	0.0001
Pillai's Trace	0.55325412	4.2551	25	855	0.0001

(continued)

Exhibit 13.2 (continued)

Hotelling-Lawley Trace	1.01867129	6.7395	25	827	0.0001
Roy's Greatest Root	0.95079644	32.5172	5	171	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

5a

Standardized Canonical Coefficients for the SOURCE OF NUTRITION INFO

	V1	V2
BOOKS	0.7915	-0.7713
PAPER	-0.1799	-0.5661
MAGAZINE	0.0900	-0.0124
PACKAGE	0.4211	1.5020
TV	-0.1538	-0.4075

5b

Standardized Canonical Coefficients for the DEMOGRAPHIC CHARACTERISTICS

	W1	W2
EDUC	0.3148	0.1037
CHILD	-0.0234	0.7891
INCOME	0.5268	-0.7248
AGE	0.1346	0.5116
EDSPOUSE	0.3017	-0.2529

Canonical Structure

6a

Correlations Between the SOURCE OR NUTRITION INFO and Their Canonical Variables

	V1	V2
BOOKS	0.9592	-0.2787
PAPER	0.7030	-0.3187
MAGAZINE	0.7085	-0.2785
PACKAGE	0.8868	0.2646
TV	0.4548	-0.4998

6b

Correlations Between the DEMOGRAPHIC CHARACTERISTICS and Their Canonical Variables

	W1	W2
EDUC	0.7972	0.1459
CHILD	0.5314	0.7030
INCOME	0.8644	-0.2947
AGE	0.6103	0.5156
EDSPOUSE	0.7425	0.1868

7a

Correlations Between the SOURCE OF NUTRITION INFO and the Canonical Variables of the DEMOGRAPHIC CHARACTERISTICS

	W1	W2
BOOKS	0.6697	-0.0533
PAPER	0.4908	-0.0609
MAGAZINE	0.4946	-0.0532
PACKAGE	0.6191	0.0506
TV	0.3175	-0.0955

7b

Correlations Between the DEMOGRAPHIC CHARACTERISTICS and the Canonical Variables of the SOURCE OF NUTRITION INFO

	V1	V2
EDUC	0.5565	0.0279
CHILD	0.3710	0.1344
INCOME	0.6034	-0.0563
AGE	0.4261	0.0986
EDSPOUSE	0.5184	0.0357

8 Standardized Variance of the SOURCE OF NUTRITION INFO

	Explained by				
	Their Own Canonical Variables			The Opposite Canonical Variables	
	Proportion	Cumulative Proportion	Canonical R-Squared	Proportion	Cumulative Proportion
1	0.5819	0.5819	0.4874	0.2836	0.2836
2	0.1153	0.6972	0.0365	0.0042	0.2878

Therefore, it is suggested that one use the canonical coefficients to determine the importance of each variable in forming the canonical variates, and the loadings to provide substantive meaning for the canonical variates. Consequently, the first canonical variate represents nutrition information obtained primarily from books and packages.

Similarly, based on the canonical coefficients and the loadings of the predictor variables, education, income, and education of the spouse are influential in forming the first canonical variate (i.e., W_1) for the predictor variables [5b, 6b]. W_1, therefore, represents

respondents' socioeconomic levels. The positive correlations between the criterion variables and W_1 and between the predictor variables and V_1 suggest that the people with a higher socioeconomic level (i.e., high education, income) usually get their nutrition information from books and packages [7a, 7b].

13.5 EXTERNAL VALIDITY

As mentioned earlier, the canonical coefficients can be quite unstable for small sample sizes and in the presence of multicollinearity in the data. In order to gain insights into the nature of instability, one can do a split sample or holdout analysis. In a holdout analysis the data set is randomly divided into two subsamples. Separate canonical analyses are run for each subsample, and high correlations between the respective canonical variates in the two samples provide evidence of the stability of the canonical coefficients. Alternatively, one could use the estimates of the canonical coefficients in one sample to predict the canonical variates in the holdout sample and to correlate the respective canonical variates. Once again, high correlations indicate stability of the canonical coefficients.

13.6 CANONICAL CORRELATION ANALYSIS AS A GENERAL TECHNIQUE

Most of the dependence methods are special cases of canonical correlation analysis. If the criterion set of variables contains only one variable then canonical correlation reduces to multiple regression analysis as there is only one dependent variable and multiple independent variables. ANOVA and two-group discriminant analysis are special cases of multiple regression; therefore, these two techniques are also special cases of canonical correlation analysis. When the criterion and predictor set of variables contain a single variable, then canonical correlation reduces to a simple correlation between the two variables. MANOVA and multiple-group discriminant analysis are also special cases of canonical correlation analysis. When the criterion variables are dummy variables representing multiple groups, then canonical correlation analysis reduces to multiple-group discriminant analysis. Finally, when the predictor variables are dummy variables representing the groups formed by the various factors, then canonical correlation analysis reduces to MANOVA. In fact, SPSS does not have a separate procedure for canonical correlation analysis. Rather, one has to use MANOVA for canonical correlation analysis.

13.7 SUMMARY

In this chapter we discussed the use of canonical correlation analysis to analyze the relationship between two sets of variables. In canonical correlation analysis, linear composites of each set of variables are formed such that the correlation between the linear composites is maximum. The linear composites are called the canonical variates and the correlation between them is called the canonical correlation. The number of linear composites that can be formed is the minimum of p and q where p is the number of variables in one set and q is the number of variables in the other set. Each set of canonical variates is uncorrelated with other sets of canonical variates. The next chapter discusses covariance structural models, which essentially test the relationships among unobservable constructs or variables.

QUESTIONS

13.1 Given the following correlation matrices compute the canonical correlation.

$$R_{xx} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$R_{yy} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$R_{xy} = \begin{pmatrix} 0 & 0 \\ 0.7 & 0 \end{pmatrix}.$$

What conclusions can you draw from the analysis?

13.2 Given the following correlation matrices, compute the canonical correlation.

$$R_{xx} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$R_{yy} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$R_{xy} = \begin{pmatrix} 0.4 & 0.6 \\ 0.5 & 0.7 \end{pmatrix}.$$

What conclusions can you draw from the analysis?

13.3 Table 8.1 gives the financial data for most-admired and least-admired firms. Develop dummy coding for the grouping variable and analyze the data using canonical correlation. Interpret the results and compare them with the results reported in Exhibit 11.1.

13.4 Table 11.5 gives the data for a drug effectiveness study. Develop dummy coding for the treatment variable and analyze the data using canonical correlation. Interpret the results and compare them to the results reported in Exhibit 11.3.

13.5 Sparks and Tucker (1971) conducted a study to determine the relationship between product use and personality. The correlations among the product use and personality traits is given in Table Q13.1 and the results of canonical correlation analysis are presented in Table Q13.2. What conclusions can you draw from these two tables?

13.6 Etgar (1976) conducted a study to determine the relationship between the power insurance companies have over agents' decisions, and the insurance companies' sources of power. The power of insurers over agents was measured by three variables which were designated as the criterion variables, and the power sources were measured by four variables which were designated as predictor variables. Table Q13.3 gives the results of canonical correlation analysis. Based on the results, the author concluded that there is a strong correlation between the predictor and criterion variables. Do you agree with the conclusions? Why or why not?

13.7 Refer to Q14.4 in Chapter 14. Assume that the set of items measuring the attitudes toward importing products is the criterion set and the set of items measuring CET is the predictor set. Analyze the data using canonical correlation and interpret the results.

Table Q13.1 **Correlation Matrix: Product Use and Personality Trait**

Product	Ascendancy	Responsibility	Emotional Stability	Sociability	Cautiousness	Original Thinking	Personal Relations	Vigor
Headache remedy	.0254	−.1391	−.2104[a]	.1490	−.0073	−.0649	−.0875	−.0907
Mouthwash	.0702	−.0983	−.1308	.1125	.1501[a]	−.0242	.0443	−.1238
Men's cologne	.1473	−.1066	−.1222	.2599[a]	.1247	.0715	−.0459	.0008
Hair spray	−.0580	−.1241	−.0725	.0388	−.0824	−.0668	−.0664	−.0159
Shampoo	.1735[a]	−.1420	.0729	.1459	−.0449	.0757	.0412	.0116
Antacid remedy	.0217	−.1521[a]	−.2692[a]	.0393	−.1222	−.0974	−.1119	−.0886
Playboy	.1293	−.0218	.0787	.2621[a]	−.1038	.0650	.0169	.1185
Alcoholic beverages	.2001[a]	−.1605[a]	.0159	.1973[a]	−.2861[a]	.0041	−.1436	.0261
Brush teeth	−.1324	−.0418	.0196	−.0624	−.0663	−.1645[a]	.0329	−.1074
Fashion adoption	.2892[a]	−.1647[a]	−.0628	.3858[a]	−.0919	.0924	.0838	.0557
Complexion aids	.0065	−.0591	−.0106	.0845	−.1131	−.0826	−.0667	−.0902
Vitamin capsules	.1384	−.1197	−.1759[a]	.1288	−.0855	.0963	−.0414	.0016
Haircut	−.0587	.0616	.0655	−.0774	−.0670	−.0247	−.0394	−.0311
Cigarettes	.0869	−.1465	−.1213	.0954	−.1313	.1408	−.0376	−.0305
Coffee	−.0413	−.0265	−.1478	−.0185	.0403	−.0781	−.0683	−.0734
Chewing gum	.1645[a]	−.1035	−.1165	.2581[a]	−.1209	−.0447	.0433	−.0446
After-shave lotion	.0506	.1016	.0429	.0751	.0091	.1288	.0168	.0676

[a]Indicates correlation coefficient is significant at the .05 level.

Source: Sparks, D. L., and W. T. Tucker (1971). "A Multivariate Analysis of Personality and Product Use," *Journal of Marketing Research*, 8(August), pp. 68–69.

Table Q13.2 **Results of the Canonical Analysis**

Variables	Canonical Coefficients 1	2	3
Criterion Set (Product Use)			
Headache remedy	−.0081	−.4433	.1123
Mouthwash	−.1598	−.4538	.2809
Men's cologne	.2231	−.1935	−.2121
Hair spray	.0664	.0706	.0857
Shampoo	.3784	.1587	−.0063
Antacid remedy	−.1421	−.1746	−.3226
Playboy	.1511	.1591	.5220
Alcoholic beverages	.4639	.3098	−.1329
Brush teeth	−.1879	−.0152	.2341
Fashion adoption	.3226	−.3993	.0856
Complexion aids	−.0243	.0925	.1799
Vitamin capsules	.2870	−.0599	−.4975
Haircut	−.1698	.1855	−.0170
Cigarettes	.4065	.0551	−.2894
Coffee	−.2441	−.2453	.1330
Chewing gum	.2051	−.1320	.1342
After-shave lotion	−.0270	.3022	.0108
Predictor Set (Personality Traits)			
Ascendancy	.0182	−.0517	−.4375
Responsibility	−.5125	.0777	−.1688
Emotional stability	.4309	.6405	.4880
Sociability	.6072	−.3597	.6199
Cautiousness	−.2869	−.5959	.2438
Original thinking	.2377	.1620	−.3076
Personal relations	−.1245	−.0567	.0369
Vigor	.1681	.2592	.0481
Roots	.3671	.3000	.1711
Canonical R	.606	.548	.413
χ^2	72.7419	56.7026	29.8417
df	24	22	20
Probability	.0000	.0002	.0752

Source: Sparks, D. L., and W. T. Tucker (1971). "A Multivariate Analysis of Personality and Product Use," *Journal of Marketing Research*, 8(August), pp. 68–69.

Appendix

Let $\mathbf{Z}$ be a $(p + q) \times 1$ random vector and let $E(\mathbf{Z}) = 0$. We partition $\mathbf{Z}$ into two subvectors: a $p \times 1$ random vector $\mathbf{X}$ and a $q \times 1$ random vector $\mathbf{Y}$. For convenience, assume that $q \leq p$. The covariance matrix of $\mathbf{Z}$ can be represented as

$$\Sigma_{ZZ} = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}.$$

Now let $W = \mathbf{a}'\mathbf{X}$ be a linear combination of the $\mathbf{X}$ components and $V = \mathbf{b}'\mathbf{Y}$ be a linear combination of the $\mathbf{Y}$ components. We select $\mathbf{a}'$ and $\mathbf{b}'$ such that the variances of W and V are equal

Table Q13.3 **Indicators of Canonical Association between Measures of Insurers' Power and of Insurers' Power Sources**

	Canonical Weights	Loadings	Percentage of Squared Loadings
Criterion Variables—Measures of Insurers' Power			
1. Control over agents' business volume	.269	.085	.7
2. Control over agents' risk mix	.680	.044	40.6
3. Control over agents' choice of suppliers	.699	.773	58.7
			100.0
Predictor Variables—Insurers' Power Sources			
1. Provision of training to agents	.260	.413	39.9
2. Provision of supportive advertising	.150	.214	10.4
3. Speed of underwriting			
Homeowners'	−1.163	−.370	31.7
Automobile	.748	−.138	4.4
4. Speed of claim handling			
Homeowners'	3.179	.117	3.2
Automobile	−3.200	−.213	10.4
			100.0

Coefficient of canonical correlation .680 (significant at .05 level).

Source: Etgar, M. (1976). "Channel Domination and Countervailing Power in Distributive Channels," *Journal of Marketing Research*, 13(August), p. 259.

to one. That is,

$$E(W^2) = \mathbf{a}'E(\mathbf{XX}')\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}_{XX}\mathbf{a} = 1 \tag{A13.1}$$

$$E(V^2) = \mathbf{b}'E(\mathbf{YY}')\mathbf{b} = \mathbf{b}'\boldsymbol{\Sigma}_{YY}\mathbf{b} = 1. \tag{A13.2}$$

The correlation between W and V is equal to

$$E(WV') = \mathbf{a}'E(\mathbf{XY}')\mathbf{b} = \mathbf{a}'\boldsymbol{\Sigma}_{XY}\mathbf{b}. \tag{A13.3}$$

The objective of canonical correlation analysis is to estimate $\mathbf{a}'$ and $\mathbf{b}'$ such that the correlation between the linear combinations W and V is maximum (i.e., maximize Eq. A13.3) subject to constraints given by Eqs. A13.1 and A13.2. This is clearly a constrained maximization problem that can be solved by using Lagrangian multipliers. Let

$$\psi = \mathbf{a}'\boldsymbol{\Sigma}_{XY}\mathbf{b} - 0.5\lambda_1(\mathbf{a}'\boldsymbol{\Sigma}_{XX}\mathbf{a} - 1) - 0.5\lambda_2(\mathbf{b}'\boldsymbol{\Sigma}_{YY}\mathbf{b} - 1). \tag{A13.4}$$

Differentiating Eq. A13.4 with respect to $\mathbf{a}$ and $\mathbf{b}$ results in the following equations:

$$\frac{\partial\psi}{\partial\mathbf{a}} = \boldsymbol{\Sigma}_{XY}\mathbf{b} - \lambda_1\boldsymbol{\Sigma}_{XX}\mathbf{a} = \mathbf{0} \tag{A13.5}$$

$$\frac{\partial\psi}{\partial\mathbf{b}} = \boldsymbol{\Sigma}'_{XY}\mathbf{a} - \lambda_2\boldsymbol{\Sigma}_{YY}\mathbf{b} = \mathbf{0}, \tag{A13.6}$$

or

$$\boldsymbol{\Sigma}_{XY}\mathbf{b} - \lambda_1\boldsymbol{\Sigma}_{XX}\mathbf{a} = \mathbf{0} \tag{A13.7}$$

$$\boldsymbol{\Sigma}'_{XY}\mathbf{a} - \lambda_2\boldsymbol{\Sigma}_{YY}\mathbf{b} = \mathbf{0}. \tag{A13.8}$$

Multiplying Eq. A13.7 by $\mathbf{a}'$ gives

$$\mathbf{a}'\Sigma_{XY}\mathbf{b} - \lambda_1\mathbf{a}'\Sigma_{XX}\mathbf{a} = 0$$
$$\mathbf{a}'\Sigma_{XY}\mathbf{b} = \lambda_1$$

as $\mathbf{a}'\Sigma_{XX}\mathbf{a} = 1$. Similarly, multiplying Eq. A13.8 by $\mathbf{b}'$ gives

$$\mathbf{b}'\Sigma_{XY}'\mathbf{a} - \lambda_2\mathbf{b}'\Sigma_{YY}\mathbf{b} = 0$$
$$\mathbf{b}'\Sigma_{XY}'\mathbf{a} = \lambda_2$$

as $\mathbf{b}'\Sigma_{YY}\mathbf{b} = 1$. It can be seen from these equations that $\lambda_1 = \lambda_2$ as $\Sigma_{XY}' = \Sigma_{YX}$. Letting $\lambda_1 = \lambda_2 = \rho$, Eqs. A13.7 and A13.8 can be rewritten as

$$-\rho\Sigma_{XX}\mathbf{a} + \Sigma_{XY}\mathbf{b} = 0$$
$$\Sigma_{YX}\mathbf{a} - \rho\Sigma_{YY}\mathbf{b} = 0.$$

These equations can be written in matrix form as

$$\begin{pmatrix} -\rho\Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & -\rho\Sigma_{YY} \end{pmatrix}\begin{pmatrix} \mathbf{a} \\ \mathbf{b} \end{pmatrix} = 0. \tag{A13.9}$$

For a nontrivial solution to Eq. A13.9,

$$\begin{vmatrix} -\rho\Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & -\rho\Sigma_{YY} \end{vmatrix} = 0. \tag{A13.10}$$

The determinant of Eq. A13.10 is a polynomial of order $p + q$ whose q nonzero roots are $\pm\rho_1, \pm\rho_2, \ldots, \pm\rho_q$ and it has $p - q$ zero roots. The first root, ρ_1, gives the correlation between W and V. The solution of Eq. A13.9 for the first root will give $\mathbf{a}'$ and $\mathbf{b}'$.

The solution can also be obtained as follows. Multiply Eq. A13.7 by $\Sigma_{YX}\Sigma_{XX}^{-1}$ and add ρ times Eq. A13.8 to get

$$\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\mathbf{b} - \rho\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XX}\mathbf{a} + \rho\Sigma_{YX}\mathbf{a} - \rho^2\Sigma_{YY}\mathbf{b} = 0,$$

or

$$\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}\mathbf{b} - \rho^2\Sigma_{YY}\mathbf{b} = 0.$$

Multiplying the preceding equation by Σ_{YY}^{-1} we get

$$(\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \rho^2\mathbf{I})\mathbf{b} = 0. \tag{A13.11}$$

Similarly, one gets

$$(\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \rho^2\mathbf{I})\mathbf{a} = 0. \tag{A13.12}$$

For nontrivial solutions of these equations the following must be true:

$$|\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY} - \rho^2\mathbf{I}| = 0 \tag{A13.13}$$

and

$$|\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} - \rho^2\mathbf{I}| = 0. \tag{A13.14}$$

That is, the solution for obtaining $\mathbf{a}'$ reduces to finding the eigenstructure of the following $p \times p$ matrix:

$$\Sigma_{XX}^{-1}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}, \tag{A13.15}$$

The solution for obtaining $\mathbf{b}'$ reduces to finding the eigenstructure of the following $q \times q$ matrix:

$$\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}. \tag{A13.16}$$

The eigenvalues of Expressions A13.15 and A13.16 are the same. The first eigenvalue gives the squared canonical correlation between W and V, the corresponding eigenvector of Expression A13.15 gives $\mathbf{a}'$, and the corresponding eigenvector of Expression A13.16 gives $\mathbf{b}'$. Vectors $\mathbf{a}'$ and $\mathbf{b}'$ are then normalized such that the variances of W and V are equal to one.

It is possible to identify another set of linear combinations (i.e., W_2 and V_2), that are uncorrelated with the previous linear combinations, such that the correlation between these two linear combinations is maximum. The second eigenvalue, ρ_2, gives the squared canonical correlation between W_2 and V_2, the corresponding eigenvector of Expression A13.15 gives $\mathbf{a}_2$, and the corresponding eigenvector of Expression A13.16 gives $\mathbf{b}_2$. Vectors $\mathbf{a}_2$ and $\mathbf{b}_2$ are then normalized such that the variances of W_2 and V_2 are equal to one. In general, the rth root (i.e., ρ_r) gives the correlation between W_r and V_r and the corresponding eigenvectors of Expressions A13.15 and A13.16 give $\mathbf{a}_r'$ and $\mathbf{b}_r'$, respectively.

A13.1 EFFECT OF CHANGE IN SCALE

The canonical correlation is invariant to a change in the scale of the $\mathbf{X}$ and the $\mathbf{Y}$ vectors. For example, let $\mathbf{X}^* = \mathbf{C}_X\mathbf{X}$ and $\mathbf{Y}^* = \mathbf{C}_Y\mathbf{Y}$ where $\mathbf{C}_X$ and $\mathbf{C}_Y$ are nonsingular matrices. The resulting covariance matrices are: $\boldsymbol{\Sigma}_{X^*X^*} = \mathbf{C}_X\boldsymbol{\Sigma}_{XX}\mathbf{C}_X'$; $\boldsymbol{\Sigma}_{X^*Y^*} = \mathbf{C}_X\boldsymbol{\Sigma}_{XY}\mathbf{C}_Y'$; and $\boldsymbol{\Sigma}_{Y^*Y^*} = \mathbf{C}_Y\boldsymbol{\Sigma}_{YY}\mathbf{C}_Y'$. Substituting these transformations in Eq. A13.10 we get

$$\begin{vmatrix} -\rho\mathbf{C}_X\boldsymbol{\Sigma}_{XX}\mathbf{C}_X' & \mathbf{C}_X\boldsymbol{\Sigma}_{XY}\mathbf{C}_Y' \\ \mathbf{C}_Y\boldsymbol{\Sigma}_{YX}\mathbf{C}_X' & -\rho\mathbf{C}_Y\boldsymbol{\Sigma}_{YY}\mathbf{C}_Y' \end{vmatrix} = 0,$$

or

$$\begin{vmatrix} \mathbf{C}_X & 0 \\ 0 & \mathbf{C}_Y \end{vmatrix} \cdot \begin{vmatrix} -\rho\boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{XY} \\ \boldsymbol{\Sigma}_{YX} & -\rho\boldsymbol{\Sigma}_{YY} \end{vmatrix} \cdot \begin{vmatrix} \mathbf{C}_X' & 0 \\ 0 & \mathbf{C}_Y' \end{vmatrix} = 0. \tag{A13.17}$$

The roots (i.e., eigenvalues) of this equation are the same as the roots of Eq. A13.10.

A13.2 ILLUSTRATIVE EXAMPLE

We use the data in Table 13.1 to illustrate the computations of the canonical correlation and the canonical variates. Covariance matrices are used to illustrate that canonical correlation analysis is scale invariant. The covariance matrices are:

اگر X و Y تعداد شان مساوی بود ، X را مربع بگیریم

$$\boldsymbol{\Sigma}_{XX} = \begin{bmatrix} 1.0372 & .5675 \\ 0.5675 & 1.0221 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_{YY} = \begin{bmatrix} 1.1068 & .5686 \\ 0.5686 & 1.0668 \end{bmatrix},$$

and

$$\boldsymbol{\Sigma}_{YX} = \begin{bmatrix} 0.7608 & 0.7943 \\ 0.7025 & 0.8452 \end{bmatrix}.$$

Expression A13.15 is equal to

$$\begin{bmatrix} 0.3417 & 0.3699 \\ 0.5189 & 0.5951 \end{bmatrix}, \tag{A13.18}$$

and the determinantal equation is equal to

مجهول ، هم تواشتراست

$$\begin{vmatrix} 0.3417 - \rho & 0.3699 \\ 0.5189 & 0.5951 - \rho \end{vmatrix} = 0.$$

Simplifying the preceding equation we get

$$(0.3417 - \rho)(0.5951 - \rho) - .3699 \times 0.5189 = 0$$
$$\rho^2 - .9368\rho + .0114 = 0.$$

Using the quadratic formula

$$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a},$$

the roots are

$$\rho_1 = \frac{.9368 + \sqrt{.9368^2 - 4 \times .0014}}{2} = .9245,$$

and

$$\rho_2 = \frac{.9368 - \sqrt{.9368^2 - 4 \times .0014}}{2} = .0123,$$

این ρ برای b هم یکی است ؟

which are the same as the squared canonical correlations reported in Exhibit 13.1 [2a, 2b]. Substituting the value of ρ_1 we get

$$\begin{pmatrix} 0.3417 - .9245 & 0.3699 \\ 0.5189 & 0.5951 - .9245 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = 0.$$

Solving the preceding equation gives:

$$-.5828a_1 + .3699a_2 = 0$$
$$.5189a_1 - .3294a_2 = 0.$$

These two equations reduce to:

$$a_1 = .6347a_2$$
$$a_1 = .6347a_2,$$

which clearly shows that only the ratio of a_1 and a_2 is unique. The solution can be obtained by arbitrarily assuming that $a_1^2 + a_2^2 = 1$, which gives $a_1 = .5358$ and $a_2 = .8443$. Note that the values of a_1 and a_2 are the same as the coefficients of Eq. A13.3. The variance of the resulting linear combination is

$$(.5358 \quad .8443)\,\Sigma_{XX} \begin{pmatrix} .5358 \\ .8443 \end{pmatrix},$$

which is equal to 1.5926. The weights a_1 and a_2 are rescaled by dividing by $\sqrt{1.5926} = 1.2620$ so that the variance of the linear combination will be 1. The rescaled weights are

$$a_1 = \frac{0.5358}{1.262} = .4246,$$

and

$$a_2 = \frac{0.8443}{1.262} = .6690,$$

which are the same as the raw coefficients reported in Exhibit 13.1 [5]. Weights for the second canonical variate are obtained by repeating the procedure for the second root, ρ_2. Estimates for vector $\mathbf{b}'$ can be similarly obtained by using Expression A13.16.

The preceding computations can easily be performed on a computer to obtain the eigenvalue and eigenvector of the nonsymmetric matrix given by Expression A13.15. However, the matrix given by Expression A13.15 is nonsymmetric and most of the computer routines (e.g., PROC

Table A13.1 **PROC IML Commands for Canonical Correlation Analysis**

```
OPTIONS NOCENTER;
TITLE Canonical Correlation Analysis For Data in Table 13.1;
PROC IML;

RXX={1.0372  0.5675,
     0.5675  1.0221};
RYY={1.1068 0.5686,
     0.5686 1.06680};
RYX={0.76080 0.79430,
     0.70250 0.84520};
RXY=RYX';

*Compute the inverse of the square root of a matrix;

CALL EIGEN(EV,EVEC,RXX);
RXXHIN=INV(EVEC*SQRT(DIAG(EV))*EVEC');
RXXH=INV(RXXHIN);
CALL EIGEN(EV,EVEC,RYY);
RYYHIN=INV(EVEC*SQRT(DIAG(EV))*EVEC');

*Compute the equivalent matrices;

WMATX=RXXHIN*RXY*INV(RYY)*RYX*RYYHIN;
WMATX(|2,1|)=WMATX(|1,2|);
WMATY=RYYHIN*RYX*INV(RXX)*RXY*RXXHIN;
WMATY(|2,1|)=WMATY(|1,2|);

*Compute the eigenvectors and eigenvalues of the equivalent;
*symmetric matrix;

CALL EIGEN(EVALX,EVECX,WMATX);
CALL EIGEN(EVALY,EVECY,WMATY);

*Compute the coefficients for the canonical equations;

COEFFX=(EVECX'*RXXHIN)';
COEFFY=(EVECY'*RYYHIN)';
PRINT EVALX,EVALY,COEFFX,COEFFY;
```

IML in SAS) for obtaining the eigenstructure require a symmetric matrix. It can be shown that the eigenvalues of Expression A13.15 are the same as that of the following symmetric matrix:

$$\Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}$$

and $\mathbf{a}_r' = \mathbf{e}_{r(X)}'\Sigma_{XX}^{-1/2}$ where $\mathbf{e}_{r(X)}'$ is the rth eigenvector of the preceding expression. Table A13.1 gives the PROC IML commands for estimating the canonical correlations and the corresponding weight vectors. Exhibit A13.1 gives the resulting output. It can be seen that the eigenvalues [1, 2], within rounding errors, are the same as those computed above and the reported squared canonical correlations in Exhibit 13.1 [2a, 2b]. The eigenvalues reported in Exhibit A13.1 [1, 2] should be the same and any differences are due to rounding errors resulting from multiple matrix inversions. Also, the eigenvectors [3, 4], which give the weights for forming the canonical variates, are, within rounding errors, similar to the ones computed above and the ones reported in Exhibit 13.1 [6a, 6b].

Exhibit A13.1

PROC IML output for canonical correlation analysis

(1)
```
EVALX
0.9093504
0.0099880
```

(2)
```
EVALY
0.9358732
0.0167108
```

(3)
```
COEFFX
0.4327126  1.0943471
0.6795788 -0.971322
```

(4)
```
COEFFY
0.5229898  0.9852245
0.5764159 -0.979072
```

CHAPTER 14

Covariance Structure Models

Consider an existing theory that hypothesizes relationships among three constructs: attitude (ATT), behavioral intentions (BI), and actual behavior (B). Figure 14.1 gives the relationship among these three constructs, which suggests that ATT affects BI which in turn affects B. Also, B is directly affected by ATT. The model given in Figure 14.1 is referred to as a structural or a path model.[1] Suppose one is interested in testing whether or not data support the hypothesized model given in Figure 14.1. This chapter discusses the available statistical techniques for estimating and testing structural models. We begin by first giving a brief introduction to structural models.

14.1 STRUCTURAL MODELS

The structural or path model depicted in Figure 14.1 can be represented by the following equations:[2]

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1 \tag{14.1}$$

$$\eta_2 = \gamma_{21}\xi_1 + \beta_{21}\eta_1 + \zeta_2. \tag{14.2}$$

These are referred to as path or structural equations, as they depict the structural relationships among the constructs. Equation 14.1 gives the relationship between ATT (ξ_1) and BI (η_1), and Eq. 14.2 gives the relationship between B (η_2), BI (η_1), and ATT (ξ_1). ATT is referred to as the exogenous variable or construct because it is not affected by any other construct in the model. All exogenous variables are represented by ξ. Constructs BI and B are known as endogenous constructs because they are affected by other constructs, which may be a combination of exogenous and/or other endogenous constructs. All endogenous constructs are represented by η. The arrows between the constructs represent how the constructs are related to each other and are known as structural paths. The structural paths are quantified by the structural coefficients. Structural coefficients between endogenous and exogenous constructs are represented by γ and those among endogenous constructs are represented by β. The first subscript of the structural coefficient refers to the construct that is being affected and the second refers to the causal construct. The amount of unexplained relationship in each equation is referred to as error in equation and is denoted by ζ. For example, β_{21} is the link

[1]The terms *structural* and *path model* are used interchangeably to represent relationships among the constructs.

[2]To be consistent with standard textbooks on structural models, we use Greek letters to represent the relationships among the constructs.

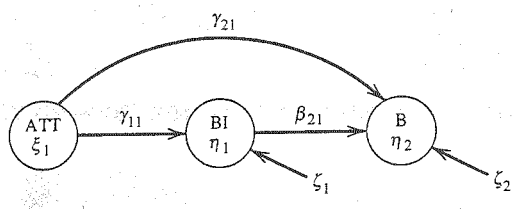


Figure 14.1 Structural or path model.

between the cause BI (η_1), and the effect B (η_2). Equations 14.1 and 14.2 are depicted graphically in Figure 14.1.

In a structural model there can be two types of constructs, unobservable and observable. Unobservable constructs, such as attitudes, intelligence, frustration, innovativeness, and consumer ethnocentric tendencies cannot be observed and therefore cannot be measured directly. On the other hand, constructs such as gender, age, and income are observable, and in principle can be measured without error. We first discuss structural models that contain observable constructs measured without error. Structural models with unobservable constructs are discussed in Section 14.3.

14.2 STRUCTURAL MODELS WITH OBSERVABLE CONSTRUCTS

Assume that ATT (ξ_1), BI (η_1), and B (η_2) can be measured without error (i.e., they are "observable" constructs). Let x_1, y_1, and y_2, respectively, be perfect measures for the observable constructs ξ_1, η_1, and η_2. Equations 14.1 and 14.2 can then be written as

$$y_1 = \gamma_{11}x_1 + \zeta_1 \tag{14.3}$$

$$y_2 = \gamma_{21}x_1 + \beta_{21}y_1 + \zeta_2. \tag{14.4}$$

Figure 14.2 gives the structural model represented by these equations. The problem reduces to estimating the parameters of the model given by Eqs. 14.3 and 14.4 (i.e., γ_{11}, γ_{21}, β_{21}, $V(x_1)$, $V(\zeta_1)$, and $V(\zeta_2)$).

14.2.1 Implied Matrix

Just as in confirmatory factor models, the elements of the covariance matrix, $\boldsymbol{\Sigma}$, can be written as functions of model parameters. It can be shown (see the Appendix) that:

$$V(y_1) = \gamma_{11}^2\phi_{11} + \psi_{11}$$
$$V(y_2) = \gamma_{21}^2\phi_{11} + \beta_{21}^2(\gamma_{11}^2\phi_{11} + \psi_{11}) + 2\gamma_{21}\beta_{21}\gamma_{11}\phi_{11} + \psi_{22}$$
$$V(x_1) = \phi_{11}$$

با کمک ماتریس (و رجوع به ضمیمه) می‌توان به طور ذهنی به توان ۲ رسانده و محاسبه کرد

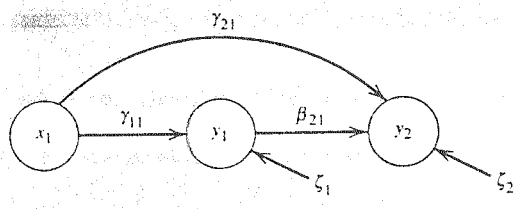


Figure 14.2 Structural model with observable constructs.

$$
\begin{aligned}
Cov(y_2, y_1) &= \gamma_{21}\gamma_{11}\phi_{11} + \beta_{21}(\gamma_{11}^2\phi_{11} + \psi_{11}) \\
Cov(x_1, y_1) &= \gamma_{11}\phi_{11} \\
Cov(x_1, y_2) &= \gamma_{21}\phi_{11} + \beta_{21}\gamma_{11}\phi_{11},
\end{aligned} \tag{14.5}
$$

where ψ_{11} and ψ_{22} are, respectively, $V(\zeta_1)$ and $V(\zeta_2)$ and others are as defined before. It can be seen that there are six equations (one for each element of Σ) and six parameters that need to be estimated (the parameters are: γ_{11}, γ_{21}, β_{21}, ϕ_{11}, ψ_{11}, and ψ_{22}). Models in which the number of parameters to be estimated is equal to the number of equations are known as *saturated models*. Obviously, saturated models have zero degrees of freedom and will result in a perfect overall fit of the model to the data. However, a perfect overall fit does not imply that all the variance of an endogenous construct is explained by the exogenous and/or other endogenous constructs. Further discussion of this issue is provided in Section 14.2.3. The parameters in Eqs. 14.3 and 14.4 can be estimated using the ordinary least squares (OLS) option of the SYSLIN procedure in SAS.[3] Alternatively, one can use the maximum likelihood estimation procedure in LISREL. The following section shows how LISREL can be used to estimate the parameters of the structural model.

14.2.2 Representing Structural Equations as LISREL Models

The structural model given by Eqs. 14.3 and 14.4 can be represented in matrix form as

$$
\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \gamma_{11} \\ \gamma_{21} \end{pmatrix} x_1 + \begin{pmatrix} 0 & 0 \\ \beta_{21} & 0 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \tag{14.6}
$$

or

$$
\mathbf{y} = \mathbf{\Gamma x} + \mathbf{B y} + \boldsymbol{\zeta}. \tag{14.7}
$$

Table 14.1 gives the LISREL notation for the parameter matrices of the structural model represented by Eqs. 14.6 or 14.7. In the table, PHI gives the covariance matrix of the exogenous constructs, the BE matrix contains the structural links among the endogenous constructs, the GA matrix contains the structural coefficients between the exogenous and the endogenous constructs, and the PSI matrix is the covariance matrix of errors in equations.

Table 14.1 **Representation of Parameter Matrices of the Structural Model in LISREL**

Parameter Matrix	LISREL Notation	Order
$\mathbf{\Phi}$	PHI	NX × NX
$\mathbf{B}$	BE	NY × NY
$\mathbf{\Gamma}$	GA	NY × NX
$\mathbf{\Psi}$	PSI	NY × NY

Note: $\mathbf{\Phi}$ is the covariance matrix for ξ (i.e., the exogenous constructs); $\mathbf{B}$ is the matrix of β coefficients; $\mathbf{\Gamma}$ is the matrix of γ coefficients; and $\mathbf{\Psi}$ is the covariance matrix of ζ (i.e., errors in equation).

[3]Two-stage least squares can be used if there are feedback loops or paths in the model and three-stage least squares can be used if the errors in the equation are correlated.

14.2.3 An Empirical Illustration

Table 14.2 gives a hypothetical covariance matrix for the model given in Figure 14.2. The covariance matrix was computed by assuming known values for model parameters. This obviously results in a perfect overall fit of the model.[4] Table 14.3 gives the commands for the LISREL procedure in SPSS. The commands before the LISREL command are standard SPSS commands for reading covariance matrices.

In the MODEL command NY specifies the number of endogenous and NX the number of exogenous variables. BE = FU specifies that the BE matrix is a full matrix and GA = FU specifies that the GA matrix is a full matrix. PSI = DI indicates that the PSI (i.e., errors in the equations) matrix is a diagonal matrix. Elements to be fixed or freed are specified as usual by the PA command. Formats for other LISREL commands in this table are described in Chapter 6. Exhibit 14.1 gives the partial output of the LISREL program.

LISREL Estimates (Maximum Likelihood)

This part of the output gives unstandardized maximum likelihood estimates of the model parameters. From the parameter estimates of the coefficients, the structural equations

Table 14.2 **Hypothetical Covariance Matrix for the Model Given in Figure 14.2**

	y_1	y_2	x_1
y_1	16.000	8.960	6.400
y_2	8.960	16.000	4.160
x_1	6.400	4.160	4.000

Table 14.3 **LISREL Commands for the Model Given in Figure 14.2**

```
TITLE LISREL IN SPSS
MATRIX DATA VARIABLES=Y1 Y2 X1 / CONTENTS N COV
BEGIN DATA
insert data here
END DATA
LISREL
 /"TITLE STRUCTURAL MODEL WITH NO MEASUREMENT ERRORS"
 /DATA NI=3 NO=200 MA=CM
 /MODEL NY=2 NX=1 BE=FU GA=FU PSI=DI
 /PA BE
 /0 0
 /1 0
 /PA GA
 /1
 /1
 /ST 0.5 ALL
 /OU ALL TO
FINISH
```

[4]A perfect overall fit would also result because the number of parameters to be estimated is equal to the number of equations. That is, the model is saturated.

Exhibit 14.1

LISREL output for the covariance matrix given in Table 14.2

```
1TITLE STRUCTURAL MODEL WITH NO MEASUREMENT ERRORS
0LISREL ESTIMATES (MAXIMUM LIKELIHOOD)

                                  (1a)                                  (1b)
0    BETA                                   GAMMA        COVARIANCE MATRIX OF Y AND X
0              Y1      Y2                     X1                Y1      Y2      X1
+            -----   ------               --------          ------  ------  ------
   Y1        0.000   0.000          Y1    1.600      Y1    16.000
   Y2        0.400   0.000          Y2    0.400      Y2     8.960  16.000
                                                     X1     6.400   4.160   4.000

(1c)0    PSI
0              Y1          Y2
+         --------    --------
             5.760      10.752

(1d)0    SQUARED MULTIPLE CORRELATIONS FOR STRUCTURAL EQUATIONS
0              Y1          Y2
+         --------    --------
             0.640       0.328

(1e)0    TOTAL COEFFICIENT OF DETERMINATION FOR STRUCTURAL EQUATIONS  IS  0.648

(2)0      CHI-SQUARE WITH    0 DEGREES OF FREEDOM =      0.00 (P = 1.00)
0                   GOODNESS OF FIT INDEX =1.000
                 ROOT MEAN SQUARE RESIDUAL =      0.000

(3)0     FITTED RESIDUALS
0              Y1          Y2          X1
+         --------    --------    --------
   Y1       0.000
   Y2       0.000       0.000
   X1       0.000       0.000       0.000

(4)-T-VALUES
0    BETA                                   GAMMA          PSI
0              Y1          Y2                   X1          Y1      Y2
+         --------    --------             -------       ------  ------
   Y1       0.000       0.000          Y1   18.762        9.950   9.950
   Y2       4.120       0.000          Y2    2.060

(5)-TOTAL AND INDIRECT EFFECTS

(5a)0    TOTAL EFFECTS OF X ON Y      STANDARD ERRORS FOR TOTAL EFFECTS OF X ON Y
0              X1                                  X1
          --------                             --------
   Y1       1.600                        Y1      0.085
   Y2       1.040                        Y2      0.121

(5b)0    INDIRECT EFFECTS OF X ON Y   STANDARD ERRORS FOR INDIRECT EFFECTS OF X ON Y
0              X1                                  X1
+         --------                             ------
   Y1       0.000                        Y1      0.000
   Y2       0.640                        Y2      0.159
```

(continued)

Exhibit 14.1 (continued)

```
5c 0    TOTAL EFFECTS OF Y ON Y        STANDARD ERRORS FOR TOTAL EFFECTS OF Y ON Y
   0              Y1         Y2                          Y1         Y2
   +        --------   --------                     --------   --------
      Y1       0.000      0.000            Y1        0.000      0.000
      Y2       0.400      0.000            Y2        0.097      0.000

5d 0    INDIRECT EFFECTS OF Y ON Y     STANDARD ERRORS FOR INDIRECT EFFECTS OF Y ON Y
   0              Y1         Y2                          Y1         Y2
   +        --------   --------                     --------   --------
      Y1       0.000      0.000            Y1        0.000      0.000
      Y2       0.000      0.000            Y2        0.000      0.000

6 -STANDARDIZED SOLUTION
   0    BETA                                GAMMA
   0              Y1         Y2                          X1
   +        --------   --------                     --------
      Y1       0.000      0.000          Y1          0.800
      Y2       0.400      0.000          Y2          0.200
```

can be written as [1a]

$$y_1 = 1.6x_1 \tag{14.8}$$

$$y_2 = 0.4x_1 + 0.4y_1 \tag{14.9}$$

Variance of the errors in Eqs. 14.8 and 14.9 are, respectively, 5.760 and 10.752 [1c]. The amount of variance of each endogenous variable that is accounted for by the exogenous and/or other endogenous variables is given by

$$\frac{V(y_i) - \psi_{ii}}{V(y_i)}.$$

From this equation, the amount of variance of the endogenous construct y_1 that is accounted for by x_1 is equal to 0.640 (i.e., $(16-5.760)/16$) [1b, 1c], and is the same as the squared multiple correlation for the structural equation [1d]. Interpretation of squared multiple correlation (SMR) is analogous to the R^2 statistic in multiple regression analysis, as SMR represents the amount of variance accounted for in the endogenous (i.e., dependent) variable by the set of exogenous and/or other endogenous (i.e., independent) variables. Therefore, it is clear that SMR for structural equations gives R^2 or the coefficient of determination for the respective equation. That is, 64% of the variance in y_1 is explained by x_1 and 32.8% of the variance in y_2 is explained by y_1 and x_1 [1d].

The statistic, *total coefficient of determination for structural equations*, is a measure of the amount of total variance of all the endogenous variables in the system of structural equations that is explained or accounted for by the set of exogenous and/or other endogenous variables. As reported in the output, a total of 64.8% [1e] of the total variation in all the endogenous variables is accounted for by the exogenous and/or endogenous variables in the structural model. The following formula is used to obtain the total coefficient of determination

$$\frac{|\mathrm{Cov}(y)| - |\Psi|}{|\mathrm{Cov}(y)|}$$

where $\mathrm{Cov}(y)$ is the covariance matrix of the endogenous variables and can be obtained from the covariance matrix of y and x [1b] and $|.|$ is the determinant of the respective matrix.

Overall Model Fit

As discussed in Chapter 6, the χ^2 statistic examines the overall fit of the model to the data. Since there are zero degrees of freedom, the χ^2 value is zero, GFI = 1, and RMSR = 0 [2]. A perfect overall fit simply implies that, given the hypothesized model, the covariance matrix implied from its parameter estimates is equal to the sample covariance matrix, and therefore the residual matrix is equal to zero [3]. The equivalence of the implied or the fitted covariance matrix, $\hat{\Sigma}$, and the sample covariance matrix, S, has nothing to do with how much variance is explained in each equation or in the overall structural model by the various structural parameters. All it suggests is that the model fits the data. The structural paths in the model could be strong or weak. In other words, a perfectly fitting model could have weak or strong structural paths. A weak structural model implies that the paths among the constructs are weak while a strong structural model implies that the structural paths among the constructs are strong. The strength of the structural paths in the model is determined by R^2, or coefficient of determination. The higher the R^2, the stronger the structural paths among the constructs and vice versa.

t-Values

The t-values can be used to determine the statistical significance of the parameter estimates (i.e., to test if the parameter estimates are significantly different from zero). All the estimated parameters are statistically significant at the .05 level [4].

Total and Indirect Effects

The endogenous constructs in structural models are either directly or indirectly affected by other constructs. In Figure 14.2, for example, y_2 is directly affected by x_1 and y_1 *and* is indirectly affected by x_1 through y_1. The total effect of an endogenous construct is the sum of all the possible ways (i.e., direct and indirect) it is affected by other endogenous and exogenous constructs. These different effects on endogenous constructs are reported in the output [5]. Following is a discussion of these effects and how they are computed.

Consider the endogenous construct y_2. Substituting Eq. 14.8 in Eq. 14.9 we get

$$\begin{aligned} y_2 &= 0.4x_1 + 0.4 \times 1.6x_1 \\ &= 0.4x_1 + 0.64x_1. \end{aligned}$$

The first term in the above equation gives the direct effect of x_1 on y_2 and the second term gives the indirect effect of x_1 on y_2. That is, the direct effect of x_1 on y_2 is equal to 0.40 and it is equal to its respective structural coefficient (i.e., γ_{21}) [1a]. The indirect effect of x_1 on y_2 is 0.640 and it is the same as reported in the output [5b]. The total effect of x_1 on y_2 is 1.040 (i.e., 0.640 + 0.40) and it is also the same as reported in the output [5a]. [5]

The reported standard errors of the direct and indirect effects can be used to compute the t-values for testing the statistical significance of each effect [5a, 5b, 5c, and 5d]. Table 14.4 presents a summary of the various effects of the row constructs on the column constructs and their t-values. The direct effects are taken from the beta and gamma matrices [1a, 1b]. Note that in the absence of indirect effects, the total effects consist of only the direct effects, and any discrepancies between the t-values of the total and the direct effects reported in the output and the table are due to rounding errors.

[5]Computation of direct and indirect effects for nonrecursive and other complex models involves complex matrix manipulations, which are discussed in the Appendix to this chapter. However, the conceptual meaning of direct and indirect effects remains the same.

Table 14.4 **Summary of Total, Direct, and Indirect Effects**

	y_1		y_2	
Construct	Effect	t-value	Effect	t-value
x_1				
Direct	1.600	18.762	0.400	2.060
Indirect	0.000	0.000	0.640	4.025
Total	1.600	18.823	1.040	8.595
y_1				
Direct	—	—	0.400	4.120
Indirect	—	—	0.000	0.000
Total	—	—	0.400	4.124

Note: The table gives the effects of row constructs on column constructs. Values of the direct effects are taken from the β and γ matrices.

Standardized Solution

The standardized solution reported in the output is obtained by standardizing the variance of the constructs to one [6]. In the standardized solution, the structural coefficients will be between -1 and 1 and their interpretation is similar to the standardized regression coefficients in regression analysis.

Note that the model discussed so far contained only observable constructs. The estimates of the parameters for a model with observable constructs can be estimated using the SYSLIN procedure in SAS. In the present case LISREL was used for estimating the parameters. Therefore, LISREL is simply a computer program for estimating the parameters of various models. *It is not a technique.* However, LISREL is a powerful multipurpose program that facilitates the estimation of a variety of models such as the confirmatory factor models and structural models with unobservable constructs. Other computer programs with similar capabilities are CALIS in SAS and EQS in BIOMED. In almost all applications that use LISREL or similar programs for estimating the parameters of the structural models, the structural models contain unobservable constructs. Structural models with unobservable constructs are discussed in the following sections.

14.3 STRUCTURAL MODELS WITH UNOBSERVABLE CONSTRUCTS

As discussed earlier, a number of constructs used in behavioral and social sciences are unobservable. Figure 14.3 gives a structural model with unobservable or latent constructs. Note that the figure can be conceptually divided into two parts or submodels. The first part consists of the structural model depicting the relationships among the latent constructs and is the same as Figure 14.2 except that the unobservable exogenous construct is represented by ξ_1, and the unobservable endogenous constructs are represented by η_1 and η_2. As before, the structural relationships among the unobservable constructs can be represented by the following equations:

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1 \tag{14.10}$$

$$\eta_2 = \gamma_{21}\xi_1 + \beta_{21}\eta_1 + \zeta_2, \tag{14.11}$$

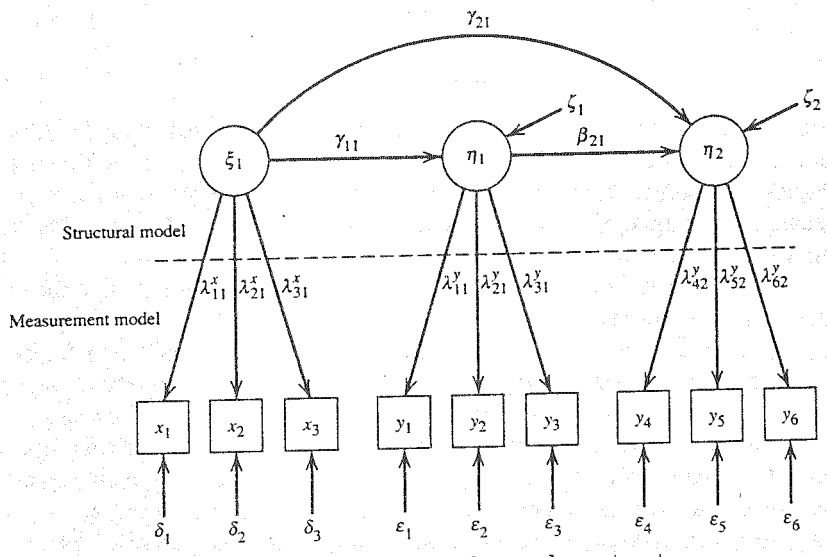


Figure 14.3 Structural model with unobserved constructs.

or, in matrix form, as

$$\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}\gamma_{11}\\ \gamma_{21}\end{pmatrix}\xi_1 + \begin{pmatrix}0 & 0\\ \beta_{21} & 0\end{pmatrix}\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} + \begin{pmatrix}\zeta_1\\ \zeta_2\end{pmatrix} \tag{14.12}$$

or

$$\eta \leftarrow \mathbf{B} = \boldsymbol{\Gamma}\boldsymbol{\xi} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta}. \tag{14.13}$$

Note that Eqs. 14.10 and 14.11 are the same as Eqs. 14.1 and 14.2. Since the constructs η_1, η_2, and ξ_1 are unobservable, we must now have indicators to measure the constructs. The second part of the model represents how the constructs are related to their indicators. This is called the measurement model and it is represented by the following equations:

$$\begin{aligned} x_1 &= \lambda^x_{11}\xi_1 + \delta_1;\ x_2 = \lambda^x_{21}\xi_1 + \delta_2;\ x_3 = \lambda^x_{31}\xi_1 + \delta_3\\ y_1 &= \lambda^y_{11}\eta_1 + \epsilon_1;\ y_2 = \lambda^y_{21}\eta_1 + \epsilon_2;\ y_3 = \lambda^y_{31}\eta_1 + \epsilon_3\\ y_4 &= \lambda^y_{42}\eta_2 + \epsilon_4;\ y_5 = \lambda^y_{52}\eta_2 + \epsilon_5;\ y_6 = \lambda^y_{62}\eta_2 + \epsilon_6. \end{aligned} \tag{14.14}$$

These equations can be represented in matrix form as

$$\begin{pmatrix}x_1\\ x_2\\ x_3\end{pmatrix} = \begin{pmatrix}\lambda^x_{11}\\ \lambda^x_{21}\\ \lambda^x_{31}\end{pmatrix}\xi_1 + \begin{pmatrix}\delta_1\\ \delta_2\\ \delta_3\end{pmatrix}, \tag{14.15}$$

or

$$\mathbf{x} = \boldsymbol{\Lambda}^x\boldsymbol{\xi} + \boldsymbol{\Theta}_\delta,$$

and

$$\begin{pmatrix}y_1\\ y_2\\ y_3\\ y_4\\ y_5\\ y_6\end{pmatrix} = \begin{pmatrix}\lambda^y_{11} & 0\\ \lambda^y_{21} & 0\\ \lambda^y_{31} & 0\\ 0 & \lambda^y_{42}\\ 0 & \lambda^y_{52}\\ 0 & \lambda^y_{62}\end{pmatrix}\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} + \begin{pmatrix}\epsilon_1\\ \epsilon_2\\ \epsilon_3\\ \epsilon_4\\ \epsilon_5\\ \epsilon_6\end{pmatrix}, \tag{14.16}$$

or

$$\mathbf{y} = \Lambda^y \boldsymbol{\eta} + \Theta_\epsilon,$$

where x represents indicators of the exogenous latent construct ξ, y represents indicators of the endogenous latent constructs ηs, λ^x represents loadings or structural coefficients between the exogenous construct and its respective indicators, λ^y represents loadings or structural coefficients among the endogenous constructs and their indicators, and δ and ϵ represent measurement errors.

Equation 14.15 gives the relationship between the latent construct ξ_1 and its indicators, x, and Eq. 14.16 gives the relationship between the latent constructs η_1 and η_2 and their indicators, y. Note that the measurement part of the model is similar to a factor model, where each factor represents a latent construct measured by its respective indicators. Therefore, the structural model with unobservable constructs can be viewed as a composite of two models: a structural or path model and a factor model. The structural model represents the relationships among the constructs, and the measurement model represents relationships among the unobservable constructs and their indicators.

Conceivably one could estimate the parameters of each model separately. That is, first use a confirmatory factor model to estimate the parameters of the measurement model and then use the estimated factor scores to estimate the parameters of the structural model. However, a more efficient procedure is to estimate parameters of both the models *jointly or simultaneously*. This can be done using LISREL or other comparable packages. Representation of the parameter matrices in LISREL is given in Table 14.5. Note that the parameter matrices are combinations of the matrices for the structural models discussed earlier and the confirmatory factor models discussed in Chapter 6.

14.3.1 Empirical Illustration

A hypothetical covariance matrix was computed for the model given in Figure 14.3 by assuming known values for the parameters. Table 14.6 gives the LISREL commands and the covariance matrix. Once again, notice that the commands are a combination of the confirmatory factor model commands discussed in Chapter 6 and the commands for

Table 14.5 **Representation of Parameter Matrices of the Structural Model with Unobservable Constructs in LISREL**

Parameter Matrix	LISREL Notation	Order
Λ_x	LX	NX × NK
Λ_y	LY	NY × NE
Θ_δ	TD	NX × NX
Θ_ϵ	TE	NY × NY
Φ	PHI	NK × NK
B	BE	NK × NK
Γ	GA	NE × NK
Ψ	PSI	NK × NK

Notes: Λ_x is the matrix of loadings of the x indicators, Λ_y is the matrix of loadings of the y indicators, Θ_δ is the covariance matrix of δ's, Θ_ϵ is the covariance matrix of ϵ's, Φ is the covariance matrix of ξ's, B is the matrix of β coefficients, Γ is the matrix of γ coefficients, and Ψ is the covariance matrix of ψ's.

Table 14.6 **LISREL Commands for Structural Model with Unobservable Constructs**

```
TITLE STRUCTURAL MODEL WITH UNOBSERVABLE CONSTRUCTS
MATRIX DATA VARIABLES=Y1,TO,Y6 X1,TO,X3 / CONTENTS=N COV
BEGIN DATA
200 200 200 200 200 200 200 200 200
4.000
3.240 4.000
3.240 3.240 4.000
1.814 1.814 1.814 4.000
1.814 1.814 1.814 3.240 4.000
1.814 1.814 1.814 3.240 3.240 4.000
2.304 2.304 2.304 1.498 1.498 1.498 4.000
2.304 2.304 2.304 1.498 1.498 1.498 2.560 4.000
2.304 2.304 2.304 1.498 1.498 1.498 2.560 2.560 4.000
END DATA
LISREL
 /TITLE "STRUCTURAL MODEL WITH UNOBSERVABLE CONSTRUCTS -- FIGURE 14.3"
 /DA NI=9 NO=200 MA=CM
 /MO NY=6 NX=3 NE=2 NK=1 BE=FU GA=FU
 /PA LX
 /0
 /1
 /1
 /PA LY
 /0 0
 /1 0
 /1 0
 /0 0
 /0 1
 /0 1
 /PA BE
 /0 0
 /1 0
 /PA GA
 /1
 /1
 /PA TD
 /1 1 1
 /PA TE
 /1 1 1 1 1 1
 /PA PHI
 /1
 /PA PSI
 /1
 /0 1
 /ST .5 ALL
 /VA 1.0 LX(1,1) LY(1,1) LY(4,2)
 /OU SE TV RS EF SS SC TO
FINISH
```

structural models with observable constructs discussed in the first part of this chapter. Note further that one indicator of each construct is fixed to one for defining the scale of the respective latent construct and for model identification purposes. Exhibit 14.2 gives the partial LISREL output.

Maximum Likelihood Estimates, Overall Fit, and Statistical Significance

The reliabilities (i.e., squared multiple correlation) [2] for each of the y indicators are high. Similar conclusions can be reached for the x indicators [3]. Using Eq. 6.17 and the completely standardized estimates [7], the reliabilities of the η_1, η_2, and ξ_1 constructs are, respectively, equal to 0.927, 0.927, and 0.842 and are quite high.

Interpretation of the structural coefficients is similar to that discussed before. Specifically, 64% of the variance in η_1 is accounted for by the exogenous construct ξ_1 and about 32.8% of the variance in η_2 is accounted for by ξ_1 and η_1 [4]. The total variance in the model that is accounted for by all the structural coefficients is equal to 64.8% [5]. The overall fit of the model is perfect, not because the model is saturated but because the sample covariance matrix was computed assuming known values for the parameters [4a]. All the parameter estimates are statistically significant at an alpha level of 0.05 [5].

Total, Direct, and Indirect Effects

The various effects in structural models with unobservable constructs can be classified as: (1) effects among the constructs; (2) effects between the exogenous constructs (i.e., ξs) and their indicators (λ^xs); (3) effects between the endogenous constructs and their indicators (λ^ys); and (4) effects between the exogenous constructs (i.e., ξs) and indicators of the endogenous (i.e., ηs) constructs. These effects are reported in the output. Note that the unstandardized effects are reported. A brief discussion of each of these effects is provided in the following section (see the Appendix for a detailed discussion of the various effects).

EFFECTS AMONG THE CONSTRUCTS. Each exogenous construct affects the endogenous construct directly and/or indirectly. Following are the various effects among the constructs for the model given in Figure 14.3:

1. The direct effect of ξ_1 on η_1 is given by γ_{11}; however, ξ_1 does not have any indirect effect on η_1. Therefore, γ_{11} gives the total effect of ξ_1 on η_1, and it is equal to 0.900 [1d, 6a].
2. The direct effect of ξ_1 on η_2 is given by γ_{21} and is equal to 0.225 [1d]. The indirect effect of ξ_1 on η_2 is through η_1 and is given by $\gamma_{11}\beta_{21}$ which equals 0.360 (i.e., $.900 \times .400$) [1c, 1d, 6c]. The total effect of ξ_1 on η_2 is, therefore, given by $\gamma_{21} + \gamma_{11}\beta_{21}$ and is equal to 0.585 (i.e., $.225 + .360$) [6a].
3. The direct effect of η_1 on η_2 is given by β_{21} and there is no indirect effect. The total effect then would be β_{21} and it is equal to 0.400 [1c, 6e].

EFFECTS OF THE EXOGENOUS CONSTRUCTS ON THEIR INDICATORS. Each of the x indicators is directly affected by ξ and the effect is given by the respective loading (i.e., λ^x). That is, the Λ_x matrix gives the direct effect of the exogenous constructs on their indicators. The x indicators (i.e., the indicators of the exogenous constructs) are never affected indirectly and, therefore, the λ^xs also give the total effects [1b].

Exhibit 14.2

LISREL output for structural model with unobservable constructs

```
(1) 0LISREL ESTIMATES (MAXIMUM LIKELIHOOD)

                    (1a)                                  (1b)
 0        LAMBDA Y                                    LAMBDA X
 0              ETA 1      ETA 2                         KSI 1
 +           --------   --------                      -------
       Y1      1.000      0.000            X1          1.000
       Y2      1.000      0.000            X2          1.000
       Y3      1.000      0.000            X3          1.000
       Y4      0.000      1.000
       Y5      0.000      1.000
       Y6      0.000      1.000
                                                      (1d)
(1c) 0        BETA                                   GAMMA
 0              ETA 1      ETA 2                         KSI 1
 +           --------   --------                      --------
    ETA 1      0.000      0.000            ETA 1        0.900
    ETA 2      0.400      0.000            ETA 2        0.225

(2) 0  SQUARED MULTIPLE CORRELATIONS FOR Y - VARIABLES
 0           Y1         Y2         Y3         Y4         Y5         Y6
 +     --------   --------   --------   --------   --------   --------
          0.810      0.810      0.810      0.810      0.810      0.810

 0  TOTAL COEFFICIENT OF DETERMINATION FOR Y - VARIABLES IS  0.993

(3) 0  SQUARED MULTIPLE CORRELATIONS FOR X - VARIABLES
 0           X1         X2         X3
 +     --------   --------   --------
          0.640      0.640      0.640

 0  TOTAL COEFFICIENT OF DETERMINATION FOR X - VARIABLES IS  0.842

(4) 0  SQUARED MULTIPLE CORRELATIONS FOR STRUCTURAL EQUATIONS
 0        ETA 1      ETA 2
 +     --------   --------
          0.640      0.328

(4a) 0  TOTAL COEFFICIENT OF DETERMINATION FOR STRUCTURAL EQUATIONS  IS  0.648

 0   CHI-SQUARE WITH  24 DEGREES OF FREEDOM =     0.00 (P = 1.00)
 0                GOODNESS OF FIT INDEX =1.000
         ADJUSTED GOODNESS OF FIT INDEX =1.000
              ROOT MEAN SQUARE RESIDUAL =     0.000

(5) -T-VALUES
 0     LAMBDA Y                     LAMBDA X          BETA
 0           ETA 1      ETA 2          KSI 1                ETA 1   ETA 2
 +        --------   --------         -------              -----   -----
    Y1      0.000      0.000     X1    0.000     ETA 1     0.000   0.000
    Y2     18.998      0.000     X2   11.662     ETA 2     3.064   0.000
    Y3     18.998      0.000     X3   11.662
    Y4      0.000      0.000
    Y5      0.000     18.706
    Y6      0.000     18.706
```

(continued)

Exhibit 14.2 (continued)

```
0        GAMMA                PHI                    PSI
0              KSI 1                  KSI 1                ETA 1   ETA 2
+            --------                 -----              ------   -----
    ETA 1     10.581          KSI 1   6.434      ETA 1   5.908
    ETA 2      1.488                              ETA 2   0.000   7.651

0      THETA EPS
0              Y1        Y2        Y3        Y4        Y5        Y6
+         --------  --------  --------  --------  --------  --------
             6.783     6.783     6.783     6.475     6.475     6.475

0      THETA DELTA
0              X1        X2        X3
+         --------  --------  --------
             7.276     7.276     7.276
```

6 -TOTAL AND INDIRECT EFFECTS

6a TOTAL EFFECTS OF KSI ON ETA

```
0              KSI 1
+            --------
    ETA 1      0.900
    ETA 2      0.585
```

6b STANDARD ERRORS FOR TOTAL EFFECTS OF KSI ON ETA

```
              KSI 1
              -------
    ETA 1     0.085
    ETA 2     0.088
```

6c INDIRECT EFFECTS OF KSI ON ETA

```
0              KSI 1
+            --------
    ETA 1      0.000
    ETA 2      0.360
```

6d STANDARD ERRORS FOR INDIRECT EFFECTS OF KSI ON ETA

```
              KSI 1
              ------
    ETA 1     0.000
    ETA 2     0.120
```

6e 0TOTAL EFFECTS OF ETA ON ETA

```
0              ETA 1      ETA 2
+            --------   --------
    ETA 1      0.000      0.000
    ETA 2      0.400      0.000
```

6f STANDARD ERRORS FOR TOTAL EFFECTS OF ETA ON ETA

```
              ETA 1      ETA 2
              --------   ------
    ETA 1     0.000      0.000
    ETA 2     0.131      0.000
```

6g TOTAL EFFECTS OF ETA ON Y

```
0              ETA 1      ETA 2
+            --------   --------
        Y1     1.000      0.000
        Y2     1.000      0.000
        Y3     1.000      0.000
        Y4     0.400      1.000
        Y5     0.400      1.000
        Y6     0.400      1.000
```

6h STANDARD ERRORS FOR TOTAL EFFECTS OF ETA ON Y

```
              ETA 1      ETA 2
            --------   --------
        Y1    0.000      0.000
        Y2    0.053      0.000
        Y3    0.053      0.000
        Y4    0.131      0.000
        Y5    0.131      0.053
        Y6    0.131      0.053
```

(continued)

Exhibit 14.2 (continued)

6i INDIRECT EFFECTS OF ETA ON Y

	ETA 1	ETA 2
Y1	0.000	0.000
Y2	0.000	0.000
Y3	0.000	0.000
Y4	0.400	0.000
Y5	0.400	0.000
Y6	0.400	0.000

6j STANDARD ERRORS FOR INDIRECT EFFECTS OF ETA ON Y

	ETA 1	ETA 2
Y1	0.000	0.000
Y2	0.000	0.000
Y3	0.000	0.000
Y4	0.131	0.000
Y5	0.131	0.000
Y6	0.131	0.000

6k TOTAL EFFECTS OF KSI ON Y

	KSI 1
Y1	0.900
Y2	0.900
Y3	0.900
Y4	0.585
Y5	0.585
Y6	0.585

6l STANDARD ERRORS FOR TOTAL EFFECTS OF KSI ON Y

	KSI 1
Y1	0.085
Y2	0.085
Y3	0.085
Y4	0.088
Y5	0.088
Y6	0.088

7 -COMPLETELY STANDARDIZED SOLUTION

LAMBDA Y

	ETA 1	ETA 2
Y1	0.900	0.000
Y2	0.900	0.000
Y3	0.900	0.000
Y4	0.000	0.900
Y5	0.000	0.900
Y6	0.000	0.900

LAMBDA X

	KSI 1
X1	0.800
X2	0.800
X3	0.800

BETA

	BETA 1	BETA 2
ETA 1	0.000	0.000
ETA 2	0.400	0.000

GAMMA

	KSI 1
ETA 1	0.800
ETA 2	0.200

CORRELATION MATRIX OF ETA AND KSI

	ETA 1	ETA 2	KSI 1
ETA 1	1.000		
ETA 2	0.560	1.000	
KSI 1	0.800	0.520	1.000

PSI

	ETA 1	ETA 2
ETA 1	0.360	
ETA 2	0.000	0.672

THETA EPS

Y1	Y2	Y3	Y4	Y5	Y6
0.190	0.190	0.190	0.190	0.190	0.190

THETA DELTA

X1	X2	X3
0.360	0.360	0.360

EFFECTS OF THE ENDOGENOUS CONSTRUCTS ON THEIR INDICATORS. Each of the y indicators is directly affected by its respective constructs. The direct effects of η_1 and η_2 on their respective indicators are given by the respective loadings. That is, the Λ_y matrix gives the direct effects of the endogenous constructs on their indicators [1a].

The indicators of η_2 are also indirectly affected by η_1 through η_2, and the indirect effect is given by the product of β_{21} and the respective λ^ys. For example, the indirect effect of η_1 on y_4 is equal to $\beta_{21}\lambda^y_{42}$ and is equal to .4 (i.e., .4 × 1.0) [6i], which is also equal to the total effect of η_1 on y_4 [6g].

Table 14.7 **Summary of the Results for Structural Model with Unobservable Constructs**

Overall Model Fit

Statistic	Value	Statistic	Value
Chi-square	0.000	*df*	0.000
GFI	1.000	AGFI	1.000
NCP	0.000	MDN	1.000
RNI	1.000	TLI	1.000
RMR	0.000		

Measurement Model Results

Constructs and Indicators	Completely Standardized Loadings	Reliabilities
η_1		0.927
y_1	0.900	0.910
y_2	0.900[a]	0.910
y_3	0.900[a]	0.910
η_2		0.927
y_4	0.900	0.910
y_5	0.900[a]	0.910
y_6	0.900[a]	0.910
ξ_1		0.842
x_1	0.800	0.640
x_2	0.800[a]	0.640
x_3	0.800[a]	0.640

Structural Model Results

Parameters	Standardized Estimate
Exogenous paths	
γ_{11}	0.800[a]
γ_{21}	0.200[a]
Endogenous paths	
β_{21}	0.400[a]
Coefficient of determination	
All structural equations	0.648
η_1	0.640
η_2	0.328

[a]Significant at $p < .01$.

EFFECTS OF EXOGENOUS CONSTRUCTS ON INDICATORS OF THE ENDOGENOUS CONSTRUCTS. The y indicators of η_1 are indirectly affected by ξ_1, and the indirect effect is given by the product of γ_{11} times the respective loading. For example, the indirect effect of ξ_1 on y_1 is given by $\gamma_{11}\lambda^y_{11}$ and is equal to 0.900 (i.e., $.90 \times 1.00$) [1a, 1d] which is also equal to the total effect of ξ_1 on y_1 [6k]. The indicators of η_2 are indirectly affected by ξ_1 through η_1 and η_2, and also through η_2. For instance, the total effect of ξ_1 on y_4 is given by

$$\gamma_{11}\beta_{21}\lambda^y_{42} + \gamma_{21}\lambda^y_{42},$$

and is equal to [1a, 1c, 1d, 6k]

$$.90 \times .40 \times 1.0 + .225 \times 1.0 = .585.$$

The first term in this expression gives the indirect effect of ξ_1 on y_4 through η_1 and η_2, and the second term gives the indirect effect of ξ_1 on y_4 through η_2.

Completely Standardized Solution

In the completely standardized solution the estimates are standardized with respect to the variances of the constructs and also with respect to the variances of the indicators [7]. In reporting the results, most researchers typically provide a summary of the results, which includes the fit statistics, and results of the measurement and structural model. Table 14.7 presents an example of such a reporting.

14.4 AN ILLUSTRATIVE EXAMPLE

In this section we present an application of structural equation modeling with unobservable constructs by discussing its application to coupon usage behavior. Shimp and Kavas (1984) postulated a model to study this facet of consumer behavior. The model is presented in Figure 14.4 and a brief discussion follows (for the time being ignore the dotted paths). The model suggests that actual coupon usage behavior (B) is affected by behavioral intentions (BI), which in turn is affected by attitude towards the act (AACT) and subjective norm (SN). AACT is the outcome of cognitive structures (AACTCOG) and SN is the outcome of normative structures (SNCOG). The cognitive and normative structures are measured by single items. Data collected from a two-state consumer panel resulted in a total sample size of 533 respondents. Exhibit 14.3 gives partial LISREL output.

14.4.1 Assessing the Overall Model Fit

The χ^2 statistic is significant; however, as discussed in Chapter 6, one typically resorts to other fit indices such as the GFI, AGFI, NCP, MDN, RNI, and TLI for assessing model fit [1]. Table 14.8 gives the values of the fit indices. Since the fit indices are less than the recommended cutoff value of 0.90, the researchers concluded that the model could be improved based on theory and the modification indices given in the output.

Model Respecification

The modification indices can be used to respecify the hypothesized model. As noted in Chapter 6, the modification index of a fixed parameter gives the approximate decrease in the χ^2 if the fixed parameter is freed (i.e., it is estimated). Examination of the

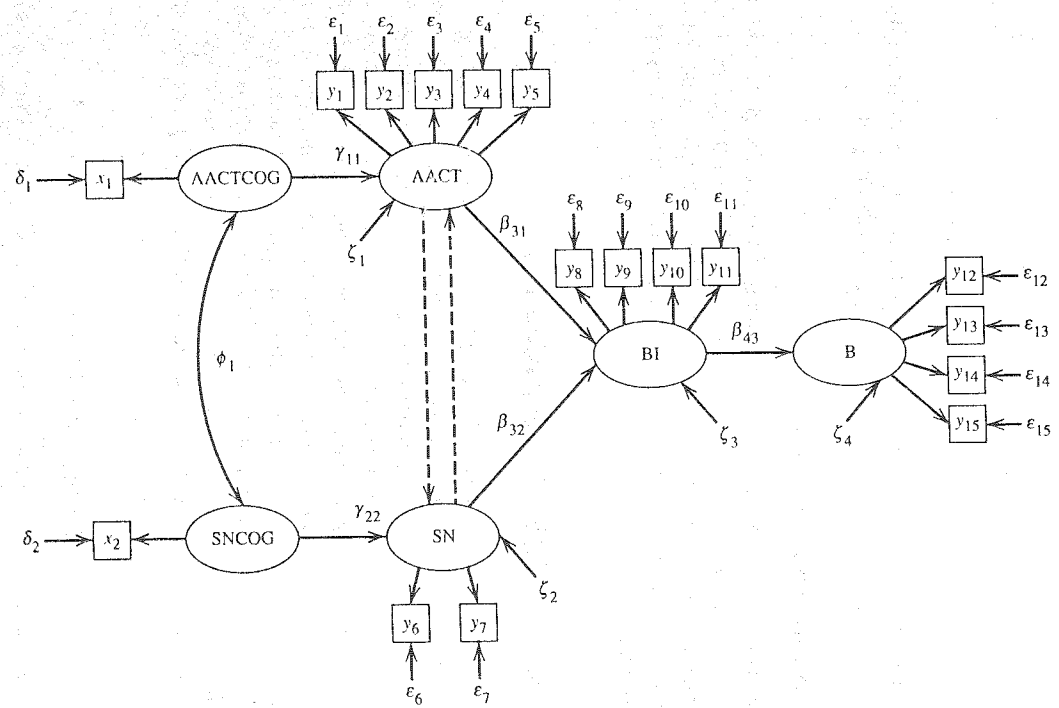


Figure 14.4 Coupon usage model. *Source:* Shimp, T. A. and A. Kavas (1984). "The Theory of Reasoned Action Applied to Coupon Usage," *Journal of Consumer Research,* 11 (December), p. 797.

Exhibit 14.3

LISREL output for coupon usage model

```
(1) 0        CHI-SQUARE WITH 115 DEGREES OF FREEDOM =   775.43 (P = .000)
    0                GOODNESS OF FIT INDEX =0.874
            ADJUSTED GOODNESS OF FIT INDEX =0.832
                 ROOT MEAN SQUARE RESIDUAL =     0.183


(2) 0        MODIFICATION INDICES FOR BETA
    0              AACT         SN         BI          B
    +          --------   --------   --------   --------
        AACT      0.000    210.298    176.803     25.541
          SN    165.862      0.000     64.462     19.978
          BI      0.000      0.000      0.000      0.517
           B      0.096      0.477      0.000      0.000

    0        MODIFICATION INDICES FOR GAMMA
    0            AACTCOG      SNCOG
    +          --------   --------
        AACT      0.000     53.053
          SN      9.059      0.000
          BI      2.742     11.192
           B      0.026      2.210

    0NO NON-ZERO MODIFICATION INDICES FOR PHI
    0NO NON-ZERO MODIFICATION INDICES FOR PSI
    0NO NON-ZERO MODIFICATION INDICES FOR THETA EPS

    0        MODIFICATION INDICES FOR THETA DELTA
    0                X1         X2
    +          --------   --------
              53.051      9.059

    0    MAXIMUM MODIFICATION INDEX IS  210.30 FOR ELEMENT ( 1, 2) OF BETA
```

modification indices suggests that inclusion of the crossover paths between AACT and SN (i.e., β_{12} and β_{21}), and a crossover path between BI and AACT (i.e., β_{13}) would improve model fit [2]. Obviously, these paths must have a theoretical support. Shimp and Kavas (1984) provide a theoretical reasoning for the inclusion of β_{12} and β_{21}. The dotted paths in Figure 14.4 represent these crossover paths. It is important that all model extensions be well grounded in theory. The analysis was rerun by freeing (i.e., estimating) the two parameters. We do not provide the LISREL output as all the relevant information can be summarized in a table. Table 14.9 gives the overall goodness-of-fit measures, the measurement model results, and the structural model results for the respecified model. The fit indices suggest a good model fit, implying that the data fit the hypothesized model.

14.4.2 Assessing the Measurement Model

All the factor loadings are quite high and statistically significant. The reliabilities of the constructs and their indicators are also acceptable. The construct reliabilities were computed using Eq. 6.17. Note that reliabilities of single-item constructs are one, as they are assumed to be measured without error. Overall, the measurement model appears to be acceptable.

Table 14.8 **Goodness-of-Fit Indices for the Coupon Usage Model**

Statistic	Value	Statistic	Value
Chi-square	775.430	*df*	115
GFI	0.874	RGFI	.896
AGFI	0.832	RAGFI	.860
NCP	1.239	MDN	0.538
RNI	0.892	TLI	0.872
RMR	0.183		

Table 14.9 **Summary of the Results for the Respecified Coupon Usage Model**

Overall Model Fit

Statistic	Value	Statistic	Value
Chi-square	446.280	*df*	113
GFI	0.914	RGFI	.936
AGFI	0.884	RAGFI	.913
NCP	0.625	MDN	0.732
RNI	0.945	TLI	0.934
RMR	0.044		

Structural Model Results

Parameters	Standardized Estimate
Exogenous paths	
γ_{11}	0.025
γ_{21}	0.387[a]
Endogenous paths	
β_{31}	0.343[a]
β_{32}	0.387[a]
β_{43}	0.696[a]
β_{21}	0.262[a]
β_{12}	0.683[a]
Coefficient of determination	
All structural equations	0.482
η_1	0.631
η_2	0.501
η_3	0.480
η_4	0.484

(continued)

Assessing the Structural Model

All the hypothesized paths are statistically significant, supporting the hypotheses related to the structural equations. The variances accounted for by structural equations range from a low of 48% to a high of 63.1% and the overall variance accounted for by the system of structural equations is 48.2%. These results suggest that all the relationships are quite strong.

Table 14.9 (continued)

Measurement Model Results

Constructs and Indicators	Completely Standardized Loadings	Reliabilities
η_1 (AACT)		0.941
y_1	0.868[a]	0.754
y_2	0.796[a]	0.634
y_3	0.875[a]	0.766
y_4	0.893[a]	0.797
y_5	0.929[a]	0.863
η_2 (SN)		0.760
y_6	0.821[a]	0.673
y_7	0.744[a]	0.554
η_3 (BI)		0.873
y_8	0.769[a]	0.591
y_9	0.784[a]	0.615
y_{10}	0.851[a]	0.725
y_{11}	0.837[a]	0.701
η_4 (B)		0.840
y_{12}	0.758[a]	0.574
y_{13}	0.793[a]	0.629
y_{14}	0.773[a]	0.597
y_{15}	0.689[a]	0.475
ξ_1 (AACTCOG)	1.000	1.000
ξ_2 (SNCOG)	1.000	1.000
ϕ	0.085[a]	

[a]Significant at $p < .01$.

An obvious question is whether the improvement in the fit of the respecified model is statistically significant. That is, are the additional parameters statistically significant? The statistical significance of the improvement in fit, and therefore the estimates of the additional parameters, can be assessed by the χ^2 difference tests for nested models. Two models are said to be nested if one of the models can be obtained by placing restrictions on the parameters of the other model. In the present case the original model is nested within the respecified model. That is, the original model can be obtained from the respecified model by constraining the parameters β_{21} and β_{12} to be equal to zero. The chi-square difference test is described below.

The difference in the χ^2 values and the degrees of freedom for the two models are equal to 329.15 (775.430 − 446.280) and 2, respectively. For nested models, the difference in the chi squares is distributed as a χ^2 statistic with degree of freedom equal to the difference in the degrees of freedom of the two models. If the χ^2 difference value is statistically significant, then the respecified model with the additional paths is assumed to have a statistically significant improvement over the original model. Since the χ^2 of 329.150 with 2 *df* is statistically significant, the respecified model has a statistically significant improvement over the previous model. Furthermore, each of the crossover paths included in the respecified model is statistically significant, suggesting that there are significant crossover effects between AACT and SN.

14.5 SUMMARY

This chapter discussed structural models. Structural or path models depict the relationship among a number of constructs. If the constructs in the structural model are measured without measurement error, then the parameters can be estimated using the standard statistical packages (e.g., SAS). This chapter, however, discussed the use of LISREL, a well-known computer program for estimating the parameters of the model. In the case when model constructs are measured with error, the resulting model is a combination of a factor model and the structural model and is typically referred to as the structural model with unobservable variables. The factor model depicts the relationship between the unobservable constructs and its measures (i.e., indicators) and the structural model presents the relationships between the unobservable constructs. Once again, the use of LISREL for estimating the parameters of the structural model with unobservable constructs is illustrated.

QUESTIONS

14.1 What are the assumptions that must be made (about observed variables, error terms, causal relationships, etc.) for an effective use of structural models?

14.2 Figure Q14.1 shows a structural model. Substitute standard notation for the letters A through D and a through w. Use the standard notation to represent the model in equation form. Classify the parameters as belonging to either the measurement part or the structural part of the model.

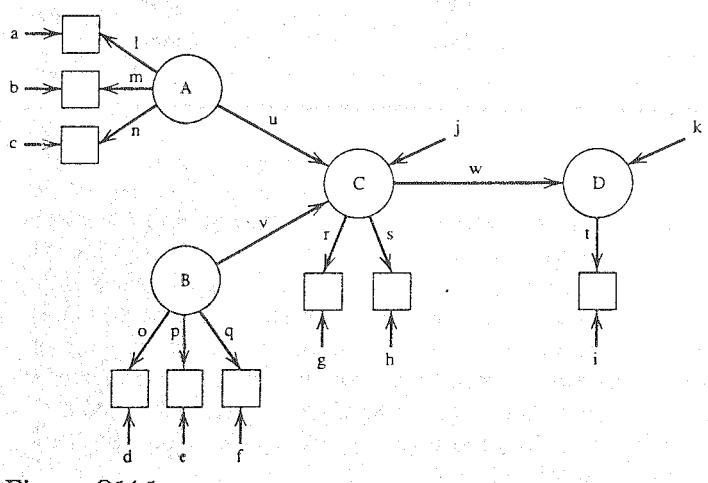


Figure Q14.1

14.3 File INTPERF.DAT gives the covariances between 12 indicators used to measure various aspects of intelligence and class performance of 347 high school juniors.

Indicators 1 to 4 are scores from tests designed to test the quantitative ability of the students. Indicators 5 to 8 are scores from tests designed to test the verbal ability of students. Indicators 9 to 12 are scores on four variations of a general intelligence test. It is believed that the performances of students on the general intelligence tests are a function of the students' quantitative and verbal abilities.

The structural model shown in Figure Q14.2 was proposed to test the above theory and determine the relative strength of the effects of quantitative and verbal abilities on

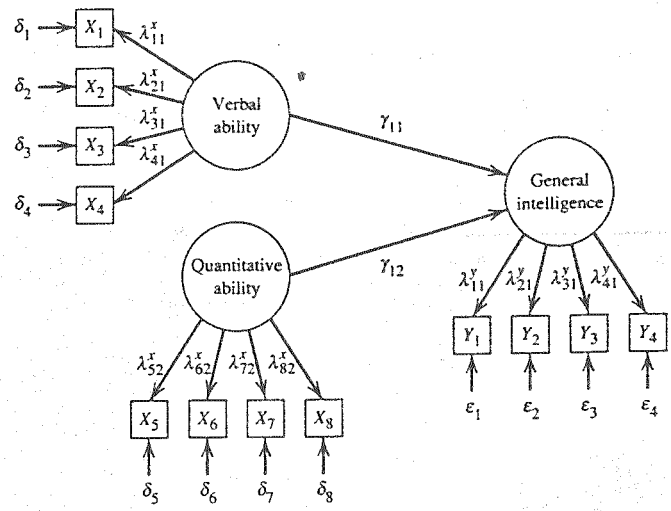


Figure Q14.2 Note that X_1 to X_8 correspond to indicators 1 to 8 and Y_1 to Y_4 correspond to indicators 9 to 12.

general intelligence levels. Estimate the parameters of the model shown in the figure and interpret the results.

How can you modify the model to improve the overall fit? Provide support for any modifications suggested by you.

14.4 In a study conducted to examine consumer ethnocentric tendencies (CET), 667 subjects were asked to indicate their attitudes toward importing products. The ethnocentric tendencies of these consumers were measured using seven indicators. File CETNEW.DAT gives the covariances among the indicators (X_1–X_7 measure ethnocentric tendencies and Y_1–Y_5 are attitudinal measures).

It is proposed that CET affect the attitudes toward importing products. Draw a structural model that represents the relationship between CET and consumer attitudes toward importing foreign products. Use the covariance data to estimate the parameters of the model. Interpret the results.

14.5 Compare the results obtained in Question 14.4 with those obtained using canonical correlation analysis in Question 13.7. What are the conceptual differences between canonical correlation analysis and structural equation modeling?

14.6 File PERFSAT.DAT gives the Pearson correlations for eight observed variables. The data came from a study on performance and satisfaction. Bagozzi (1980) formulated a structural equation model to study the relationship between performance and satisfaction in an industrial sales force. His model was designed to answer such questions as: "Is there a link between performance and job satisfaction? Does performance influence satisfaction, or does satisfaction influence performance?" Figure Q14.3 presents the path diagram for the causal model finally adopted by Bagozzi.

The latent constructs shown in the figure are as follows:

ξ_1 = achievement motivation
ξ_2 = task specific self-esteem
ξ_3 = verbal intelligence
η_1 = performance
η_2 = job satisfaction.

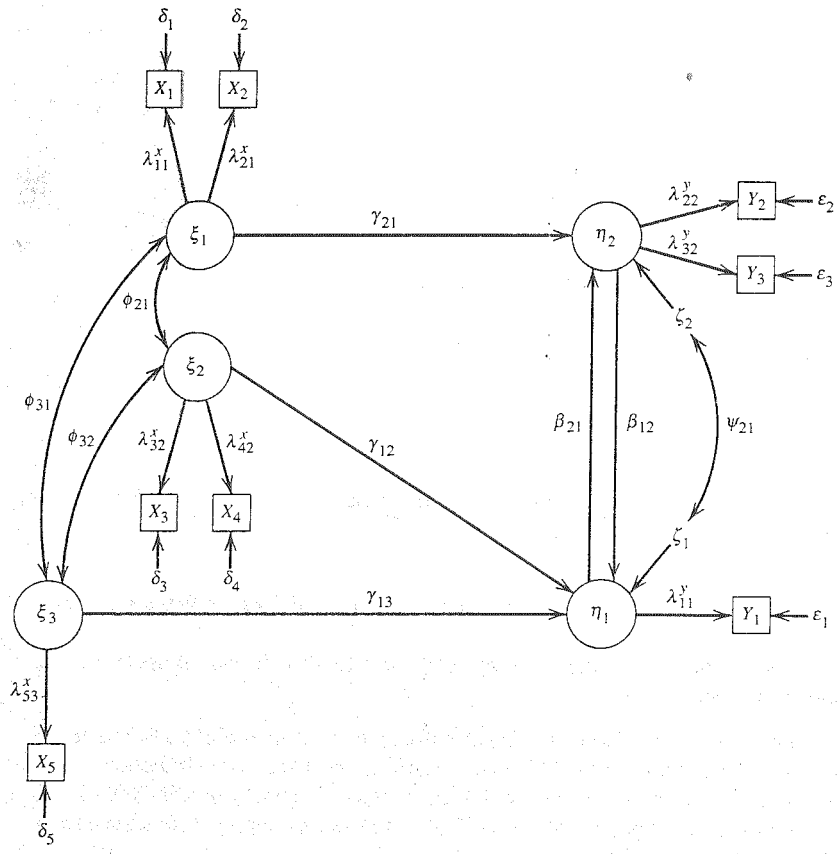


Figure Q14.3

According to the model, ξ_1 is measured by two indicators (X_1 and X_2), ξ_2 is measured by two indicators (X_3 and X_4), ξ_3 is measured by a single indicator (X_5), η_1 is measured by a single indicator (Y_1), and η_2 is measured by two indicators (Y_2 and Y_3).

Estimate the model parameters and interpret the results.

14.7 In a study designed to determine the predictors of drinking and driving behavior among 18- to 24-year-old males, the model shown in Figure Q14.4 was proposed.

The constructs shown in the figure are as follows:

ξ_1 = attitude toward drinking and driving

ξ_2 = social norms pertaining to drinking and driving

ξ_3 = perceived control over drinking and driving

η_1 = intentions to drink and drive

η_2 = drinking and driving behavior.

Attitude is measured by five indicators (X_1–X_5), social norms are measured by three indicators (X_6–X_8), perceived control is measured by four indicators (X_9–X_{12}), intentions are measured by two indicators (Y_1–Y_2), and behavior is measured using two indicators (Y_3–Y_4). File DRINKD.DAT presents the covariance matrix between the indicators (sample size = 356).

Use the covariance data to estimate the parameters of the structural model. Comment on the model fit. What modifications can you make to the model to improve the model fit? Interpret the results.

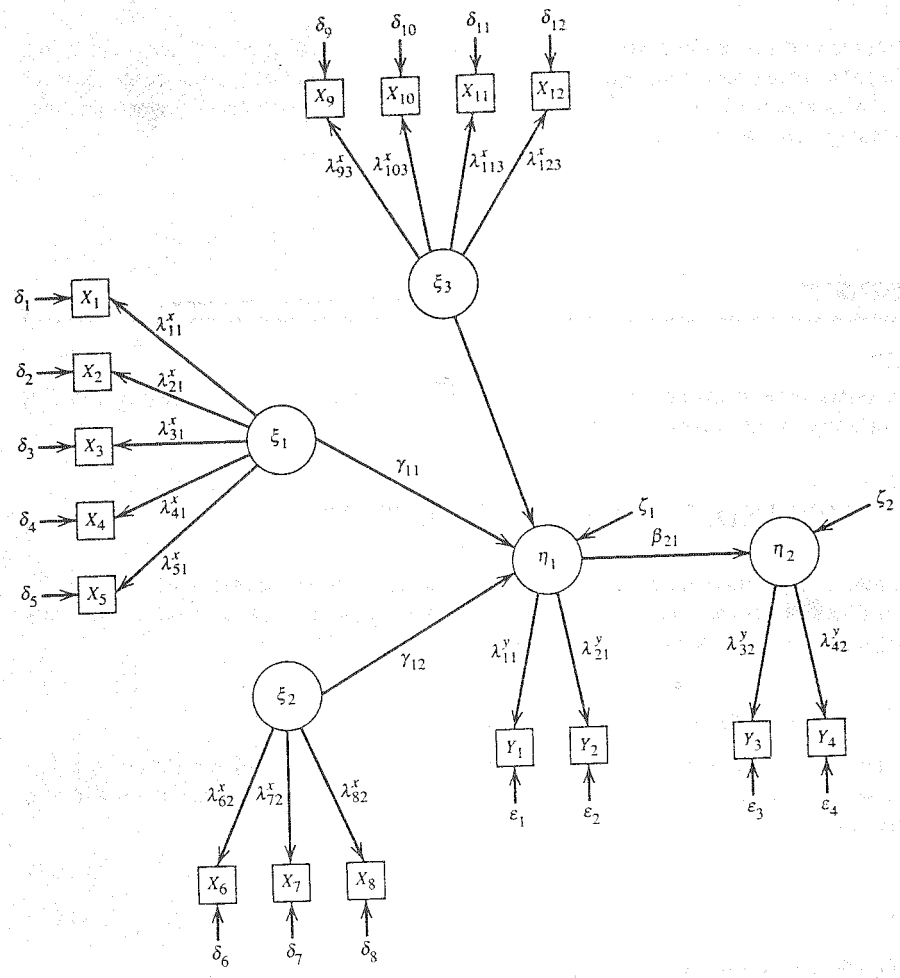


Figure Q14.4

14.8 XYZ National Bank conducted a survey of 423 customers to determine how satisfied they are with their credit cards. The bank believes that *overall satisfaction* with credit cards is a function of customer satisfaction with the following four processes: *application, billing, customer service*, and *late payment handling*. The bank also believes that *overall satisfaction* in turn determines whether the customer intends to continue using the credit card (*intent to continue*) and whether the customer will recommend the card to a friend (*recommendation*). Draw a structural model representing the relationships between the constructs, as proposed by the bank.

In its survey, the bank used four indicators to measure satisfaction with application (X_1–X_4), three indicators to measure satisfaction with billing (X_5–X_7), four indicators to measure satisfaction with customer service (X_8–X_{11}), two indicators to measure satisfaction with late payment handling (X_{12}–X_{13}), two indicators to measure overall satisfaction (Y_1–Y_2), two indicators to measure recommendation (Y_3–Y_4), and two indicators to measure intent to continue (Y_5–Y_6). File BANKSAT.DAT gives the covariance matrix between the indicators.

Use the covariance data to estimate the parameters of the structural model and interpret the results.

14.9 Assume that the study described in Question 14.4 was replicated in Korea and in the United States using a sample size of 300. File CET.DAT gives the covariance matrices for the two samples. Do a group analysis to compare the structural model of the two samples. What conclusions can you draw from your analysis?

Appendix

In this appendix we discuss the procedures for computing the implied covariance matrix and model effects (e.g., direct, indirect, and total effects).

A14.1 IMPLIED COVARIANCE MATRIX

In this section the computational procedures for obtaining the implied covariance matrix from the parameters of a given model are discussed. We first discuss models with observable constructs and then discuss models with unobservable constructs.

A14.1.1 Models with Observable Constructs

Consider the model given in Figure 14.3, which is represented by the following structural equations (also see Eqs. 14.1 and 14.2), and assume that the error terms are uncorrelated with the latent constructs.[1]

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1 \tag{A14.1}$$

$$\eta_2 = \gamma_{21}\xi_1 + \beta_{21}\eta_1 + \zeta_2. \tag{A14.2}$$

The variance of η_1 is given by

$$\begin{aligned} V(\eta_1) &= E(\eta_1^2) \\ &= E[(\gamma_{11}\xi_1 + \zeta_1)^2] \\ &= E[\gamma_{11}^2\xi_1^2 + \zeta_1^2 + 2\gamma_{11}\xi_1\zeta_1] \\ &= \gamma_{11}^2 E(\xi_1^2) + E(\zeta_1^2) + 2\gamma_{11}E(\xi_1\zeta_1) \\ &= \gamma_{11}^2\phi_{11} + \Psi_{11} + 0 \\ &= \gamma_{11}^2\phi_{11} + \Psi_{11}. \end{aligned} \tag{A14.3}$$

The covariance between ξ_1 and η_1 can be obtained by taking the expected value of Eq. A14.1 after multiplying it by ξ_1. That is,

$$\begin{aligned} Cov(\xi_1\eta_1) &= E[\gamma_{11}\xi_1^2 + \zeta_1\xi_1] \\ &= \gamma_{11}E(\xi_1^2) + E(\zeta_1\xi_1) \\ &= \gamma_{11}\phi_{11} + 0 \\ &= \gamma_{11}\phi_{11}. \end{aligned} \tag{A14.4}$$

[1]To be consistent with most textbooks, we use Greek letters to represent constructs and Roman letters to represent measures or indicators of the constructs. Therefore, η_1, η_2, and ξ_1, respectively, represent y_1, y_2, and x_1.

The variance of η_2 is given by

$$\begin{aligned}
V(\eta_2) &= E(\eta_2^2) \\
&= E[(\gamma_{21}\xi_1 + \beta_{21}\eta_1 + \zeta_2)^2] \\
&= E[\gamma_{21}^2\xi_1^2 + \beta_{21}^2\eta_1^2 + \zeta_2^2 + 2\gamma_{21}\beta_{21}\xi_1\eta_1 + 2\gamma_{21}\xi_1\zeta_2 + 2\beta_{21}\eta_1\zeta_2] \\
&= \gamma_{21}^2E(\xi_1^2) + \beta_{21}^2E(\eta_1^2) + E(\zeta_2^2) + 2\gamma_{21}\beta_{21}E(\xi_1\eta_1) + 2\gamma_{21}E(\xi_1\zeta_2) + 2\beta_{21}E(\eta_1\zeta_2) \\
&= \gamma_{21}^2\phi_{11} + \beta_{21}^2V(\eta_1) + \Psi_{22} + 2\gamma_{21}\beta_{21}Cov(\xi_1\eta_1) + 0 + 0 \\
&= \gamma_{21}^2\phi_{11} + \beta_{21}^2(\gamma_{11}^2\phi_{11} + \Psi_{11}) + 2\gamma_{21}\beta_{21}\gamma_{11}\phi_{11} + \Psi_{22}. \qquad \text{(A14.5)}
\end{aligned}$$

The covariance between η_1 and η_2 can be obtained by multiplying Eq. A14.2 by η_1 and taking its expected value

$$\begin{aligned}
Cov(\eta_1\eta_2) &= E[\gamma_{21}\xi_1\eta_1 + \beta_{21}\eta_1^2 + \zeta_2\eta_1] \\
&= \gamma_{21}E(\xi_1\eta_1) + \beta_{21}E(\eta_1^2) + E(\zeta_2\eta_1) \\
&= \gamma_{21}\gamma_{11}\phi_{11} + \beta_{21}V(\eta_1) + 0 \\
&= \gamma_{21}\gamma_{11}\phi_{11} + \beta_{21}(\gamma_{11}^2\phi_{11} + \Psi_{11}). \qquad \text{(A14.6)}
\end{aligned}$$

The covariance matrix between ξ_1 and η_2 is obtained by taking the expected value of Eq. A14.2 after multiplying it by ξ_1. That is,

$$\begin{aligned}
Cov(\xi_1\eta_2) &= E(\gamma_{21}\xi_1^2 + \beta_{21}\xi_1\eta_1 + \xi_1\zeta_2) \\
&= \gamma_{21}E(\xi_1^2) + \beta_{21}E(\xi_1\eta_1) + E(\xi_1\zeta_2) \\
&= \gamma_{21}\phi_{11} + \beta_{21}\gamma_{11}\phi_{11}. \qquad \text{(A14.7)}
\end{aligned}$$

Equations A14.3 to A14.7 give the necessary elements of the covariance matrix implied by the model parameters, and are the same as given by Eq. 14.5, except that y_1, y_2, and x_1, respectively, represent η_1, η_2, and ξ_1.

Implied Covariance Matrix Using Matrix Algebra

Equations A14.1 and A14.2 can be represented in matrix form as

$$\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} = \begin{pmatrix}\gamma_{11}\\ \gamma_{21}\end{pmatrix}\xi_1 + \begin{pmatrix}0 & 0\\ \beta_{21} & 0\end{pmatrix}\begin{pmatrix}\eta_1\\ \eta_2\end{pmatrix} + \begin{pmatrix}\zeta_1\\ \zeta_2\end{pmatrix}, \qquad \text{(A14.8)}$$

or

$$\begin{aligned}
\boldsymbol{\eta} &= \boldsymbol{\Gamma\xi} + \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\zeta} \\
\boldsymbol{\eta} - \mathbf{B}\boldsymbol{\eta} &= \boldsymbol{\Gamma\xi} + \boldsymbol{\zeta} \\
(\mathbf{I} - \mathbf{B})\boldsymbol{\eta} &= \boldsymbol{\Gamma\xi} + \boldsymbol{\zeta} \\
\boldsymbol{\eta} &= (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma\xi} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta}. \qquad \text{(A14.9)}
\end{aligned}$$

The covariance matrix, $\boldsymbol{\Sigma}_{\eta\eta}$, between the endogenous constructs (i.e., η_1 and η_2) is given by

$$\begin{aligned}
\boldsymbol{\Sigma}_{\eta\eta} &= E(\boldsymbol{\eta\eta}') \\
&= E[(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma\xi} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta}][(\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma\xi} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\zeta}]' \\
&= (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma}E(\boldsymbol{\xi\xi}')\boldsymbol{\Gamma}'(\mathbf{I} - \mathbf{B})^{-1'} + (\mathbf{I} - \mathbf{B})^{-1}E(\boldsymbol{\zeta\zeta}')(\mathbf{I} - \mathbf{B})^{-1'} \\
&= (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Gamma\Phi\Gamma}'(\mathbf{I} - \mathbf{B})^{-1'} + (\mathbf{I} - \mathbf{B})^{-1}\boldsymbol{\Psi}(\mathbf{I} - \mathbf{B})^{-1'} \\
&= (\mathbf{I} - \mathbf{B})^{-1}[\boldsymbol{\Gamma\Phi\Gamma}' + \boldsymbol{\Psi}](\mathbf{I} - \mathbf{B})^{-1'}. \qquad \text{(A14.10)}
\end{aligned}$$

The covariance matrix between the exogenous constructs, $\boldsymbol{\Sigma}_{\xi\xi}$, is given by

$$\begin{aligned}
\boldsymbol{\Sigma}_{\xi\xi'} &= E(\boldsymbol{\xi\xi}') \\
&= \boldsymbol{\Phi}. \qquad \text{(A14.11)}
\end{aligned}$$

The covariance matrix, $\Sigma_{\eta\xi}$, between exogenous and endogenous constructs is given by

$$\begin{aligned}\Sigma_{\eta\xi} &= E(\eta\xi') \\ &= E[(\mathbf{I}-\mathbf{B})^{-1}\Gamma\xi\xi' + (\mathbf{I}-\mathbf{B})^{-1}\zeta\xi'] \\ &= (\mathbf{I}-\mathbf{B})^{-1}\Gamma E(\xi\xi') + (\mathbf{I}-\mathbf{B})^{-1}E(\zeta\xi') \\ &= (\mathbf{I}-\mathbf{B})^{-1}\Gamma\Phi + \mathbf{0} \\ &= (\mathbf{I}-\mathbf{B})^{-1}\Gamma\Phi. \end{aligned} \tag{A14.12}$$

The covariance matrix, Σ, of the model is equal to

$$\Sigma = \begin{pmatrix} \Sigma_{\eta\eta} & \Sigma_{\eta\xi} \\ \Sigma_{\xi\eta} & \Sigma_{\xi\xi} \end{pmatrix}, \tag{A14.13}$$

or

$$\Sigma = \begin{pmatrix} (\mathbf{I}-\mathbf{B})^{-1}[\Gamma\Phi\Gamma' + \Psi](\mathbf{I}-\mathbf{B})^{-1'} & (\mathbf{I}-\mathbf{B})^{-1}\Gamma\Phi \\ \Gamma'(\mathbf{I}-\mathbf{B})^{-1'}\Phi & \Phi \end{pmatrix}. \tag{A14.14}$$

The preceding equations can be used to obtain the covariance matrix of any structural model with observable constructs. The following section presents an example.

An Illustrative Example

Figure A14.1 gives the structural model represented by Eqs. A14.1 and A14.2. The figure also gives a hypothetical set of parameter values. The values of these parameters were used to generate the hypothetical covariance matrix given in Table 14.2. The parameter matrices are

$$\Phi = (4); \qquad \Gamma = \begin{pmatrix} 1.60 \\ 0.40 \end{pmatrix}; \qquad \beta = \begin{pmatrix} 0 & 0 \\ 0.40 & 0 \end{pmatrix}; \qquad \Psi = \begin{pmatrix} 5.76 & 0 \\ 0 & 10.752 \end{pmatrix}.$$

The PROC IML procedure in SAS can be used for computing the covariance matrix. Exhibit A14.1 gives the resulting output. Note that the covariance matrix is the same as that given in Table 14.2.

A14.1.2 Models with Unobservable Constructs

The model given in Figure A14.2, which is the same as that given by Figure 14.3, can be represented by the following equations:

$$\begin{aligned} \eta &= \Gamma\xi + \mathbf{B}\eta + \zeta \\ \mathbf{x} &= \Lambda_x\xi + \delta \\ \mathbf{y} &= \Lambda_y\eta + \epsilon. \end{aligned} \tag{A14.15}$$

The first equation gives the structural model and the last two equations give the measurement part of the model. Since the constructs in the model are unobservable and are measured by their respective indicators, the implied covariance matrix contains covariances among the indicators of the constructs. The covariance matrix, Σ_{xx}, among the indicators of the exogenous constructs

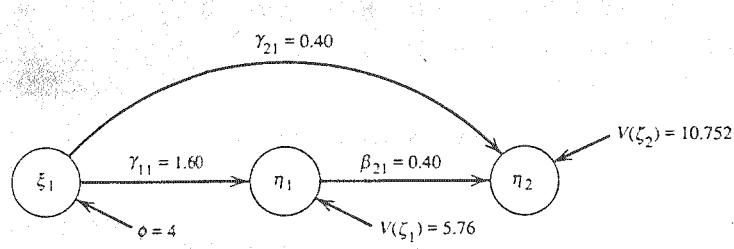


Figure A14.1 Structural model with observable constructs.

Exhibit A14.1

Covariance matrix for structural model with observable constructs

```
SAS                                    14:43 TUESDAY, DECEMBER 22, 1992    1

CYY        COL1       COL2

ROW1    16.0000     8.9600
ROW2     8.9600    16.0000


CYX        COL1

ROW1     6.4000
ROW2     4.1600


CXX        COL1

ROW1     4.0000


COV        COL1       COL2       COL3

ROW1    16.0000     8.9600     6.4000
ROW2     8.9600    16.0000     4.1600
ROW3     6.4000     4.1600     4.0000
```

Note:
CYY: Covariance matrix among the endogenous constructs.
CXX: Covariance matrix among the exogenous constructs.
CYX: Covariance matrix between the exogenous and the endogenous constructs.

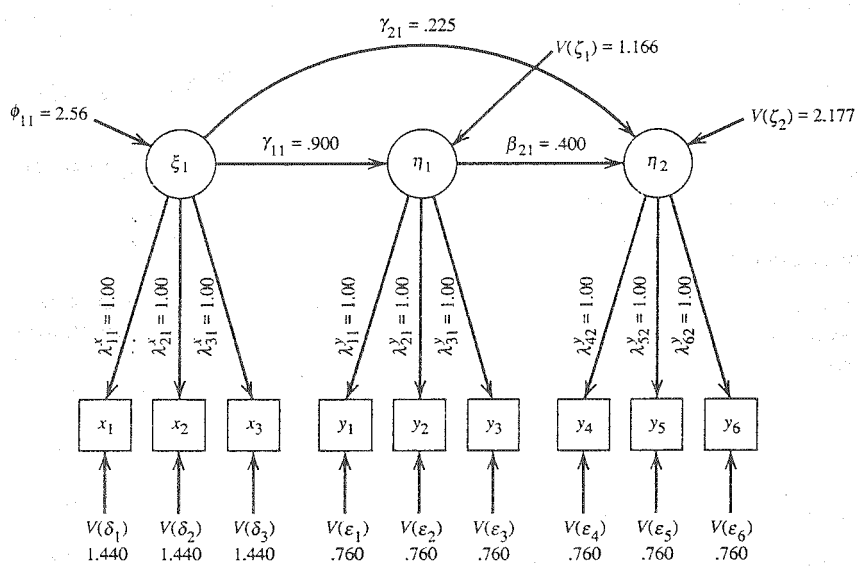


Figure A14.2 Structural model with unobservable constructs.

is equal to

$$\begin{aligned}\Sigma_{xx} = Cov(\mathbf{xx}) &= E(\mathbf{xx}') \\ &= E[(\Lambda_x\xi + \delta)(\Lambda_x\xi + \delta)'] \\ &= \Lambda_x E(\xi\xi')\Lambda_x' + \Lambda_x E(\xi\delta') + \Lambda_x E(\delta\xi) + E(\delta\delta') \\ &= \Lambda_x\Phi\Lambda_x' + 0 + 0 + \Theta_\delta \\ &= \Lambda_x\Phi\Lambda_x' + \Theta_\delta. \end{aligned} \tag{A14.16}$$

The covariance matrix, Σ_{yy}, among the indicators of the endogenous construct is given by

$$\begin{aligned}\Sigma_{yy} = Cov(\mathbf{yy}) &= E(\mathbf{yy}') \\ &= E[(\Lambda_y\eta + \epsilon)(\Lambda_y\eta + \epsilon)'] \\ &= \Lambda_y E(\eta\eta')\Lambda_y' + \Lambda_y E(\eta\epsilon') + \Lambda_y E(\epsilon\eta) + E(\epsilon\epsilon') \\ &= \Lambda_y\Sigma_{\eta\eta}\Lambda_y' + 0 + 0 + \Theta_\epsilon.\end{aligned}$$

Substituting Eq. A14.10 in this equation, we get

$$\Sigma_{yy} = \Lambda_y[(\mathbf{I} - \mathbf{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)(\mathbf{I} - \mathbf{B})^{-1'}]\Lambda_y' + \Theta_\epsilon. \tag{A14.17}$$

And the covariance matrix, Σ_{xy}, among the indicators of exogenous and endogenous constructs is given by

$$\begin{aligned}\Sigma_{xy} = Cov(\mathbf{xy}) &= E(\mathbf{xy}') \\ &= E[(\Lambda_x\xi + \delta)(\Lambda_y\eta + \epsilon)'] \\ &= \Lambda_x E(\xi\eta')\Lambda_y' + \Lambda_x E(\xi\epsilon') + E(\delta\eta')\Lambda_y' + E(\delta\epsilon') \\ &= \Lambda_x\Sigma_{\xi\eta}\Lambda_y' + 0 + 0 + 0.\end{aligned} \tag{A14.18}$$

Substituting Eq. A14.12 in this equation we get

$$\Lambda_x\Phi\Gamma'(\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y'. \tag{A14.19}$$

Therefore, the covariance matrix, Σ, for the model with unobservable constructs is equal to

$$\Sigma = \begin{pmatrix}\Sigma_{yy} & \Sigma_{yx} \\ \Sigma_{xy} & \Sigma_{xx}\end{pmatrix} \tag{A14.20}$$

or

$$\Sigma = \begin{pmatrix}\Lambda_y[(\mathbf{I} - \mathbf{B})^{-1}(\Gamma\Phi\Gamma' + \Psi)(\mathbf{I} - \mathbf{B})^{-1'}]\Lambda_y' + \Theta_\epsilon & \Lambda_y(\mathbf{I} - \mathbf{B})^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'(\mathbf{I} - \mathbf{B})^{-1'}\Lambda_y' & \Lambda_x\Phi\Lambda_x' + \Theta_\delta\end{pmatrix}. \tag{A14.21}$$

The preceding equations can be used to obtain the covariance matrix for any model given the parameter values. Following is an illustrative example.

An Illustrative Example

Consider the model given in Figure A14.2 along with a hypothetical set of parameter values. The parameter matrices for the model in Figure 14.4 that were used to generate the hypothetical covariance matrix given in Table 14.6 are

$$\Lambda_x' = (1 \quad 1 \quad 1),$$

$$\Lambda_y' = \begin{pmatrix}1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1\end{pmatrix},$$

$$\Theta_\delta' = (1.440 \quad 1.440 \quad 1.440),$$

$$\Theta_\epsilon' = (0.760 \quad 0.760 \quad 0.760 \quad 0.760 \quad 0.760 \quad 0.760),$$

$$\Phi = (2.560); \quad \Psi = \begin{pmatrix}1.166 & 0 \\ 0 & 2.177\end{pmatrix}; \quad \Gamma = \begin{pmatrix}0.90 \\ 0.225\end{pmatrix}; \quad \mathbf{B} = \begin{pmatrix}0 & 0 \\ 0.400 & 0\end{pmatrix}.$$

Exhibit A14.2

Covariance matrix for structural model with unobservable constructs

1 SAS 14:40 TUESDAY, DECEMBER 22, 1992 1

CYY	COL1	COL2	COL3	COL4	COL5	COL6
ROW1	3.9996	3.2396	3.2396	1.8142	1.8142	1.8142
ROW2	3.2396	3.9996	3.2396	1.8142	1.8142	1.8142
ROW3	3.2396	3.2396	3.9996	1.8142	1.8142	1.8142
ROW4	1.8142	1.8142	1.8142	3.9997	3.2397	3.2397
ROW5	1.8142	1.8142	1.8142	3.2397	3.9997	3.2397
ROW6	1.8142	1.8142	1.8142	3.2397	3.2397	3.9997

CXY	COL1	COL2	COL3	COL4	COL5	COL6
ROW1	2.3040	2.3040	2.3040	1.4976	1.4976	1.4976
ROW2	2.3040	2.3040	2.3040	1.4976	1.4976	1.4976
ROW3	2.3040	2.3040	2.3040	1.4976	1.4976	1.4976

CXX	COL1	COL2	COL3
ROW1	4.0000	2.5600	2.5600
ROW2	2.5600	4.0000	2.5600
ROW3	2.5600	2.5600	4.0000

COV	COL1	COL2	COL3	COL4	COL5	COL6	COL7	COL8	COL9
ROW1	3.9996	3.2396	3.2396	1.8142	1.8142	1.8142	2.3040	2.3040	2.3040
ROW2	3.2396	3.9996	3.2396	1.8142	1.8142	1.8142	2.3040	2.3040	2.3040
ROW3	3.2396	3.2396	3.9996	1.8142	1.8142	1.8142	2.3040	2.3040	2.3040
ROW4	1.8142	1.8142	1.8142	3.9997	3.2397	3.2397	1.4976	1.4976	1.4976
ROW5	1.8142	1.8142	1.8142	3.2397	3.9997	3.2397	1.4976	1.4976	1.4976
ROW6	1.8142	1.8142	1.8142	3.2397	3.2397	3.9997	1.4976	1.4976	1.4976
ROW7	2.3040	2.3040	2.3040	1.4976	1.4976	1.4976	4.0000	2.5600	2.5600
ROW8	2.3040	2.3040	2.3040	1.4976	1.4976	1.4976	2.5600	4.0000	2.5600
ROW9	2.3040	2.3040	2.3040	1.4976	1.4976	1.4976	2.5600	2.5600	4.0000

Note:
CYY: Covariance matrix among the indicators of the endogenous constructs.
CYX: Covariance matrix between the indicators of the exogenous and the endogenous constructs.
CXX: Covariance matrix among the indicators of the exogenous constructs.
COV: Covariance matrix among the indicators.

Exhibit A14.2 gives the PROC IML output. The covariance matrix given in the table is, within rounding errors, the same as that given in Table 14.6.

A14.2 MODEL EFFECTS

In many instances the researcher is interested in determining a number of effects in the structural model, which can be classified as: (1) effects among the endogenous constructs (i.e., how one endogenous construct affects other endogenous construct(s)); (2) effects of the exogenous

constructs on the endogenous constructs (i.e., how the exogenous constructs affect the various endogenous constructs); and (3) effects of the constructs on the indicators. Each of these effects is discussed in the following sections using simple models. However, the formulae given are general and can be used to obtain the various effects for any structural models.

A14.2.1 Effects among the Endogenous Constructs

Consider the structural model given in Figure A14.3. The figure, depicting only the relationships among the constructs, can be represented by the following equations

$$\eta_1 = \gamma_{11}\xi_1 + \zeta_1 \tag{A14.22}$$

$$\eta_2 = \gamma_{22}\xi_2 + \beta_{21}\eta_1 + \zeta_2 \tag{A14.23}$$

$$\eta_3 = \beta_{31}\eta_1 + \beta_{32}\eta_2 + \zeta_3. \tag{A14.24}$$

Concentrating on the paths among the endogenous constructs, it can be seen from Figure A14.3 that some of the endogenous constructs are directly and/or indirectly affected by other endogenous constructs. For example, η_2 is directly affected by η_1, and it is not indirectly affected by any other endogenous construct. On the other hand, η_3 is directly affected by η_1 and η_2, and it is also indirectly affected by η_1 via η_2. The total effect of a given construct is the sum of all its direct and indirect effects. The following section discusses the direct and indirect effects in greater detail and illustrates the computational procedures.

Direct Effects

As discussed above, direct effects result when one construct directly affects another construct. The direct effects can be obtained directly from the structural equations. In Eq. A14.23 the direct effect of η_1 on η_2 is given by the respective structural coefficient, β_{21}. Similarly, in Eq. A14.24 the direct effect of η_1 and η_2 on η_3 are, respectively, given by β_{31} and β_{32}. It is obvious that the direct effects among the endogenous constructs are given by the $\mathbf{B}$ matrix. For the model given in Figure A14.3, the following $\mathbf{B}$ matrix gives the effect of the column construct on the row construct:

$$\mathbf{B} = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{21} & 0 & 0 \\ \beta_{31} & \beta_{32} & 0 \end{pmatrix}. \tag{A14.25}$$

Indirect Effects

The indirect effect of one construct on another construct *must* be through one or more other construct(s). For example, in Figure A14.3 the indirect effect of η_1 on η_3 is via η_2. That is, η_1 indirectly affects η_3 through η_2, or in Figure A14.4 the indirect effect of η_1 on η_4 is through η_2 and η_3. The order of an indirect effect is denoted by the length of the effect, and the length

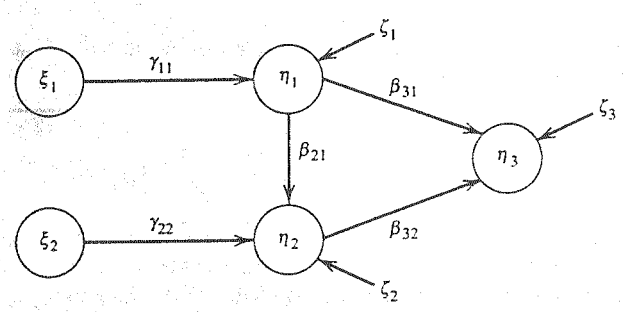


Figure A14.3 Structural model.

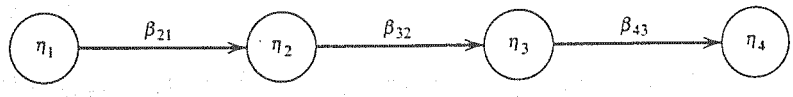


Figure A14.4 Indirect effects of length three.

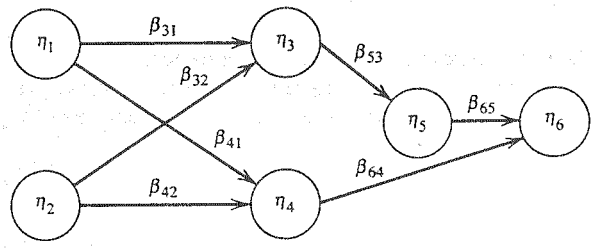


Figure A14.5 Multiple indirect effects.

of an effect is defined as the number of links or paths between the two constructs. For instance, in Figure A14.4 the indirect effect of η_1 on η_3 is of length two as there are two links between η_1 and η_3. One link is between η_1 and η_2 and the other link is between η_2 and η_3. Also, the indirect link between η_1 and η_4 is of length three as there are three links between η_1 and η_4. It is also possible that a given construct may have multiple indirect effects. As an example, in Figure A14.5 η_1 indirectly affects η_6 through η_3 and η_5, as well as through η_4. The total indirect effect of a construct, therefore, is equal to the sum of all its indirect effects.

Indirect effects are equal to the product of the structural coefficients of the links between the effects. As an example, the indirect effects of η_1 on η_6 in Figure A14.5 are given by $\beta_{31}\beta_{53}\beta_{65}$ and $\beta_{41}\beta_{64}$. The total indirect effect of η_1 on η_6 is therefore equal to $\beta_{31}\beta_{53}\beta_{65} + \beta_{41}\beta_{64}$. In general, it has been shown that indirect effects of length or order k are given by $\mathbf{B}^k$, where $\mathbf{B}$ is the matrix of beta coefficients. That is, the indirect effect of length two for the model given in Figure A14.3 is given by

$$\mathbf{B}^2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \beta_{21}\beta_{32} & 0 & 0 \end{pmatrix}.$$

And the indirect effects of length 3 are given by

$$\mathbf{B}^3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

That is, as is obvious from Figure A14.3, there are no indirect effects of length three. The total indirect effects are given by

$$\mathbf{B}^2 + \mathbf{B}^3 + \cdots + \mathbf{B}^\infty,$$

which has been shown to be equal to

$$(\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I} - \mathbf{B}. \tag{A14.26}$$

Total Effects

The total effect is the sum of direct and indirect effects. For example, the total effect of η_1 on η_2 is the sum of the direct effect and the indirect effects of η_1 on η_2. From Eqs. A14.25 and A14.26, the total effects are

$$(\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I} - \mathbf{B} + \mathbf{B}$$

or

$$(\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I}. \tag{A14.27}$$

A14.2.2 Effects of Exogenous Constructs on Endogenous Constructs

Direct Effects

From Figure A14.3 it can be seen that the exogenous construct, ξ_1, directly affects η_1 and the exogenous construct, ξ_2, directly affects η_2. The direct effects of ξ_1 and ξ_2, respectively, are γ_{11} and γ_{22}. The direct effects of exogenous constructs on the endogenous constructs are given by the Γ matrix and are equal to

$$\Gamma = \begin{pmatrix} \gamma_{11} & 0 \\ 0 & \gamma_{22} \end{pmatrix}. \tag{A14.28}$$

Indirect Effects

The following indirect effects can be identified in Figure A14.3:

1. ξ_1 indirectly affects η_2 through η_1 and the indirect effect is given by $\gamma_{11}\beta_{21}$.
2. ξ_1 indirectly affects η_3 through η_1, and through η_1 and η_2. These effects are, respectively, given by $\gamma_{11}\beta_{31}$ and $\gamma_{11}\beta_{21}\beta_{32}$.
3. The indirect effect of ξ_2 on η_3 is through η_2 and is given by $\gamma_{22}\beta_{32}$.

In general, it can be shown that the indirect effects of the exogenous constructs on the endogenous constructs are given by

$$[(\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I}]\Gamma, \tag{A14.29}$$

and, therefore, the total effects are given by

$$(\mathbf{I} - \mathbf{B})^{-1}\Gamma - \Gamma + \Gamma$$

or

$$(\mathbf{I} - \mathbf{B})^{-1}\Gamma. \tag{A14.30}$$

A14.2.3 Effects of the Constructs on Their Indicators

Consider the model given in Figure A14.2. The effects of the constructs on the indicators can be classified as: (1) the direct effect of each construct on its respective indicators; (2) the indirect effect of exogenous constructs on indicators of endogenous constructs; and (3) the indirect effect of the endogenous construct on the indicators of other endogenous constructs. These effects are discussed below.

Direct Effects

The direct effect of the exogenous construct on its indicators is given by the respective λ coefficient. For example, the effect of ξ_1 on x_1 is given by λ^x_{11}, and the indirect effect of η_1 on y_1 is given by λ^y_{11}. Therefore, the direct effects of exogenous and endogenous constructs on their indicators are, respectively, given by the Λ_x and Λ_y matrices.

Indirect Effects

The indicators of the exogenous constructs are not indirectly affected. Only the indicators of endogenous constructs are indirectly affected. In Figure A14.2, y_1 is indirectly affected by ξ_1 via η_1 and the effect is equal to $\gamma_{11}\lambda^y_{11}$. Similarly, y_4 is indirectly affected by ξ_1 via η_1 and η_2

and the effect is given by $\gamma_{11}\beta_{21}\lambda^y_{42}$. The indicator y_4 is also indirectly affected by η_1 through η_2, and this effect is given by $\beta_{21}\lambda^y_{42}$. The total indirect effect of y_4 is equal to $\gamma_{11}\beta_{21}\lambda^y_{42} + \beta_{21}\lambda^y_{42}$. The total effect of y_4 is equal to the sum of all indirect and direct effects and is equal to

$$\lambda^y_{42} + \gamma_{11}\beta_{21}\lambda^y_{42} + \beta_{21}\lambda_{42}.$$

In general, the indirect effects on the indicators of endogenous constructs are given by

$$\mathbf{\Lambda}_y[(\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I}],$$

and the total effects are given by

$$\mathbf{\Lambda}_y[(\mathbf{I} - \mathbf{B})^{-1} - \mathbf{I}] + \mathbf{\Lambda}_y$$

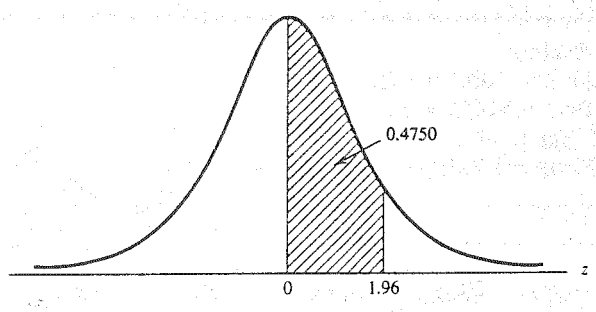
or

$$\mathbf{\Lambda}_y(\mathbf{I} - \mathbf{B})^{-1}. \tag{A14.31}$$

Statistical Tables

Table T.1 **Standard Normal Probabilities**

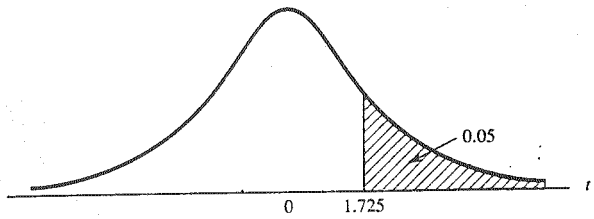
Example
$\Pr(0 \le z \le 1.96) = 0.4750$
$\Pr(z \ge 1.96) = 0.5 - 0.4750$
$= 0.025$



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990

Table T.2 Students' t-Distribution Critical Points

Example
$\Pr(t > 2.086) = 0.025$
$\Pr(t > 1.725) = 0.05$
for $df = 20$
$\Pr(|t| > 1.725) = 0.10$



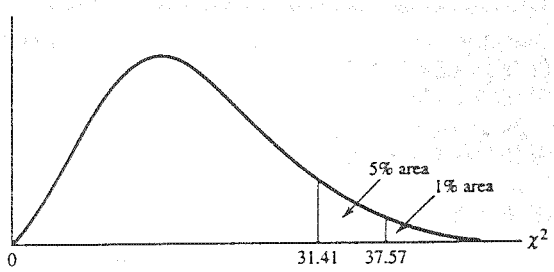
Pr / df	0.25 0.50	0.10 0.20	0.05 0.10	0.025 0.05	0.01 0.02	0.005 0.010	0.001 0.002
1	1.000	3.078	6.314	12.706	31.821	63.657	318.31
2	0.816	1.886	2.920	4.303	6.965	9.925	22.327
3	0.765	1.638	2.353	3.182	4.541	5.841	10.214
4	0.741	1.533	2.132	2.776	3.747	4.604	7.173
5	0.727	1.476	2.015	2.571	3.365	4.032	5.893
6	0.718	1.440	1.943	2.447	3.143	3.707	5.208
7	0.711	1.415	1.895	2.365	2.998	3.499	4.785
8	0.706	1.397	1.860	2.306	2.896	3.355	4.501
9	0.703	1.383	1.833	2.262	2.821	3.250	4.297
10	0.700	1.372	1.812	2.228	2.764	3.169	4.144
11	0.697	1.363	1.796	2.201	2.718	3.106	4.025
12	0.695	1.356	1.782	2.179	2.681	3.055	3.930
13	0.694	1.350	1.771	2.160	2.650	3.012	3.852
14	0.692	1.345	1.761	2.145	2.624	2.977	3.787
15	0.691	1.341	1.753	2.131	2.602	2.947	3.733
16	0.690	1.337	1.746	2.120	2.583	2.921	3.686
17	0.689	1.333	1.740	2.110	2.567	2.898	3.646
18	0.688	1.330	1.734	2.101	2.552	2.878	3.610
19	0.688	1.328	1.729	2.093	2.539	2.861	3.579
20	0.687	1.325	1.725	2.086	2.528	2.845	3.552
21	0.686	1.323	1.721	2.080	2.518	2.831	3.527
22	0.686	1.321	1.717	2.074	2.508	2.819	3.505
23	0.685	1.319	1.714	2.069	2.500	2.807	3.485
24	0.685	1.318	1.711	2.064	2.492	2.797	3.467
25	0.684	1.316	1.708	2.060	2.485	2.787	3.450
26	0.684	1.315	1.706	2.056	2.479	2.779	3.435
27	0.684	1.314	1.703	2.052	2.473	2.771	3.421
28	0.683	1.313	1.701	2.048	2.467	2.763	3.408
29	0.683	1.311	1.699	2.045	2.462	2.756	3.396
30	0.683	1.310	1.697	2.042	2.457	2.750	3.385
40	0.681	1.303	1.684	2.021	2.423	2.704	3.307
60	0.679	1.296	1.671	2.000	2.390	2.660	3.232
120	0.677	1.289	1.658	1.980	2.358	2.167	3.160
∞	0.674	1.282	1.645	1.960	2.326	2.576	3.090

Note: The smaller probability shown at the head of each column is the area in one tail; the larger probability is the area in both tails.

Source: From E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians*, vol. 1, 3d ed., table 12, Cambridge University Press, New York, 1966. Reproduced by permission of the editors and trustees of *Biometrika*.

Table T.3 χ^2 Critical Points

Example
$\Pr(\chi^2 > 23.8277) = 0.25$
$\Pr(\chi^2 > 31.4104) = 0.05$
for $df = 20$
$\Pr(\chi^2 > 37.5662) = 0.01$



df \ Pr	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32330	2.70554	3.84146	5.02389	6.63490	7.87944	10.828
2	2.77259	4.60517	5.99146	7.37776	9.21034	10.5966	13.816
3	4.10834	6.25139	7.81473	9.34840	11.3449	12.8382	16.266
4	5.38527	7.77944	9.48773	11.1433	13.2767	14.8603	18.467
5	6.62568	9.23636	11.0705	12.8325	15.0863	16.7496	20.515
6	7.84080	10.6446	12.5916	14.4494	16.8119	18.5476	22.458
7	9.03715	12.0170	14.0671	16.0128	18.4753	20.2777	24.322
8	10.2189	13.3616	15.5073	17.5345	20.0902	21.9550	26.125
9	11.3888	14.6837	16.9190	19.0228	21.6660	23.5894	27.877
10	12.5489	15.9872	18.3070	20.4832	23.2093	25.1882	29.588
11	13.7007	17.2750	19.6751	21.9200	24.7250	26.7568	31.264
12	14.8454	18.5493	21.0261	23.3367	26.2170	28.2995	32.909
13	15.9839	19.8119	22.3620	24.7356	27.6882	29.8195	34.528
14	17.1169	21.0641	23.6848	26.1189	29.1412	31.3194	36.123
15	18.2451	22.3071	24.9958	27.4884	30.5779	32.8013	37.697
16	19.3689	23.5418	26.2962	28.8454	31.9999	34.2672	39.252
17	20.4887	24.7690	27.5871	30.1910	33.4087	35.7185	40.790
18	21.6049	25.9894	28.8693	31.5264	34.8053	37.1565	42.312
19	22.7178	27.2036	30.1435	32.8523	36.1909	38.5823	43.820
20	23.8277	28.4120	31.4104	34.1696	37.5662	39.9968	45.315
21	24.9348	29.6151	32.6706	35.4789	38.9322	41.4011	46.797
22	26.0393	30.8133	33.9244	36.7807	40.2894	42.7957	48.268
23	27.1413	32.0069	35.1725	38.0756	41.6384	44.1813	49.728
24	28.2412	33.1962	36.4150	39.3641	42.9798	45.5585	51.179
25	29.3389	34.3816	37.6525	40.6465	44.3141	46.9279	52.618
26	30.4346	35.5632	38.8851	41.9232	45.6417	48.2899	54.052
27	31.5284	36.7412	40.1133	43.1945	46.9629	49.6449	55.476
28	32.6205	37.9159	41.3371	44.4608	48.2782	50.9934	56.892
29	33.7109	39.0875	42.5570	45.7223	49.5879	52.3356	58.301
30	34.7997	40.2560	43.7730	46.9792	50.8922	53.6720	59.703
40	45.6160	51.8051	55.7585	59.3417	63.6907	66.7660	73.402
50	56.3336	63.1671	67.5048	71.4202	76.1539	79.4900	86.661
60	66.9815	74.3970	79.0819	83.2977	88.3794	91.9517	99.607
70	77.5767	85.5270	90.5312	95.0232	100.425	104.215	112.317
80	88.1303	96.5782	101.879	106.629	112.329	116.321	124.839
90	98.6499	107.565	113.145	118.136	124.116	128.299	137.208
100	109.141	118.498	124.342	129.561	135.807	140.169	149.449
$Z^{\dagger}$	+0.6745	+1.2816	+1.6449	+1.9600	+2.3263	+2.5758	+3.0902

†For df greater than 100, the expression

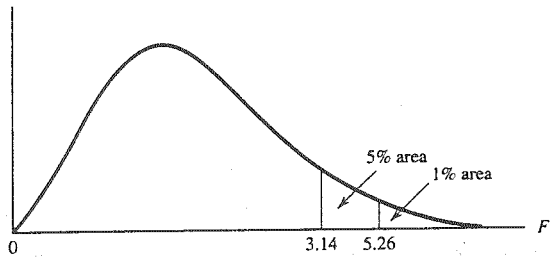
$$\sqrt{2\chi^2} - \sqrt{(2k-1)} = Z$$

follows the standardized normal distribution, where k represents the degrees of freedom.

Source: From E. S. Pearson and H. O. Hartley, eds., *Biometrika Tables for Statisticians*, vol. 1, 3d ed., table 8, Cambridge University Press, New York, 1966. Reproduced by permission of the editors and trustees of *Biometrika*.

Table T.4 *F*-Distribution

Example
Pr ($F > 1.59$) = 0.25
Pr ($F > 2.42$) = 0.10
for *df* $N_1 = 10$
Pr ($F > 3.14$) = 0.05
and $N_2 = 9$
Pr ($F > 5.26$) = 0.01



df for Denominator N_2	Pr	*df* for Numerator N_1 1	2	3	4	5	6	7	8	9	10	11	12
1	.25	5.83	7.50	8.20	8.58	8.82	8.98	9.10	9.19	9.26	9.32	9.36	9.41
	.10	39.9	49.5	53.6	55.8	57.2	58.2	58.9	59.4	59.9	60.2	60.5	60.7
	.05	161	200	216	225	230	234	237	239	241	242	243	244
2	.25	2.57	3.00	3.15	3.23	3.28	3.31	3.34	3.35	3.37	3.38	3.39	3.39
	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.40	9.41
	.05	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4
	.01	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4
3	.25	2.02	2.28	2.36	2.39	2.41	2.42	2.43	2.44	2.44	2.44	2.45	2.45
	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.22
	.05	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.76	8.74
	.01	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	27.1
4	.25	1.81	2.00	2.05	2.06	2.07	2.08	2.08	2.08	2.08	2.08	2.08	2.08
	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.91	3.90
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.94	5.91
	.01	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.4
5	.25	1.69	1.85	1.88	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89	1.89
	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.28	3.27
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.71	4.68
	.01	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.96	9.89
6	.25	1.62	1.76	1.78	1.79	1.79	1.78	1.78	1.78	1.77	1.77	1.77	1.77
	.10	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.92	2.90
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
	.01	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.79	7.72
7	.25	1.57	1.70	1.72	1.72	1.71	1.71	1.70	1.70	1.69	1.69	1.69	1.68
	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.68	2.67
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.60	3.57
	.01	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.54	6.47
8	.25	1.54	1.66	1.67	1.66	1.66	1.65	1.64	1.64	1.63	1.63	1.63	1.62
	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.52	2.50
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.31	3.28
	.01	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.73	5.67
9	.25	1.51	1.62	1.63	1.63	1.62	1.61	1.60	1.60	1.59	1.59	1.58	1.58
	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.40	2.38
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.10	3.07
	.01	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.18	5.11

Source: E. S. Pearson and H. O. Hartley, eds. *Biometrika Tables for Statisticians*, vol. 1, 3d ed., table 18, p. 558, Cambridge University Press, New York, 1966. Reproduced by permission of the editors and trustees of *Biometrika*.

Table T.4 (Continued)

df for Numerator N_1													*df* for Denominator N_2
15	20	24	30	40	50	60	100	120	200	500	∞	Pr	
9.49	9.58	9.63	9.67	9.71	9.74	9.76	9.78	9.80	9.82	9.84	9.85	.25	
61.2	61.7	62.0	62.3	62.5	62.7	62.8	63.0	63.1	63.2	63.3	63.3	.10	1
246	248	249	250	251	252	252	253	253	254	254	254	.05	
3.41	3.43	3.43	3.44	3.45	3.45	3.46	3.47	3.47	3.48	3.48	3.48	.25	
9.42	9.44	9.45	9.46	9.47	9.47	9.47	9.48	9.48	9.49	9.49	9.49	.10	2
19.4	19.4	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	19.5	.05	
99.4	99.4	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	99.5	.01	
2.46	2.46	2.46	2.47	2.47	2.47	2.47	2.47	2.47	2.47	2.47	2.47	.25	
5.20	5.18	5.18	5.17	5.16	5.15	5.15	5.14	5.14	5.14	5.14	5.13	.10	3
8.70	8.66	8.64	8.62	8.59	8.58	8.57	8.55	8.55	8.54	8.53	8.53	.05	
26.9	26.7	26.6	26.5	26.4	26.4	26.3	26.2	26.2	26.2	26.1	26.1	.01	
2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	2.08	.25	
3.87	3.84	3.83	3.82	3.80	3.80	3.79	3.78	3.78	3.77	3.76	3.76	.10	4
5.86	5.80	5.77	5.75	5.72	5.70	5.69	5.66	5.66	5.65	5.64	5.63	.05	
14.2	14.0	13.9	13.8	13.7	13.7	13.7	13.6	13.6	13.5	13.5	13.5	.01	
1.89	1.88	1.88	1.88	1.88	1.88	1.87	1.87	1.87	1.87	1.87	1.87	.25	
3.24	3.21	3.19	3.17	3.16	3.15	3.14	3.13	3.12	3.12	3.11	3.10	.10	5
4.62	4.56	4.53	4.50	4.46	4.44	4.43	4.41	4.40	4.39	4.37	4.36	.05	
9.72	9.55	9.47	9.38	9.29	9.24	9.20	9.13	9.11	9.08	9.04	9.02	.01	
1.76	1.76	1.75	1.75	1.75	1.75	1.74	1.74	1.74	1.74	1.74	1.74	.25	
2.87	2.84	2.82	2.80	2.78	2.77	2.76	2.75	2.74	2.73	2.73	2.72	.10	6
3.94	3.87	3.84	3.81	3.77	3.75	3.74	3.71	3.70	3.69	3.68	3.67	.05	
7.56	7.40	7.31	7.23	7.14	7.09	7.06	6.99	6.97	6.93	6.90	6.88	.01	
1.68	1.67	1.67	1.66	1.66	1.66	1.65	1.65	1.65	1.65	1.65	1.65	.25	
2.63	2.59	2.58	2.56	2.54	2.52	2.51	2.50	2.49	2.48	2.48	2.47	.10	7
3.51	3.44	3.41	3.38	3.34	3.32	3.30	3.27	3.27	3.25	3.24	3.23	.05	
6.31	6.16	6.07	5.99	5.91	5.86	5.82	5.75	5.74	5.70	5.67	5.65	.01	
1.62	1.61	1.60	1.60	1.59	1.59	1.59	1.58	1.58	1.58	1.58	1.58	.25	
2.46	2.42	2.40	2.38	2.36	2.35	2.34	2.32	2.32	2.31	2.30	2.29	.10	8
3.22	3.15	3.12	3.08	3.04	3.02	3.01	2.97	2.97	2.95	2.94	2.93	.05	
5.52	5.36	5.28	5.20	5.12	5.07	5.03	4.96	4.95	4.91	4.88	4.86	.01	
1.57	1.56	1.56	1.55	1.55	1.54	1.54	1.53	1.53	1.53	1.53	1.53	.25	
2.34	2.30	2.28	2.25	2.23	2.22	2.21	2.19	2.18	2.17	2.17	2.16	.10	9
3.01	2.94	2.90	2.86	2.83	2.80	2.79	2.76	2.75	2.73	2.72	2.71	.05	
4.96	4.81	4.73	4.65	4.57	4.52	4.48	4.42	4.40	4.36	4.33	4.31	.01	

(continued)

Table T.4 (Continued)

df for Denominator N_2	Pr	*df* for Numerator N_1 1	2	3	4	5	6	7	8	9	10	11	12
10	.25	1.49	1.60	1.60	1.59	1.59	1.58	1.57	1.56	1.56	1.55	1.55	1.54
	.10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.30	2.28
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.94	2.91
	.01	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.77	4.71
11	.25	1.47	1.58	1.58	1.57	1.56	1.55	1.54	1.53	1.53	1.52	1.52	1.51
	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.23	2.21
	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.82	2.79
	.01	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.46	4.40
12	.25	1.46	1.56	1.56	1.55	1.54	1.53	1.52	1.51	1.51	1.50	1.50	1.49
	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.17	2.15
	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.72	2.69
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.22	4.16
13	.25	1.45	1.55	1.55	1.53	1.52	1.51	1.50	1.49	1.49	1.48	1.47	1.47
	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.12	2.10
	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.63	2.60
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	4.02	3.96
14	.25	1.44	1.53	1.53	1.52	1.51	1.50	1.49	1.48	1.47	1.46	1.46	1.45
	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.08	2.05
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.57	2.53
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.86	3.80
15	.25	1.43	1.52	1.52	1.51	1.49	1.48	1.47	1.46	1.46	1.45	1.44	1.44
	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.04	2.02
	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.51	2.48
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.73	3.67
16	.25	1.42	1.51	1.51	1.50	1.48	1.47	1.46	1.45	1.44	1.44	1.44	1.43
	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	2.01	1.99
	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.46	2.42
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.62	3.55
17	.25	1.42	1.51	1.50	1.49	1.47	1.46	1.45	1.44	1.43	1.43	1.42	1.41
	.10	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.98	1.96
	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.41	2.38
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.52	3.46
18	.25	1.41	1.50	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.42	1.41	1.40
	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.96	1.93
	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.37	2.34
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.43	3.37
19	.25	1.41	1.49	1.49	1.47	1.46	1.44	1.43	1.42	1.41	1.41	1.40	1.40
	.10	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.94	1.91
	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.34	2.31
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.36	3.30
20	.25	1.40	1.49	1.48	1.46	1.45	1.44	1.43	1.42	1.41	1.40	1.39	1.39
	.10	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.92	1.89
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.31	2.28
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.29	3.23

Table T.4 (Continued)

df for Numerator N_1													*df* for Denominator N_2
15	20	24	30	40	50	60	100	120	200	500	∞	Pr	
1.53	1.52	1.52	1.51	1.51	1.50	1.50	1.49	1.49	1.49	1.48	1.48	.25	
2.24	2.20	2.18	2.16	2.13	2.12	2.11	2.09	2.08	2.07	2.06	2.06	.10	10
2.85	2.77	2.74	2.70	2.66	2.64	2.62	2.59	2.58	2.56	2.55	2.54	.05	
4.56	4.41	4.33	4.25	4.17	4.12	4.08	4.01	4.00	3.96	3.93	3.91	.01	
1.50	1.49	1.49	1.48	1.47	1.47	1.47	1.46	1.46	1.46	1.45	1.45	.25	
2.17	2.12	2.10	2.08	2.05	2.04	2.03	2.00	2.00	1.99	1.98	1.97	.10	11
2.72	2.65	2.61	2.57	2.53	2.51	2.49	2.46	2.45	2.43	2.42	2.40	.05	
4.25	4.10	4.02	3.94	3.86	3.81	3.78	3.71	3.69	3.66	3.62	3.60	.01	
1.48	1.47	1.46	1.45	1.45	1.44	1.44	1.43	1.43	1.43	1.42	1.42	.25	
2.10	2.06	2.04	2.01	1.99	1.97	1.96	1.94	1.93	1.92	1.91	1.90	.10	12
2.62	2.54	2.51	2.47	2.43	2.40	2.38	2.35	2.34	2.32	2.31	2.30	.05	
4.01	3.86	3.78	3.70	3.62	3.57	3.54	3.47	3.45	3.41	3.38	3.36	.01	
1.46	1.45	1.44	1.43	1.42	1.42	1.42	1.41	1.41	1.40	1.40	1.40	.25	
2.05	2.01	1.98	1.96	1.93	1.92	1.90	1.88	1.88	1.86	1.85	1.85	.10	13
2.53	2.46	2.42	2.38	2.34	2.31	2.30	2.26	2.25	2.23	2.22	2.21	.05	
3.82	3.66	3.59	3.51	3.43	3.38	3.34	3.27	3.25	3.22	3.19	3.17	.01	
1.44	1.43	1.42	1.41	1.41	1.40	1.40	1.39	1.39	1.39	1.38	1.38	.25	
2.01	1.96	1.94	1.91	1.89	1.87	1.86	1.83	1.83	1.82	1.80	1.80	.10	14
2.46	2.39	2.35	2.31	2.27	2.24	2.22	2.19	2.18	2.16	2.14	2.13	.05	
3.66	3.51	3.43	3.35	3.27	3.22	3.18	3.11	3.09	3.06	3.03	3.00	.01	
1.43	1.41	1.41	1.40	1.39	1.39	1.38	1.38	1.37	1.37	1.36	1.36	.25	
1.97	1.92	1.90	1.87	1.85	1.83	1.82	1.79	1.79	1.77	1.76	1.76	.10	15
2.40	2.33	2.29	2.25	2.20	2.18	2.16	2.12	2.11	2.10	2.08	2.07	.05	
3.52	3.37	3.29	3.21	3.13	3.08	3.05	2.98	2.96	2.92	2.89	2.87	.01	
1.41	1.40	1.39	1.38	1.37	1.37	1.36	1.36	1.35	1.35	1.34	1.34	.25	
1.94	1.89	1.87	1.84	1.81	1.79	1.78	1.76	1.75	1.74	1.73	1.72	.10	16
2.35	2.28	2.24	2.19	2.15	2.12	2.11	2.07	2.06	2.04	2.02	2.01	.05	
3.41	3.26	3.18	3.10	3.02	2.97	2.93	2.86	2.84	2.81	2.78	2.75	.01	
1.40	1.39	1.38	1.37	1.36	1.35	1.35	1.34	1.34	1.34	1.33	1.33	.25	
1.91	1.86	1.84	1.81	1.78	1.76	1.75	1.73	1.72	1.71	1.69	1.69	.10	17
2.31	2.23	2.19	2.15	2.10	2.08	2.06	2.02	2.01	1.99	1.97	1.96	.05	
3.31	3.16	3.08	3.00	2.92	2.87	2.83	2.76	2.75	2.71	2.68	2.65	.01	
1.39	1.38	1.37	1.36	1.35	1.34	1.34	1.33	1.33	1.32	1.32	1.32	.25	
1.89	1.84	1.81	1.78	1.75	1.74	1.72	1.70	1.69	1.68	1.67	1.66	.10	18
2.27	2.19	2.15	2.11	2.06	2.04	2.02	1.98	1.97	1.95	1.93	1.92	.05	
3.23	3.08	3.00	2.92	2.84	2.78	2.75	2.68	2.66	2.62	2.59	2.57	.01	
1.38	1.37	1.36	1.35	1.34	1.33	1.33	1.32	1.32	1.31	1.31	1.30	.25	
1.86	1.81	1.79	1.76	1.73	1.71	1.70	1.67	1.67	1.65	1.64	1.63	.10	19
2.23	2.16	2.11	2.07	2.03	2.00	1.98	1.94	1.93	1.91	1.89	1.88	.05	
3.15	3.00	2.92	2.84	2.76	2.71	2.67	2.60	2.58	2.55	2.51	2.49	.01	
1.37	1.36	1.35	1.34	1.33	1.33	1.32	1.31	1.31	1.30	1.30	1.29	.25	
1.84	1.79	1.77	1.74	1.71	1.69	1.68	1.65	1.64	1.63	1.62	1.61	.10	20
2.20	2.12	2.08	2.04	1.99	1.97	1.95	1.91	1.90	1.88	1.86	1.84	.05	
3.09	2.94	2.86	2.78	2.69	2.64	2.61	2.54	2.52	2.48	2.44	2.42	.01	

(continued)

Table T.4 (Continued)

df for Denominator N_2	Pr	*df* for Numerator N_1 1	2	3	4	5	6	7	8	9	10	11	12
22	.25	1.40	1.48	1.47	1.45	1.44	1.42	1.41	1.40	1.39	1.39	1.38	1.37
	.10	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.88	1.86
	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.26	2.23
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.18	3.12
24	.25	1.39	1.47	1.46	1.44	1.43	1.41	1.40	1.39	1.38	1.38	1.37	1.36
	.10	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.85	1.83
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.21	2.18
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.09	3.03
26	.25	1.38	1.46	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.37	1.36	1.35
	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86	1.84	1.81
	.05	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.18	2.15
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09	3.02	2.96
28	.25	1.38	1.46	1.45	1.43	1.41	1.40	1.39	1.38	1.37	1.36	1.35	1.34
	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84	1.81	1.79
	.05	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.15	2.12
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03	2.96	2.90
30	.25	1.38	1.45	1.44	1.42	1.41	1.39	1.38	1.37	1.36	1.35	1.35	1.34
	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.79	1.77
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.13	2.09
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.91	2.84
40	.25	1.36	1.44	1.42	1.40	1.39	1.37	1.36	1.35	1.34	1.33	1.32	1.31
	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.73	1.71
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.04	2.00
	.01	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.73	2.66
60	.25	1.35	1.42	1.41	1.38	1.37	1.35	1.33	1.32	1.31	1.30	1.29	1.29
	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.68	1.66
	.05	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.95	1.92
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.56	2.50
120	.25	1.34	1.40	1.39	1.37	1.35	1.33	1.31	1.30	1.29	1.28	1.27	1.26
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.62	1.60
	.05	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.91	1.87	1.83
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.40	2.34
200	.25	1.33	1.39	1.38	1.36	1.34	1.32	1.31	1.29	1.28	1.27	1.26	1.25
	.10	2.73	2.33	2.11	1.97	1.88	1.80	1.75	1.70	1.66	1.63	1.60	1.57
	.05	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93	1.88	1.84	1.80
	.01	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50	2.41	2.34	2.27
∞	.25	1.32	1.39	1.37	1.35	1.33	1.31	1.29	1.28	1.27	1.25	1.24	1.24
	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.57	1.55
	.05	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.79	1.75
	.01	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.25	2.18

Table T.4 (Continued)

df for Numerator N_1													*df* for Denominator N_2
15	20	24	30	40	50	60	100	120	200	500	∞	Pr	
1.36	1.34	1.33	1.32	1.31	1.31	1.30	1.30	1.30	1.29	1.29	1.28	.25	
1.81	1.76	1.73	1.70	1.67	1.65	1.64	1.61	1.60	1.59	1.58	1.57	.10	22
2.15	2.07	2.03	1.98	1.94	1.91	1.89	1.85	1.84	1.82	1.80	1.78	.05	
2.98	2.83	2.75	2.67	2.58	2.53	2.50	2.42	2.40	2.36	2.33	2.31	.01	
1.35	1.33	1.32	1.31	1.30	1.29	1.29	1.28	1.28	1.27	1.27	1.26	.25	
1.78	1.73	1.70	1.67	1.64	1.62	1.61	1.58	1.57	1.56	1.54	1.53	.10	24
2.11	2.03	1.98	1.94	1.89	1.86	1.84	1.80	1.79	1.77	1.75	1.73	.05	
2.89	2.74	2.66	2.58	2.49	2.44	2.40	2.33	2.31	2.27	2.24	2.21	.01	
1.34	1.32	1.31	1.30	1.29	1.28	1.28	1.26	1.26	1.26	1.25	1.25	.25	
1.76	1.71	1.68	1.65	1.61	1.59	1.58	1.55	1.54	1.53	1.51	1.50	.10	26
2.07	1.99	1.95	1.90	1.85	1.82	1.80	1.76	1.75	1.73	1.71	1.69	.05	
2.81	2.66	2.58	2.50	2.42	2.36	2.33	2.25	2.23	2.19	2.16	2.13	.01	
1.33	1.31	1.30	1.29	1.28	1.27	1.27	1.26	1.25	1.25	1.24	1.24	.25	
1.74	1.69	1.66	1.63	1.59	1.57	1.56	1.53	1.52	1.50	1.49	1.48	.10	28
2.04	1.96	1.91	1.87	1.82	1.79	1.77	1.73	1.71	1.69	1.67	1.65	.05	
2.75	2.60	2.52	2.44	2.35	2.30	2.26	2.19	2.17	2.13	2.09	2.06	.01	
1.32	1.30	1.29	1.28	1.27	1.26	1.26	1.25	1.24	1.24	1.23	1.23	.25	
1.72	1.67	1.64	1.61	1.57	1.55	1.54	1.51	1.50	1.48	1.47	1.46	.10	30
2.01	1.93	1.89	1.84	1.79	1.76	1.74	1.70	1.68	1.66	1.64	1.62	.05	
2.70	2.55	2.47	2.39	2.30	2.25	2.21	2.13	2.11	2.07	2.03	2.01	.01	
1.30	1.28	1.26	1.25	1.24	1.23	1.22	1.21	1.21	1.20	1.19	1.19	.25	
1.66	1.61	1.57	1.54	1.51	1.48	1.47	1.43	1.42	1.41	1.39	1.38	.10	40
1.92	1.84	1.79	1.74	1.69	1.66	1.64	1.59	1.58	1.55	1.53	1.51	.05	
2.52	2.37	2.29	2.20	2.11	2.06	2.02	1.94	1.92	1.87	1.83	1.80	.01	
1.27	1.25	1.24	1.22	1.21	1.20	1.19	1.17	1.17	1.16	1.15	1.15	.25	
1.60	1.54	1.51	1.48	1.44	1.41	1.40	1.36	1.35	1.33	1.31	1.29	.10	60
1.84	1.75	1.70	1.65	1.59	1.56	1.53	1.48	1.47	1.44	1.41	1.39	.05	
2.35	2.20	2.12	2.03	1.94	1.88	1.84	1.75	1.73	1.68	1.63	1.60	.01	
1.24	1.22	1.21	1.19	1.18	1.17	1.16	1.14	1.13	1.12	1.11	1.10	.25	
1.55	1.48	1.45	1.41	1.37	1.34	1.32	1.27	1.26	1.24	1.21	1.19	.10	120
1.75	1.66	1.61	1.55	1.50	1.46	1.43	1.37	1.35	1.32	1.28	1.25	.05	
2.19	2.03	1.95	1.86	1.76	1.70	1.66	1.56	1.53	1.48	1.42	1.38	.01	
1.23	1.21	1.20	1.18	1.16	1.14	1.12	1.11	1.10	1.09	1.08	1.06	.25	
1.52	1.46	1.42	1.38	1.34	1.31	1.28	1.24	1.22	1.20	1.17	1.14	.10	200
1.72	1.62	1.57	1.52	1.46	1.41	1.39	1.32	1.29	1.26	1.22	1.19	.05	
2.13	1.97	1.89	1.79	1.69	1.63	1.58	1.48	1.44	1.39	1.33	1.28	.01	
1.22	1.19	1.18	1.16	1.14	1.13	1.12	1.09	1.08	1.07	1.04	1.00	.25	
1.49	1.42	1.38	1.34	1.30	1.26	1.24	1.18	1.17	1.13	1.08	1.00	.10	∞
1.67	1.57	1.52	1.46	1.39	1.35	1.32	1.24	1.22	1.17	1.11	1.00	.05	
2.04	1.88	1.79	1.70	1.59	1.52	1.47	1.36	1.32	1.25	1.15	1.00	.01	

Table T.5 Percent Points of the Normal Probability Plot Correlation Coefficient

	Level													
n	.000	.005	.01	.025	.05	.10	.25	.50	.75	.90	.95	.975	.99	.995
3	.866	.867	.869	.872	.879	.891	.924	.966	.991	.999	1.000	1.000	1.000	1.000
4	.784	.813	.822	.845	.868	.894	.931	.958	.979	.992	.996	.998	.999	1.000
5	.726	.803	.822	.855	.879	.902	.935	.960	.977	.988	.992	.995	.997	.998
6	.683	.818	.835	.868	.890	.911	.940	.962	.977	.986	.990	.993	.996	.997
7	.648	.828	.847	.876	.899	.916	.944	.965	.978	.986	.990	.992	.995	.996
8	.619	.841	.859	.886	.905	.924	.948	.967	.979	.986	.990	.992	.995	.996
9	.595	.851	.868	.893	.912	.929	.951	.968	.980	.987	.990	.992	.994	.995
10	.574	.860	.876	.900	.917	.934	.954	.970	.981	.987	.990	.992	.994	.995
11	.556	.868	.883	.906	.922	.938	.957	.972	.982	.988	.990	.992	.994	.995
12	.539	.875	.889	.912	.926	.941	.959	.973	.982	.988	.990	.992	.994	.995
13	.525	.882	.895	.917	.931	.944	.962	.975	.983	.988	.991	.993	.994	.995
14	.512	.888	.901	.921	.934	.947	.964	.976	.984	.989	.991	.993	.994	.995
15	.500	.894	.907	.925	.937	.950	.965	.977	.984	.989	.991	.993	.994	.995
16	.489	.899	.912	.928	.940	.952	.967	.978	.985	.989	.991	.993	.994	.995
17	.478	.903	.916	.931	.942	.954	.968	.979	.986	.990	.992	.993	.994	.995
18	.469	.907	.919	.934	.945	.956	.969	.979	.986	.990	.992	.993	.995	.995
19	.460	.909	.923	.937	.947	.958	.971	.980	.987	.990	.992	.993	.995	.995
20	.452	.912	.925	.939	.950	.960	.972	.981	.987	.991	.992	.994	.995	.995
21	.445	.914	.928	.942	.952	.961	.973	.981	.987	.991	.993	.994	.995	.996
22	.437	.918	.930	.944	.954	.962	.974	.982	.988	.991	.993	.994	.995	.996
23	.431	.922	.933	.947	.955	.964	.975	.983	.988	.991	.993	.994	.995	.996
24	.424	.926	.936	.949	.957	.965	.975	.983	.988	.992	.993	.994	.995	.996
25	.418	.928	.937	.950	.958	.966	.976	.984	.989	.992	.993	.994	.995	.996
26	.412	.930	.939	.952	.959	.967	.977	.984	.989	.992	.993	.994	.995	.996
27	.407	.932	.941	.953	.960	.968	.977	.984	.989	.992	.994	.995	.995	.996
28	.402	.934	.943	.955	.962	.969	.978	.985	.990	.992	.994	.995	.995	.996
29	.397	.937	.945	.956	.962	.969	.979	.985	.990	.992	.994	.995	.995	.996
30	.392	.938	.947	.957	.964	.970	.979	.986	.990	.993	.994	.995	.996	.996
31	.388	.939	.948	.958	.965	.971	.980	.986	.990	.993	.994	.995	.996	.996
32	.383	.939	.949	.959	.966	.972	.980	.986	.990	.993	.994	.995	.996	.996
33	.379	.940	.950	.960	.967	.973	.981	.987	.991	.993	.994	.995	.996	.996
34	.375	.941	.951	.960	.967	.973	.981	.987	.991	.993	.994	.995	.996	.996
35	.371	.943	.952	.961	.968	.974	.982	.987	.991	.993	.995	.995	.996	.997
36	.367	.945	.953	.962	.968	.974	.982	.987	.991	.994	.995	.996	.996	.997
37	.364	.947	.955	.962	.969	.975	.982	.988	.991	.994	.995	.996	.996	.997
38	.360	.948	.956	.964	.970	.975	.983	.988	.992	.994	.995	.996	.996	.997
39	.357	.949	.957	.965	.971	.976	.983	.988	.992	.994	.995	.996	.996	.997
40	.354	.949	.958	.966	.972	.977	.983	.988	.992	.994	.995	.996	.996	.997
41	.351	.950	.958	.967	.972	.977	.984	.989	.992	.994	.995	.996	.996	.997
42	.348	.951	.959	.967	.973	.978	.984	.989	.992	.994	.995	.996	.997	.997
43	.345	.953	.959	.967	.973	.978	.984	.989	.992	.994	.995	.996	.997	.997
44	.342	.954	.960	.968	.973	.978	.984	.989	.992	.994	.995	.996	.997	.997
45	.339	.955	.961	.969	.974	.978	.985	.989	.993	.994	.995	.996	.997	.997
46	.336	.956	.962	.969	.974	.979	.985	.990	.993	.995	.995	.996	.997	.997
47	.334	.956	.963	.970	.974	.979	.985	.990	.993	.995	.995	.996	.997	.997
48	.331	.957	.963	.970	.975	.980	.985	.990	.993	.995	.996	.996	.997	.997
49	.329	.957	.964	.971	.975	.980	.986	.990	.993	.995	.996	.996	.997	.997
50	.326	.959	.965	.972	.977	.981	.986	.990	.993	.995	.996	.996	.997	.997
55	.315	.962	.967	.974	.978	.982	.987	.991	.994	.995	.996	.997	.997	.997
60	.305	.965	.970	.976	.980	.983	.988	.991	.994	.995	.996	.997	.997	.998
65	.296	.967	.972	.977	.981	.984	.989	.992	.994	.996	.996	.997	.997	.998
70	.288	.969	.974	.978	.982	.985	.989	.993	.995	.996	.997	.997	.998	.998
75	.281	.971	.975	.979	.983	.986	.990	.993	.995	.996	.997	.997	.998	.998
80	.274	.973	.976	.980	.984	.987	.991	.993	.995	.996	.997	.997	.998	.998
85	.268	.974	.977	.981	.985	.987	.991	.994	.995	.997	.997	.997	.998	.998
90	.263	.976	.978	.982	.985	.988	.991	.994	.996	.997	.997	.998	.998	.998
95	.257	.977	.979	.983	.986	.989	.992	.994	.996	.997	.997	.998	.998	.998
100	.252	.979	.981	.984	.987	.989	.992	.994	.996	.997	.998	.998	.998	.998

Source: J. J. Filliben (1975). "The Probability Plot Correlation Coefficient Test for Normality," *Technometrics*, 17 (1), 113.

Table T.6 **Simulation Percentiles of b_2**

Sample Size	Percentiles											
	1	2	2.5	5	10	20	80	90	95	97.5	98	99
7	1.25	1.30	1.34	1.41	1.53	1.70	2.78	3.20	3.55	3.85	3.93	4.23
8	1.31	1.37	1.40	1.46	1.58	1.75	2.84	3.31	3.70	4.09	4.20	4.53
9	1.35	1.42	1.45	1.53	1.63	1.80	2.98	3.43	3.86	4.28	4.41	4.82
10	1.39	1.45	1.49	1.56	1.68	1.85	3.01	3.53	3.95	4.40	4.55	5.00
12	1.46	1.52	1.56	1.64	1.76	1.93	3.06	3.55	4.05	4.56	4.73	5.20
15	1.55	1.61	1.64	1.72	1.84	2.01	3.13	3.62	4.13	4.66	4.85	5.30
20	1.65	1.71	1.74	1.82	1.95	2.13	3.21	3.68	4.17	4.68	4.87	5.36
25	1.72	1.79	1.83	1.91	2.03	2.20	3.23	3.68	4.16	4.65	4.82	5.30
30	1.79	1.86	1.90	1.98	2.10	2.26	3.25	3.68	4.11	4.59	4.75	5.21
35	1.84	1.91	1.95	2.03	2.14	2.31	3.27	3.68	4.10	4.53	4.68	5.13
40	1.89	1.96	1.98	2.07	2.19	2.34	3.28	3.67	4.06	4.46	4.61	5.04
45	1.93	2.00	2.03	2.11	2.22	2.37	3.28	3.65	4.00	4.39	4.52	4.94
50	1.95	2.03	2.06	2.15	2.25	2.41	3.28	3.62	3.99	4.33	4.45	4.88

Source: R. B. D'Agostino and G. L. Tietjen (1971). "Simulation Probability Points of b_2 for Small Samples," *Biometrika*, 58 (3), 670.

Table T.7 **Simulation Probability Points of $\sqrt{b_1}$**

n	Two-sided Test					
	0.20	0.10	0.05	0.02	0.01	0.002
5	0.819	1.058	1.212	1.342	1.396	1.466
6	0.805	1.034	1.238	1.415	1.498	1.642
7	0.787	1.008	1.215	1.432	1.576	1.800
8	0.760	0.991	1.202	1.455	1.601	1.873
9	0.752	0.977	1.189	1.408	1.577	1.866
10	0.722	0.950	1.157	1.397	1.565	1.887
11	0.715	0.929	1.129	1.376	1.540	1.924
13	0.688	0.902	1.099	1.312	1.441	1.783
15	0.648	0.862	1.048	1.275	1.462	1.778
17	0.629	0.820	1.009	1.188	1.358	1.705
20	0.593	0.777	0.951	1.152	1.303	1.614
23	0.562	0.743	0.900	1.119	1.276	1.555
25	0.543	0.714	0.876	1.073	1.218	1.468
30	0.510	0.664	0.804	0.985	1.114	1.410
35	0.474	0.624	0.762	0.932	1.043	1.332

Source: R. B. D'Agostino and G. L. Tietjen (1973). "Approaches to the Null Distribution of $\sqrt{b_1}$," *Biometrika*, 60 (1), p. 172.

References

Affifi, A. A. and V. Clark (1984). *Computer-Aided Multivariate Analysis*, Lifetime Learning Publications, Belmont, CA.

Agresti, A. (1984). *Analysis of Ordinal Categorical Data*, Wiley, New York.

Agresti, A. (1990). *Categorical Data Analysis*, Wiley, New York.

Allen, S. J. and R. Hubbard (1986). "Regression Equations of the Latent Roots of Random Data Correlation Matrices with Unities on the Diagonal," *Multivariate Behavioral Research*, vol. 21, 393–398.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd ed. Wiley, New York.

Andrews, F. M., L. Klem, T. N. Davidson, P. M. O'Malley, and W. L. Rodgers (1981). *A Guide for Selecting Statistical Techniques for Analyzing Social Science Data*, Institute for Social Research, Univ. of Michigan Press, Ann Arbor.

Bagozzi, R. P. (1980). "Performance and Satisfaction in an Industrial Sales Force: An Examination of Their Antecedents and Simultaneity," *Journal of Marketing*, 44 (Spring), 65–77.

Bearden, W. O., S. Sharma, and J. E. Teel (1982). "Sample Size Effects on Chi Square and Other Statistics Used in Evaluating Causal Models," *Journal of Marketing Research*, 19 (November 1982), 425–430.

Bentler, P. M. (1982). *Theory and Implementation of EQS, A Structural Equations Program*, BMDP Statistical Software, Inc., Los Angeles.

BIOMED (1990). *BMDP Statistical Software Manual*, vols. 1 & 2, W. J. Dixon (chief ed.), University of California Press, Los Angeles.

Bollen, K. A. (1989). *Structural Equations with Latent Variables*, Wiley, New York.

Bone, P. F., S. Sharma, and T. A. Shimp (1989). "A Bootstrap Procedure for Evaluating the Goodness-of-Fit Indices of Structural Equation and Confirmatory Factor Models," *Journal of Marketing Research* (February 1989), 105–111.

Cattell, R. B. (1966). "The Meaning and Strategic Use of Factor Analysis," in R. B. Cattell (ed.), *Handbook of Multivariate Experimental Psychology*, Rand McNally, Chicago.

Cliff, N. (1988). "The Eigenvalue-Greater-than-One Rule and the Reliability of Components," *Psychological Bulletin*, 103 (2), 276–279.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, New York.

Costanza, M. C. and A. A. Affifi (1979). "Comparison of Stopping Rules in Forward Stepwise Discriminant Analysis," *Journal of the American Statistical Association*, 74, 777–785.

Cox, D. R. and E. J. Snell (1989). *The Analysis of Binary Data*, 2nd ed., Chapman & Hall, London.

D'Agostino, R. B. and G. L. Tietjen (1971). "Simulation Probability Points of b_2 in Small Samples," *Biometrika*, 58, 669–672.

D'Agostino, R. B. and G. L. Tietjen (1973). "Approaches to the Null Distribution of b_1," *Biometrika*, 60, 169–173.

Daniel, C. and F. S. Wood (1980). *Fitting Equations to Data*, Wiley, New York.

Dillon, W. R. and M. Goldstein (1984). *Multivariate Analysis*, Wiley, New York.

Efron, B. (1987). "Better Bootstrap Confidence Intervals," *Journal of the American Statistical Society*, 82 (March), 171–185.

Etgar, M. (1976). "Channel Domination and Countervailing Power in Distributive Channels," *Journal of Marketing Research*, 13 (August), 254–262.

Everitt, B. S. (1979). "A Monte Carlo Investigation of the Robustness of Hotelling's One and Two Sample T^2 Tests," *Journal of the American Statistical Association*, 74, 48–51.

Filliben, J. J. (1975). "The Probability Plot Correlation Coefficient Test for Normality," *Technometrics*, 17 (1), 111–117.

Freeman, D. H., Jr. (1987). *Applied Categorical Data Analysis*, Dekker, New York.

Gilbert, E. S. (1969). "The Effect of Unequal Variance–Covariance Matrices on Fisher's Linear Discriminant Function," *Biometrics*, 25, 505–516.

Glass, G.V. and K. Hopkins (1984). *Statistical Methods in Education and Psychology*, Prentice-Hall, Englewood Cliffs, N.J.

Glass, G. V., P. D. Peckham, and J. R. Sanders (1972). "Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analyses of Variance and Covariance," *Review of Educational Research*, 42, 237–288.

Glick, N. (1978). "Additive Estimators for Probabilities of Correct Classification," *Pattern Recognition*, 10, 211–222.

Gnandesikan, R. (1977). *Methods for Statistical Analysis of Multivariate Observations*, Wiley, New York.

Goldstein, M. and W. R. Dillon (1978). *Discrete Discriminant Analysis*, Wiley, New York.

Green, P. E. (1976). *Mathematical Tools for Applied Multivariate Analysis*, Academic Press, New York.

Green, P. E. (1978). *Analyzing Multivariate Data*, Dryden, Hinsdale, Ill.

Guttman, L. (1953). "Image Theory for the Structure of Quantitative Variates," *Psychometrika*, 18, 277–296.

Haberman, S. J. (1978). *Analysis of Qualitative Data*, Academic Press, New York.

Hakstian, A. R., J. C. Roed, and J. C. Linn (1979). "Two Sample T Procedures and the Assumption of Homogeneous Covariance Matrices," *Psychological Bulletin*, 86, 1255–1263.

Harman, H. H. (1976). *Modern Factor Analysis*, Univ. of Chicago Press, Chicago.

Hartigan, J. (1975). *Clustering Algorithms*, Wiley, New York.

Hayduk, L. A. (1987). *Structural Equation Modeling with LISREL*, Johns Hopkins Press, Baltimore.

Holloway, L. N. and O. J. Dunn (1967). "The Robustness of Hotelling's T^2," *Journal of the American Statistical Association*, 62, 124–136.

Hopkins, J. W. and P. P. F. Clay (1963). "Some Empirical Distributions of Bivariate T2 and Homoscedasticity Criterion M under Unequal Variance and Leptokurtosis," *Journal of the American Statistical Association*, 58, 1048–1053.

Horn, J. L. (1965). "A Rationale and Test for the Number of Factors in Factor Analysis," *Psychometrika*, 30, 179–186.

Hosmer, D. W., Jr., and S. Lemeshow (1989). *Applied Logistic Regression*, Wiley, New York.

Huberty, C. J. (1984). "Issues in the Use and Interpretation of Discriminant Analysis," *Psychological Bulletin*, 95 (1), 156–171.

Jackson, J. E. (1991). *A User's Guide to Principal Components*, Wiley, New York.

Johnson, N. and D. Wichern (1988). *Applied Multivariate Statistical Analysis*, Prentice-Hall, Englewood Cliffs, N.J.

Jöreskog, K. G. and D. Sörbom (1989). *Lisrel 7: A Guide to the Program and Applications*, SPSS Inc., Chicago.

Kaiser, H. F. (1970). "A Second Generation Little Jiffy," *Psychometrika*, 35 (December), 401–415.

Kaiser, H. F. and J. Rice (1974). "Little Jiffy Mark IV," *Educational and Psychological Measurement*, 34 (Spring), 111–117.

Kenny, D., and C. Judd (1986). "Consequences of Violating the Independent Assumption in Analysis of Variance," *Psychological Bulletin*, 99, 422–431.

Lachenbruch, P. A. (1967). "An Almost Unbiased Method of Obtaining Confidence Intervals for the Probability of Misclassification in Discriminant Analysis," *Biometrics*, 23, 639–645.

Lachenbruch, P. A., C. Sneeringer, and L. T. Revo (1973). "Robustness of the Linear and Quadratic Discriminant Function to Certain Types of Non-normality," *Communications in Statistics*, 1, 39–57.

Long, S. J. (1983). *Confirmatory Factor Models*, Sage, Beverly Hills, Calif.

Maiti, S. S. and B. N. Mukherjee (1990). "A Note on Distributional Properties of the Jöreskog and Sörbom Fit Indices," *Psychometrika*, 55 (December), 721–726.

Mardia, K. V. (1971). "The Effect of Non-normality on Some Multivariate Tests and Robustness to Non-normality in the Linear Model," *Biometrika*, 58, 105–121.

Marks, S. and O. J. Dunn (1974). "Discriminant Functions when Covariance Matrices are Unequal," *Journal of the American Statistical Association*, 69, 555–559.

Marsh, H. W., J. R. Balla, and R. McDonald (1988). "Goodness-of-Fit Indexes in Confirmatory Factor Analysis: The Effects of Sample Size," *Psychological Bulletin*, 103, 391–410.

McDonald, R. (1985). *Factor Analysis and Related Techniques*, Lawrence Erlbaum, Hillsdale, N.J.

McDonald, R. and H. W. Marsh (1990). "Choosing a Multivariate Model: Noncentrality and Goodness of Fit," *Psychological Bulletin*, 105, 430–445.

McDonald, R. and S. A. Mulaik (1979). "Determinacy of Common Factors: A Nontechnical Review," *Psychological Bulletin*, 86, 297–306.

McIntyre, R. M. and R. K. Blashfield (1980). "A Nearest-Centroid Technique for Evaluating the Minimum-Variance Clustering Procedure," *Multivariate Behavioral Research*, 15, 225–238.

McLachlan, G. J. (1974). "An Asymptotic Unbiased Technique for Estimating the Error Rates in Discriminant Analysis," *Biometrics*, 30, 239–249.

Milligan, G. W. (1980). "An Examination of the Effect of Six Types of Error Perturbation of Fifteen Clustering Algorithms," *Psychometrika*, 45, 325–342.

Milligan, G. W. (1981). "A Montecarlo Study of Thirty Internal Criterion Measures for Cluster Analysis," *Psychometrika*, 46, 325–342.

Milligan, G. W. (1985). "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika*, 50, 159–179.

Olson, C. L. (1974). "Comparative Robustness of Six Tests in Multivariate Analysis of Variance," *Journal of the American Statistical Association*, 69 (348), 894–907.

Punj, G. and D. W. Stewart (1983). "Cluster Analysis in Marketing Research: Review and Suggestions for Application," *Journal of Marketing Research*, 20 (May), 134–148.

Rummel, R. J. (1970). *Applied Factor Analysis*, Northwestern Univ. Press, Evanston, Ill.

SAS Institute Inc. (1993). *SAS/STAT User's Guide*, Vol. 1, version 6.

Scariano, S. and J. Davenport (1986). "The Effects of Violations of the Independence Assumption in the One Way ANOVA," *The American Statistician*, 41, 123–129.

Segal, S. (1967). *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York.

Sharma, S., S. Durvasula, and W. R. Dillon (1989). "Some Results on the Behavior of Alternate Covariance Structure Estimation Procedures in the Presence of Non-Normal Data," *Journal of Marketing Research*, 26, 214–221.

Shimp, T. A. and A. Kavas (1984). "The Theory of Reasoned Action Applied to Coupon Usage," *Journal of Consumer Research*, 11 (December), 795–809.

Shimp, T. A. and S. Sharma (1987). "Consumer Ethnocentrism: Construction and Validation of the CETSCALE," *Journal of Marketing Research*, 24 (August), 280–289.

Sneath, P. and R. Sokal (1973). *Numerical Taxonomy*, Freeman, San Francisco.

Sobel, M. F. and G. W. Bohrnstedt (1985). "Use of Null Models in Evaluating the Fit of Covariance Structure Models," in N. B. Tuma (ed.), *Sociological Methodology*, 152–178. Jossey-Bass, San Francisco.

SOLO Power Analysis (1992). Version 1.0, BMDP Statistical Software Inc., Los Angeles.

Sparks, D. L. and W. T. Tucker (1971). "A Multivariate Analysis of Personality and Product Use," *Journal of Marketing Research*, 8 (February), 67–70.

Spearman, C. (1904). "'A General Intelligence Objectivity' Objectively Determined and Measured," *American Journal of Psychology*, 15, 201–293.

Stevens, S. S. (1946). "On the Theory of Scales of Measurement," *Science*, 103, 677–680.
Stewart, D. K. and W. A. Love (1968). "A General Canonical Correlation Index," *Psychological Bulletin*, 70, 160–163.
Stewart, D. W. (1981). "The Application and Misapplication of Factor Analysis in Marketing Research," *Journal of Marketing Research*, 18 (February), 51–62.
Toussant, G. T. (1974). "Bibliography on Estimation of Misclassification," *IEEE Transactions on Information Theory*, IT-20 (July), 472–479.
Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, Mass.
Urban, G. and J. A. Hauser (1993). *Design and Marketing of New Products*, Prentice-Hall, N.J.
Velleman, P. F. and L. Wilkinson (1993). "Nominal, Ordinal, Interval, and Ratio Typologies Are Misleading," *The American Statistician*, 47 (1), 65–72.
Werts, C. E., R. L. Linn, and K. G. Jöreskog (1974). "Intraclass Reliability Estimates: Testing Structural Assumptions," *Educational and Psychological Measurement*, 34, 25–33.
Wilk, H. B., S. S. Shapiro, and H. J. Chen (1968). "A Comparative Study of Various Tests of Normality," *Journal of the American Statistical Association*, 63, 1343–1372.
Winer, B. J. (1982). *Statistical Principles in Experimental Design*, 2nd ed., McGraw-Hill, New York.
Zwick, W. R. and W. F. Velicer (1986). "Comparison of Five Rules for Determining the Number of Components to Retain," *Psychological Bulletin*, 99 (3), 432–442.

Tables, Figures, and Exhibits

CHAPTER 1

Tables

1.1 Dependence Statistical Methods 6
1.2 Independent Variables Measured Using Nominal Scale 7
1.3 Attributes and Their Levels for Checking Account Example 9
1.4 Interdependence Statistical Methods 11
1.5 Contingency Table 12

Figures

1.1 Causal model 13
1.2 Causal model for unobservable constructs 13

CHAPTER 2

Figures

2.1 Points represented relative to a reference point 18
2.2 Change in origin and axes 18
2.3 Euclidean distance between two points 19
2.4 Vectors 20
2.5 Relocation or translation of vectors 20
2.6 Scalar multiplication of a vector 21
2.7 Vector addition 21
2.8 Vector subtraction 22
2.9 Vector projections 22
2.10 Vectors in a Cartesian coordinate system 23
2.11 Trigonometric functions 24
2.12 Length and direction cosines 24
2.13 Standard basis vectors 25
2.14 Linear combinations 26
2.15 Distance and angle between any two vectors 27
2.16 Geometry of vector projections and scalar products 28
2.17 Projection of a vector onto a subspace 29
2.18 Illustrative example 29
2.19 Change in basis 31
2.20 Representing points with respect to new axes 32

CHAPTER 3

Tables

3.1 Hypothetical Financial Data 37
3.2 Contingency Table 37
3.3 Hypothetical Financial Data for Groups 40
3.4 Transposed Mean-Corrected Data 48

Figures

3.1 Distribution for random variable 43
3.2 Hypothetical scatterplot of a bivariate distribution 44
3.3 Plot of data and points as vectors 45
3.4 Mean-corrected data 46
3.5 Plot of standardized data 47
3.6 Plot of data in observation space 49
3.7 Generalized variance 50

CHAPTER 4

Tables

4.1 Original, Mean-Corrected, and Standardized Data 59
4.2 Mean-Corrected Data and New Variable (x_1^*) for a Rotation of 10° 61
4.3 Variance Accounted for by the New Variable x_1^* for Various New Axes 61
4.4 Mean-Corrected Data, and x_1^* and x_2^* for the New Axes Making an Angle of 43.261° 62
4.5 SAS Statements 67
4.6 Standardized Principal Components Scores 70
4.7 Food Price Data 71
4.8 Regression Coefficients for the Principal Components 78
A4.1 PROC IML Commands 88

Figures

4.1 Plot of mean-corrected data and projection of points onto X_1^* 60
4.2 Percent of total variance accounted for by X_1^* 62
4.3 Plot of mean-corrected data and new axes 63
4.4 Representation of observations in lower-dimensional subspace 65
4.5 Scree plots 77
4.6 Plot of principal components scores 80

Exhibits

4.1 Principal components analysis for data in Table 4.1 69
4.2 Principal components analysis for data in Table 4.7 73
4.3 Principal components analysis on standardized data 74
A4.1 PROC IML output 89

CHAPTER 5

Tables

5.1 Communalities, Pattern and Structure Loadings, and Correlation Matrix for One-Factor Model 93

5.2 Communalities, Pattern and Structure Loadings, and Correlation Matrix for Two-Factor Model 95
5.3 Communalities, Pattern and Structure Loadings, Shared Variances, and Correlation Matrix for Alternative Two-Factor Model 98
5.4 Summary of Principal Components Factor Analysis for the Correlation Matrix of Table 5.2 105
5.5 Reproduced and Residual Correlation Matrices for PCF 106
5.6 Iteration History for Principal Axis Factor Analysis 108
5.7 SAS Commands 109
5.8 List of Attributes 123
5.9 Correlation Matrix for Detergent Study 124
5.10 SPSS Commands 125
A5.1 Varimax Rotation of 350° 139
A5.2 Variance of Loadings for Varimax Rotation 139
A5.3 Varimax Rotation of 320.057° 140

Figures

5.1 Relationship between grades and intelligence 91
5.2 Two-factor model 94
5.3 Two-indicator two-factor model 99
5.4 Indeterminacy due to to estimates of communalities 100
5.5 Projection of vectors onto a two-dimensional factor space 101
5.6 Rotation of factor solution 101
5.7 Factor solution 102
5.8 Scree plot and plot of eigenvalues from parallel analysis 104
5.9 Confirmatory factor model for excellence 129
A5.1 Oblique factor model 140
A5.2 Pattern and structure loadings 141

Exhibits

5.1 Principal components analysis for the correlation matrix of Table 5.2 103
5.2 Principal axis factoring for the correlation matrix of Table 5.2 110
5.3 Quartimax rotation 121
5.4 SPSS output for detergent study 126

CHAPTER 6

Tables

6.1 Symbols Used by LISREL To Represent Parameter Matrices 149
6.2 Correlation Matrix 149
6.3 LISREL Commands for the One-Factor Model 150
6.4 LISREL Commands for the Null Model 161
6.5 Computations for *NCP, MDN, TLI,* and *RNI* for the One-Factor Model 161
6.6 LISREL Commands for the Two-Factor Model 166
6.7 Computations for *NCP, MDN, TLI,* and *RNI* for the Correlated Two-Factor Model 170
6.8 SPSS Commands for Multigroup Analysis 172
6.9 Results of Multigroup Analysis: Testing Factor Structure for Males and Females 173

6.10 Items or Statements for the 10-item CET Scale 174
Q6.1 Hypothetical Correlation Matrix 178
A6.1 Value of the Likelihood Function for Various Values of p 182
A6.2 Maximum Likelihood Estimate for the Mean of a Normal Distribution 183

Figures

6.1 One-factor model 145
6.2 Two-factor model with correlated constructs 147
6.3 EGFI as a function of the number of indicators and sample size 158
6.4 Two-factor model 165
Q6.1 Model 178
Q6.2 Models 179
A6.1 Maximum likelihood estimation procedure 182
A6.2 Maximum likelihood estimation for mean of a normal distribution 184

Exhibits

6.1 LISREL output for the one-factor model 153
6.2 LISREL output (partial) for the two-factor model 167
6.3 Two-factor model with correlated constructs 168
6.4 LISREL output for the 10-item CETSCALE 175

CHAPTER 7

Tables

7.1 Hypothetical Data 186
7.2 Similarity Matrix Containing Euclidean Distances 188
7.3 Centroid Method: Five Clusters 189
7.4 Centroid Method: Four Clusters 189
7.5 Centroid Method: Three Clusters 190
7.6 Ward's Method 194
7.7 SAS Commands 194
7.8 Within-Group Sum of Squares and Degrees of Freedom for Clusters Formed in Steps 1, 2, 3, 4, and 5 199
7.9 Summary of the Statistics for Evaluating Cluster Solution 201
7.10 Initial Cluster Centroids, Distance from Cluster Centroids, and Initial Assignment of Observations 204
7.11 Centroid of the Three Clusters and Change in Cluster Centroids 204
7.12 Distance from Centroids and First Reassignment of Observations to Clusters 204
7.13 Initial Assignment, Cluster Centroids, and Reassignment 206
7.14 Initial Assignment 206
7.15 Change in ESS Due to Reassignment 207
7.16 SAS Commands for Nonhierarchical Clustering 207
7.17 Observations Selected as Seeds for Various Combinations of Radius and Replace Options 208
7.18 RS and RMSSTD for 2-, 3-, 4-, and 5-Cluster Solutions 210
7.19 Food Nutrient Data 222
7.20 Cluster Membership for the Four-Cluster Solution 227
7.21 Cluster Centers for Hierarchical Clustering of Food Nutrient Data 227

7.22 Commands for FASTCLUS Procedure 228
7.23 Correlation Matrix 232
A7.1 Using a Nonhierarchical Clustering Technique to Refine a Hierarchical Cluster Solution 235

Figures

7.1 Plot of hypothetical data 186
7.2 Dendogram for hypothetical data 190
7.3 Plots of: (a) SPR and RS and (b) RMSSTD and CD 201
7.4 Hypothetical cluster configurations 217
7.5 City-block distance 218
7.6 Cluster analysis plots. (*a*) R square. (*b*) RMSSTD. 226

Exhibits

7.1 SAS output for cluster analysis on data in Table 7.1 195
7.2 Nonhierarchical clustering on data in Table 7.1 209
7.3 Empirical comparisons of the performance of clustering algorithms 212
7.4 Hierarchical cluster analysis for food data 223
7.5 Nonhierarchical analysis for food-nutrient data 229

CHAPTER 8

Tables

8.1 Financial Data for Most-Admired and Least-Admired Firms 238
8.2 Summary Statistics for Various Linear Combinations 240
8.3 Discriminant Score and Classification for Most-Admired and Least-Admired Firms (w_1 = .934 and w_2 = .358) 243
8.4 Means, Standard Deviations, and *t*-values for Most- and Least-Admired Firms 245
8.5 SPSS Commands for Discriminant Analysis of Data in Table 8.1 246
8.6 Misclassification Costs 257
8.7 Classification Based on Mahalanobis Distance 259
8.8 Discriminant Scores, Classification, and Posterior Probability for Unequal Priors 262
8.9 Financial Data for Most-Admired and Least-Admired Firms 267
8.10 SPSS Commands for Stepwise Discriminant Analysis 268
8.11 Correlation Matrix for Discriminating Variables 272
8.12 SPSS Commands for Holdout Validation 274
A8.1 Misclassification Costs 279
A8.2 Summary of Classification Rules 281
A8.3 Illustrative Example 285
A8.4 TCM for Various Combinations of Misclassification Costs and Priors 286

Figures

8.1 Plot of data in Table 8.1 and new axis 239
8.2 Distributions of financial ratios 239
8.3 Plot of lambda versus theta 241
8.4 Examples of linear combinations 242

8.5 Plot of discriminant scores 244
A8.1 Classification in one-dimensional space 278
A8.2 Classification in two-dimensional space 279
A8.3 Density functions for one discriminating variable 280
A8.4 TCM as a function of cutoff value 286

Exhibits

8.1 Discriminant analysis for most-admired and least-admired firms 247
8.2 Multiple regression approach to discriminant analysis 263
8.3 Stepwise discriminant analysis 269

CHAPTER 9

Tables

9.1 Hypothetical Data for Four Groups 290
9.2 Lambda for Various Angles between Z and X_1 291
9.3 SPSS Commands 294
9.4 Cases in Which Wilks' Λ Is Exactly Distributed as F 298
9.5 SPSS Commands for Range Tests 300
9.6 SPSS Commands for the Beer Example 304
A9.1 Illustrative Example 312
A9.2 Conditions and Equations for Classification Regions 316

Figures

9.1 Hypothetical scatter plot 288
9.2 Plot of data in Table 9.1 290
9.3 Plot of rotation angle versus lambda 291
9.4 Classification in variable space 292
9.5 Classification in discriminant space 292
9.6 Plot of brands in discriminant space 308
9.7 Plot of brands and attributes 308
A9.1 Classification regions for three groups 314
A9.2 Group centroids 315
A9.3 Classification regions R_1 to R_4 316

Exhibits

9.1 Discriminant analysis for data in Table 9.1 295
9.2 Range tests for data in Table 9.1 301
9.3 SPSS output for the beer example 305
9.4 Range tests for the beer example 306

CHAPTER 10

Tables

10.1 Data for Most-Successful and Least-Successful Financial Institutions 318
10.2 Contingency Table for Type and Size of Financial Institution 318
10.3 SAS Commands for Logistic Regression 321
10.4 Classification Table 326

10.5 SAS Commands for Stepwise Logistic Regression 329
10.6 Classification Table for Cutoff Value of 0.5 332
A10.1 Values of the Maximum Likelihood Function for Different Values of β_0 and β_1 340

Figure

10.1 The logistic curve 320

Exhibits

10.1 Logistic regression analysis with one categorical variable as the independent variable 322
10.2 Contingency analysis output 328
10.3 Logistic regression for categorical and continuous variables 330
10.4 Discriminant analysis for data in Table 10.1 333
10.5 Logistic regression for mutual fund data 334

CHAPTER 11

Tables

11.1 Cell Means 344
11.2 MANOVA Computations 347
11.3 SPSS Commands 351
11.4 Hypothetical Data To Illustrate the Presence of Multivariate Significance in the Absence of Univariate Significance 353
11.5 Data for Drug Effectiveness Study 355
11.6 SPSS Commands for Drug Study 355
11.7 Coefficients for the Contrasts 358
11.8 SPSS Commands for Helmert Contrasts 360
11.9 Coefficients for Correlated Contrasts 363
11.10 SPSS Commands for Correlated Contrasts 364
11.11 Summary of Significant Tests 364
11.12 Data for the Ad Study 366
11.13 SPSS Commands for the Ad Study 367
11.14 Cell Means for Multivariate *GENDER* × *AD* Interaction 369

Figures

11.1 One dependent variable and one independent variable at two levels 343
11.2 Two dependent variables and one independent variable at two levels 343
11.3 More than one independent variable and two dependent variables 345
11.4 Presence of multivariate significance in the absence of univariate significance 354
11.5 *GENDER* × *AD* interaction 370

Exhibits

11.1 MANOVA for most-admired and least-admired firms 352
11.2 Multivariate significance, but no univariate significance 354
11.3 MANOVA for drug study 356
11.4 Helmert contrasts for drug study 361

11.5 SPSS output for correlated contrasts using the sequential method 365
11.6 MANOVA for ad study 368

CHAPTER 12

Tables

12.1 Hypothetical Data Simulated from Normal Distribution 376
12.2 Financial Data for Most-Admired and Least-Admired Firms 379
12.3 SPSS Commands 379
12.4 Ordered Squared Mahalanobis Distance and Chi-Square Value 381
12.5 Transformations To Achieve Normality 383
12.6 Data for Purchase Intention Study 385

Figures

12.1 Q-Q plot for data in Table 12.1 377
12.2 Q-Q plot for transformed data 378
12.3 Chi-square plot for total sample 382
12.4 Chi-square plot for ad awareness data 386

Exhibits

12.1 Univariate normality tests for data in Table 12.1 380
12.2 Partial MANOVA output for checking equality of covariance matrices assumption 387
12.3 Partial MANOVA output for checking equality of covariance matrices assumption for transformed data 388

CHAPTER 13

Tables

13.1 Hypothetical Data 392
13.2 Correlation between Various New Variables 394
13.3 Variables W_1 and V_1 396
13.4 SAS Commands for the Data in Table 13.1 398
Q13.1 Correlation Matrix: Product Use and Personality Trait 411
Q13.2 Results of the Canonical Analysis 412
Q13.3 Indicators of Canonical Association between Measures of Insurers' Power and Insurers' Sources 413
A13.1 PROC IML Commands for Canonical Correlation Analysis 417

Figures

13.1 Plot of predictor and criterion variables 393
13.2 New axes for Y and X variables 395
13.3 Geometrical illustration in subject space 397

Exhibits

13.1 Canonical correlation analysis on data in Table 13.1 399
13.2 Canonical correlation analysis for nutrition information study 407
A13.1 PROC IML output for canonical analysis 418

CHAPTER 14

Tables

14.1 Representation of Parameter Matrices of the Structural Model in LISREL 421
14.2 Hypothetical Covariance Matrix for the Model Given in Figure 14.2 422
14.3 LISREL Commands for the Model Given in Figure 14.2 422
14.4 Summary of Total, Direct, and Indirect Effects 426
14.5 Representation of Parameter Matrices of the Structural Model with Unobservable Constructs in LISREL 428
14.6 LISREL Commands for Structural Model with Unobservable Constructs 429
14.7 Summary of the Results for Structural Model with Unobservable Constructs 434
14.8 Goodness-of-Fit Indices for the Coupon Usage Model 438
14.9 Summary of the Results for the Respecified Coupon Usage Model 438

Figures

14.1 Structural or path model 420
14.2 Structural model for observable constructs 420
14.3 Structural model with unobserved constructs 427
14.4 Coupon usage model 436
A14.1 Structural model with observable constructs 446
A14.2 Structural model with unobservable constructs 447
A14.3 Structural model 450
A14.4 Indirect effects of length three 451
A14.5 Multiple indirect effects 451

Exhibits

14.1 LISREL output for the covariance matrix given in Table 14.2 423
14.2 LISREL output for structural model with unobservable constructs 431
14.3 LISREL output for coupon usage model 437
A14.1 Covariance matrix for structural model with observable constructs 447
A14.2 Covariance matrix for structural model with unobservable constructs 449

Index

A

Adjusted goodness-of-fit index, LISREL, 159
Akaike's information criteria, 324
Alpha factor analysis, 109
Analysis of variance (ANOVA)
  monotonic analysis of variance (MONANOVA), 8–9
  multivariate analysis of variance (MANOVA), 10
  with one dependent/more than one independent variable, 7
  situations for use, 7
ANOVA, *see* Analysis of variance (ANOVA)
Association coefficients, in cluster analysis, 220
Assumptions
  equality of covariance matrices assumption, 383–386
  independence assumption, 387–388
  normality assumptions, 375
Average-linkage method, hierarchical clustering method, 192–193
Axes, in Cartesian coordinate system, 17–19

B

Backward selection, stepwise discriminant analysis, 265
Bartlett's test, 76
  purpose of, 123
  sensitivity of, 123
Basis vectors, 25, 31
Bayesian theory
  objective of, 256
  posterior possibilities based on, 281
Bernoulli trial, 340
Between-group analysis, 41–42
  sum of squares and cross products matrix, 42
BIOMED
  clustering routines, 220
  structural model estimation, 426
Bootstrap method, discriminant function validation, 274
Box' M
  checking equality of covariance matrices, 384–386
  in multiple-group MANOVA, 351, 356

C

Canonical correlation
  analytic approach to, 397–398
  canonical variates, 401–402, 404
  change in scale, effect of, 415
  computer analysis, 398–406
  examples of use, 406–409, 412–418
  external validity of, 409
  as general technique, 409
  geometric view of, 391–397
  with more than one dependent/one or more independent variables, 9
  practical significance of, 404–406
  situations for use, 9, 391
  statistical significance tests for, 402–404
Canonical discriminant function, 251
  standardized, 253–254
Cartesian coordinate system, 17–19
  change in origin and axes, 18–19
  Euclidian distance, 19
  origin and axes in, 17–19
  rectangular Cartesian axes, 17
  representation of points, 17–18
  vectors in, 23–25

Central tendency measures, mean, 36
Centroid method, hierarchical clustering method, 188–191
Chaining effect, in hierarchical clustering methods, 211, 217
Chi-square difference test, 439
Chi-square goodness of fit test, 378
Chi-square plot, 381–382
  computer program for, 389–390
Classification
  classification function method, 257–258
  classification matrix, 255–256
  classification rate, evaluation of, 258–260
  computer analysis, 256–257, 261
  cutoff-value method, 255–256
  in discriminant analysis, 242–244, 278–284
  as independent procedure, 242, 244
  in logistic regression, 326–327
  Mahalanobis distance method, 258
  misclassification errors, 256–257, 261, 311–312
  for more than two groups, 311–312
  multiple-group discriminant analysis, 293, 303–304, 311–312, 313
  multivariate normal distributions, rules for, 281–283
  practical significance of, 260
  statistical decision theory, 256–257, 279–281
  statistical tests used, 258, 260
  total probability of misclassification, 280
Cluster analysis
  average-linkage method, 192–193
  centroid method, 188–191
  comparison of hierarchical/nonhierarchical methods, 211–217
  complete-linkage or farthest-neighbor method, 192
  computer analysis of, 193–202
  dendrogram in, 190–191
  examples of, 221–232
  external validity of solution, 221
  geometrical view of, 186–187
  hierarchical clustering methods, 188–193
  loss of homogeneity in, 200
  nonhierarchical clustering, 202–211
  objective of, 187
  Q-factor analysis, 187
  reliability of solution, 221
  root-mean-square total-sample standard deviation, 197, 198
  R-squared, 198, 200
  semipartial R-squared, 198, 200
  similarity measures, 187–188, 218–220
  single-linkage or nearest-neighbor method, 191
  situations for use, 12, 185
  Ward's method, 193
Common factors, 96, 108
Communality, 92
Communality estimation problem, and factor analysis, 136
Complete-linkage method, hierarchical clustering method, 192, 217
Computer programs, *see* BIOMED; Statistical Analysis System (SAS); Statistical Package for the Social Sciences (SPSS)
Concordant pair, 325, 326
Confirmatory factor analysis, 128
  LISREL, 148–177
  objectives of, 148
  situations for use, 144
Confusion matrix, 255–256
Conjoint analysis
  monotonic analysis of variance (MONANOVA), 8–9
  with one dependent/more than one independent variable, 8–9
Constrained analysis, LISREL, 171, 173
Contingency table analysis, in logistic regression, 327–328
Contrasts
  computer analysis, 360–366
  correlated contrasts, 363–366
  Helmert contrasts, 360–361
  multivariate significance tests for, 359–360, 363
  orthogonal contrasts, 357–363
  univariate significance tests for, 357–359, 360, 362–363
Correlated contrasts, in multiple-group MANOVA, 363–366
Correlation coefficient
  in cluster analysis, 220
  for standardized data, 39
Correlation matrix
  in confirmatory factor analysis, 144–145
  use in, 144–145
Correspondence analysis, situations for use, 12
Covariance matrix
  equality of covariance matrices assumption, 383–387
  and factor analysis, 144–145
  one-factor model with, 145–147
Cutoff-value method, 244, 255–256

D

Data analytic methods
dependence methods, 4–10
interdependence methods, 4, 10–12
structural models, 13–14
Data manipulations
computer procedures for, 55–57
degrees of freedom, 36–38
generalized variance, 39
group analysis, 40–42
mean, 36
mean-corrected data, 36
standardization of data, 39
sum of cross products, 39
sum of squares, 38–39
variance, 38
Degrees of freedom, 36–38
computation of, 37–38
Dendrogram, for clustering process, 190–191
Dependence methods, 4–10
analysis of variance, 7
canonical correlation, 9
conjoint analysis, 8–9
discrete discriminant analysis, 8
discrete multiple-group discriminant analysis, 10
discriminant analysis, 7–8
logistic regression, 8
for more than one dependent/more than one independent variables, 9–10
multiple-group discriminant analysis, 10
multivariate analysis of variance, 10
for one dependent/more than one independent variable, 5–9
for one dependent/one independent variable, 5, 6
regression, 5
situations for use of, 4
Detrended normal plot, 378–379
Dimensional reduction, principal components analysis as, 64–65
Direction cosines, 25
Discordant pair, 325
Discrete discriminant analysis, with one dependent/more than one independent variable, 8
Discrete multiple-group discriminant analysis
with more than one dependent/one or more independent variables, 10
situations for use, 10
Discriminant analysis, *see also* Multiple-group discriminant analysis; Two-group discriminant analysis
compared with logistic regression, 332–333
discrete discriminant analysis, 8
discrete multiple-group discriminant analysis, 10
multiple-group discriminant analysis, 10
compared to multivariate analysis of variance (MANOVA), 350
with one dependent/more than one independent variable, 7–8
situations for use, 8, 237
stepwise discriminant analysis, 246, 264–273
Discriminant function, multiple-group discriminant analysis, 294–303
assessment of importance of, 303
computation options, 294, 297
estimate of, 297–299
group differences, examination of, 307
labeling of, 307
number needed, 299–303
practical significance of, 302–303
statistical significance of, 299–302
Discriminant function, two-group discriminant analysis, 250–254
bootstrap method for validation of, 274
canonical discriminant function, 251
computation options, 250–251
estimate of, 251–252
holdout method for validation of, 273
linear discriminant function, 242
meaning of, 242
practical significance of, 253
standardized canonical discriminant function, 253–254
statistical significance of, 252–253
U-method for validation of, 273–274
Discriminant score, meaning of, 242
Discriminant variables, 238
assessment of importance of, 253–254
Distance measures, 218–220
Euclidian distance, 19, 219
Mahalanobis distance, 44–45, 220
Minkowski distance, 218
statistical distance, 42–44
Distinguishability of observations, 45

E

Effect size
in MANOVA, 348–349
multivariate, 349
univariate, 349
Eigenstructure of covariance matrix, 84–85
computer analysis, 87–89

Equality of covariance matrices
assumption, 383–387
errors and violation of, 383–384
tests for checking equality, 384–387
Equivalent vectors, 20
Error sums of squares, 193
Euclidian distance, 19
in cluster analysis, 219
in similarity matrix, 187–188
for standardized data, 219
and statistical distance, 43–44
Event, 325
Exploratory factor analysis, 128
situations for use, 144

F

Factor analysis, *see also* Confirmatory factor analysis
alpha factor analysis, 109
appropriateness of data for, 116, 123, 125
choosing technique for, 108
common factors, 96, 108
communalities problem, estimation of, 100
and communality estimation problem, 136
computer estimation of, 109–115
concepts/terms related to, 90–93
confirmatory factor analysis, 128
exploratory factor analysis, 128
factor extraction methods, 141–142
factor indeterminacy, 97–98, 136
factor rotation problem, 97, 100–102, 136
factor rotations, types of, 137–141
factor scores, 96, 142–143
factor solution, 117–118
fundamental factor analysis equation, 136
geometric view of, 99–102
image analysis, 109
interpretation of factor structure, 125
model with more than two factors, 96, 102, 135–136
number of factors needed, 116–117
objectives of, 99
one-factor model, 93, 132–133
principal axis factoring (PAF), 107, 142
compared to principal components analysis, 125–128
principal components factoring (PCF), 103–107, 141–142
representation of factors, 118–119
situations for use, 11, 90
two-factor model, 93–96, 133–135
Factor extraction methods, 141–142
principal axis factoring, 142
principal components factoring, 141–142
Factor rotation problem, 97, 100–102, 136
basis of, 97
geometric view of, 100–102
Factor rotations, 137–141
oblique rotation, 140–141
orthogonal rotation, 137
quartimax rotation, 120–121, 137
varimax rotation, 119–120, 138, 139–140
Factor scores, 96, 142–143
Farthest-neighbor method, hierarchical clustering method, 192, 217
F-distribution
in multiple-group discriminant analysis, 294
table of, 460–465
Fisher's linear discriminant function, 245, 277–278
computation of, 277–278
Fisher's Z transformation, 383
Forward selection, stepwise discriminant analysis, 265
F-ratio
in multiple-group discriminant analysis, 294, 301
in multiple-group MANOVA, 351
relationship to Wilks'Λ, 348
in stepwise discriminant analysis, 266, 271
F-test, in multiple-group discriminant analysis, 293, 294
Fundamental factor analysis equation, 136

G

Generalized variance, 39, 50–51
equality to determinant of covariance matrix, 54–55
geometric representation of, 50–51
Geometric concepts
Cartesian coordinate system, 17–19
vectors, 19–33
Goodness-of-fit measures
chi-square, 378
LISREL, 157–159
in logical regression, 324
Gram-Schmidt orthonormalization procedure, 32
Group analysis, 40–42
between-group analysis, 41–42
within-group analysis, 40–41

H

Helmert contrasts, 360–361
Heuristic measures, of model fit, 157–160
Hierarchical clustering methods, 188–193
  average-linkage method, 192–193
  centroid method, 188–191
  chaining effect in, 211, 217
  complete-linkage or farthest-neighbor method, 192
  computer analysis, 193–202
  evaluation of, 211–217
  example of, 221–228
  single-linkage or nearest-neighbor method, 191
  Ward's method, 193
Holdout method, discriminant function validation, 273

I

Image analysis, 109
Implied covariance matrix, 444–449
  matrix algebra, 445–446
  models with observable constructs, 444–446
  models with unobservable constructs, 446–449
Independence assumption, 387–388
  lack of tests for, 388
Indicator, 91
Interdependence methods, 4, 10–12
  cluster analysis, 12
  correspondence analysis, 12
  factor analysis, 11
  loglinear models, 12
  principal components analysis, 11
  situations for use of, 4, 10
Interval scale, use of, 2–3

K

Kaiser-Meyer-Olkin measure, 116
K-means clustering, 205
Kolmogorov-Smirnov test, 378, 379
Kurtosis
  and leptokurtic distribution, 375
  normalization of, 375
  of univariate normal distribution, 375

L

Latent constructs
  meaning of, 13
  and structural models, 13, 14
Latent factor, 91
Leptokurtic distributions, nature of, 375
Linear combination, vectors, 26–27
Linear discriminant function, 242
  Fisher's, 245, 277–278
LISREL, 148–177
  adjusted goodness-of-fit index, 159
  commands, 150–152
  constrained analysis, 171, 173
  estimated model parameters, evaluation of, 162–164
  example of use, 174–178
  goodness-of-fit indices, 157–159
  initial estimates, 152–157
  maximum likelihood estimates, 422–424
  McDonald's transformation of the noncentrality parameter (MDN), 159–160
  model fit, evaluation of, 157–162
  model information and parameter specifications, 152
  model respecification, 164–165
  modification indices, 164
  multigroup analysis, 170–173
  null hypothesis test, 157
  null model, 160, 161
  one-factor model, 153–156
  parameter estimates, 162–163
  relative goodness-of-fit index, 159
  relative noncentrality index, 160
  rescaled noncentrality parameter, 159, 160
  residual matrix, 160–162
  root mean square residual, 159
  squared multiple correlation, 163–164
  structural model in, 421–434
  terminology related to, 148–149
  total coefficient of determination, 164
  Tucker-Lewis index, 160
  two-factor model, 165–170
  unconstrained analysis, 171
Loadings, 404
Logistic regression
  classification, 326–327
  with combination categorical/continuous independent variables, 328–332
  compared with discriminant analysis, 332–333
  computer analysis, 321–335
  contingency table analysis, 327–328
  example as illustration of, 333–335
  example of use, 333–335
  logistic regression model, 319–321

Logistic regression (*continued*)
maximum likelihood estimation procedure in, 321, 324–325, 339–341
model fit, assessment of, 323–324
model information, 321
multiple logistic regression, 320
with one categorical variable, 321–327
with one dependent/more than one independent variable, 8
parameter estimates, 324–325
predicted probabilities, association of, 325–326
probability and odds in, 317–321
situations for use, 8, 317
stepwise selection procedure, 329, 331–332
Logit function, 320
Logit transformation, 383
Loglinear models, situations for use, 12
Loss of homogeneity, in cluster analysis, 200

M

Mahalanobis distance, 44–45
as classification method, 258
in cluster analysis, 220
definition of, 44
formula for, 44
in MANOVA, 343
squared distance in stepwise discriminant analysis, 266
MANOVA, *see* Multivariate analysis of variance (MANOVA)
Maximum likelihood estimation technique
computation of, 181–185
computer analysis, 148–173
in LISREL, 422–424
in logistic regression, 321, 324–325, 339–341
McDonald's transformation of the noncentrality parameter (MDN), 159–160
Mean, computation of, 36
Mean-corrected data, nature of, 36
Measure of I, 91
Measurement model, 14
Measurement scales
interval scale, 2–3
nominal scale, 2
and number of variables, 3–4
ordinal scale, 2
ratio scale, 3
Minimum average partial correlation (MAP), 117
Minkowski distance, 218
Modification indices, LISREL, 164
Monotonic analysis of variance (MONANOVA), situations for use, 8–9
Multicollinearity
population-based, 273
sample-based, 273
and stepwise discriminant analysis, 272–273
Multigroup analysis, with LISREL, 170–173
Multiple-group discriminant analysis
analytical approach to, 293–294
classification, 293, 303–304, 311–312, 313
computer analysis, 294–307
discrete multiple-group discriminant analysis, 10
discriminant function, 294–303
F-test in, 293, 294
geometric view of, 287–293
with more than one dependent/one or more independent variables, 10
multivariate normal distribution, 312–316
new axes, identification of, 289, 293
number of discriminant functions needed, 288–289
significance of variables, estimation of, 294
situations for use, 10
Multiple-group MANOVA, 355–366
computer analysis, 355–366
correlated contrasts, 363–366
multivariate effects, 356
orthogonal contrasts, 356–363
univariate effects, 356
Multiple regression
as canonical correlation, 9
discriminant analysis, 262–263
with one dependent/more than one independent variable, 5–9
Multivariate analysis
number of variables, 5
objectives of, 238
Multivariate analysis of variance (MANOVA)
see also Multiple-group MANOVA; Two-group MANOVA
analytic computations for, 346–350
computer analysis, 350–370
compared to discriminant analysis, 350
effect size, 348–350

geometric view of, 342–346
with more than one dependent/one or more independent variables, 10
multiple-group, 355–366
and multivariate effect size, 349–350
multivariate significance tests in, 346–348
with one independent/one dependent variable, 343
with one independent/p dependent variables, 344–346
with one independent/two or more dependent variables, 343–344
power of test in, 349–350
situations for use, 10, 342
two-group, 350–355
with two independent variables, 366–370
and univariate effect size, 349
univariate significance tests in, 348–349
Multivariate effect size, 349–350
Multivariate normal distributions
classification rules for, 281–283
multiple-group discriminant analysis, 312–316
skewness of, 375
two-group discriminant analysis, 281–283
Multivariate normality assumption, 8
and discriminant analysis, 263–264
Multivariate normality tests, 380–383
graphical test, 380–383
transformations, 383
Multivariate significance tests, 252
for contrasts, 359–360, 363
in multivariate analysis of variance (MANOVA), 346–348
in two-group MANOVA, 351, 353

N

Naive prediction rule, 260
Nearest-neighbor method, hierarchical clustering method, 191
Newton-Raphson method, 340
No-event, 325
Nominal scale, use of, 2, 7
Nonhierarchical clustering, 202–211
algorithms in, 203–207
cluster solution, evaluation/interpretation of, 210
computer analysis, 207–211
evaluation of, 217
example of, 228–232
K-means clustering, 205
method to obtain initial seeds, 202–203
reassignment rules, 203
steps in, 202
Normality assumptions, 375
Norm of vector, 20
Null hypothesis
and power of test, 349–350, 375
and Type I and Type II errors, 374–375
χ^2 statistic for testing of, 157, 162
Null model, 160
Null vector, 21

O

Oblique basis, vectors, 31
Oblique factor rotation, 140–141
Observation space, graphical representation of data in, 47–50
Odds, in logistic regression, 318–321
One-factor model
computation of, 132–133
with covariance matrix, 145–147
with LISREL, 153–156
situations for use, 93
Ordinal scale, use of, 2
Origin, in Cartesian coordinate system, 17–19
Orthogonal contrasts
computer analysis, 360–363
multiple-group MANOVA, 356–363
multivariate significance tests for, 359–360
situations for use, 357
univariate significance tests for, 357–359
Orthogonal factor model, 94
Orthogonal factor rotation, 137–140
Orthonormal vectors, 25, 31, 32

P

Parallel analysis, 77, 79
Parallelogram law of vector addition, 22
Pattern loadings, 91, 94
Pearson product moment correlation, 39
as similarity measure, 220
Percent points of normal probability plot correlation coefficient, table of, 466
Perceptual map, purpose of, 307
Population-based multicollinearity, 273
Power of test
in MANOVA, 350
purpose of, 349–350, 375

Principal axis factoring, 107
for factor extraction, 141
Principal components, 63, 66
Principal components analysis
algebraic approach to, 67–71
computer analysis, 67–71
as dimensional reducing technique, 64–65
eigenstructure of covariance matrix, 84–85
compared to factor analysis, 125–128
geometric view of, 59–66
goals of, 58, 66
interpretation of principal components, 79–80
issues related to use of, 71–81
number of components to extract, 76–79
and objective of study, 75–76
singular value decomposition, 85–86
situations for use, 11, 58
spectral decomposition of matrix, 86–87
compared to two-group discriminant analysis, 241–242
type of data, effect on analysis, 72–75
Principal components factoring, 103–107
for factor extraction, 141–142
Principal components scores, 63, 66
use of, 80
Probabilities, in logistic regression, 317–321
Projection vector, 23, 27–28
Pythagorean theorem, Euclidian distance computation, 19

Q

Q-factor analysis, situations for use, 187
Q-Q plot, 376–378
Quartimax factor rotation, 120–121, 137

R

Ratio scale, use of, 3
Ray's V, in stepwise discriminant analysis, 266
Rectangular Cartesian axes, 17
Reflection, of vectors, 21
Regression
logistic regression, 8
multiple regression, 5–9
simple regression, 5
Relative goodness-of-fit index, LISREL, 159
Relative noncentrality index, 160
Reliability, cluster analysis, 221
Rescaled noncentrality parameter, 159, 160
Residual matrix, LISREL, 160–162
Root-mean-square residual, 106–107, 118
LISREL, 159
Root-mean-square total-sample standard deviation
of the cluster, 198, 230
formula for, 197
R-squared, cluster analysis, 198, 200

S

Sample-based multicollinearity, 273
SAS, *see* Statistical Analysis System (SAS)
Saturated models, 421
Scalar product, of two vectors, 20–21, 27–28
Scale invariant, meaning of, 46
Schwartz's criterion, 324
Scree plot, 76–77
Scree plot test, 79
Semipartial R-squared, cluster analysis, 198, 200
Shapiro-Wilk test, 378, 379
Significance tests
for main effects, 370
MANOVA for two independent variables, 367–370
multivariate significance tests, 346–348, 351, 353
univariate significance tests, 348–349, 353–355
Similarity measures, 218–220
association coefficients, 220
correlation coefficient, 220
distance measures, 218–220
Simple regression, 5
Simulation percentiles of b_2, table of, 467
Simulation probability points of $\sqrt{b_1}$, table of, 467
Single-linkage method, hierarchical clustering method, 191
Singular value decomposition, 85–86
Skewness
of multivariate normal distribution, 375
of univariate normal distribution, 375
Space
observation space, 47–50
variable space, 45–46
Spectral decomposition of matrix, 86–87
SPSS, *see* Statistical Package for the Social Sciences (SPSS)

Squared multiple correlation
computation of, 181
LISREL, 163–164
Square-root transformation, 383
Standard basis vectors, 25
Standardization of data, 39
correlation coefficient, 39
Euclidian distance for standardized data, 219
graphical representations in space, 47–50
Standardized canonical discriminant function, 253–254
Standard normal probabilities, table of, 457
Statistical Analysis System (SAS)
canonical correlation, 398–406
chi-square plot, 389–390
cluster analysis, hierarchical, 193–202
cluster analysis, nonhierarchical, 207–210
data manipulations, 55–57
factor analysis, 109–115
logistic regression, 321–335
maximum likelihood estimation technique, 148
measurement models estimation, 14
ordinary least squares estimation, 421
principal components analysis, 67–71
structural model estimation, 426
Statistical decision theory, 256–257
classification rules, development of, 279–281
Statistical distance, 42–44
and Euclidian distance, 43–44
squared statistical distance, 43
Statistical Package for the Social Sciences (SPSS), 67
classification, 256–257, 261
confirmatory factor analysis, LISREL, 148–177
discriminant analysis, 245–262
Helmert contrasts, 360–361
MANOVA, multiple group, 355–366
MANOVA, two-group, 350–355
MANOVA, for two independent variables, 366–371
measurement models estimation, 14
multiple-group discriminant analysis, 294–308
orthogonal contrasts, 360–363
stepwise discriminant analysis, 267–273
univariate normality, 378–379
Statistical significance
of discriminant function, 252–253, 299–302
structural models, 430, 439
tests for canonical correlations, 402–404
Statistical tables
F-distribution, 460–465
percent points of normal probability plot correlation coefficient, 466
simulation percentiles of b_2, 467
simulation probability points of $\sqrt{b_1}$, 467
standard normal probabilities, 457
student's t-distribution critical points, 458
χ^2 critical points, 459
Statistical tests
multivariate normality tests, 380–383
power of test, 375
Type I and Type II errors, 374–375
univariate normality tests, 375–380
Stepwise discriminant analysis, 264–273
backward selection, 265
computer analysis, 267–273
cutoff values for selection criteria, 266–267
forward selection, 265
F-ratio in, 266, 271
Mahalanobis squared distance in, 266
and multicollinearity, 272–273
Ray's V in, 266
selection criteria, 265–266
stepwise selection, 265
Wilks'Λ test statistic in, 251, 266
Stepwise selection procedure, in logistic regression, 329, 331–332
Structural correlations, 404
Structural models, 13–14, 435–437
assessment of, 438–439
direct effects, 425, 450, 452
effects among constructs, 430
effects among endogenous constructs, 450–452
effects of constructs on its indicators, 452–453
effects of endogenous constructs on indicators, 434
effects of exogenous constructs on endogenous constructs, 452
effects of exogenous constructs on indicators, 430
effects of exogenous constructs on indicators of endogenous constructs, 435
estimation procedures in computer packages, 14

Structural models (*continued*)
examples of use, 13, 435–439
and implied covariance matrix, 444–449
indirect effects, 425, 450–451, 452, 452–453
and latent constructs, 13, 14
LISREL estimation, 421–434
measurement model, assessment of, 437
model fit, assessment of, 435–437
model respecification, 435–437
with observable constructs, 420–426, 444–446
overall model fit, 425
saturated models, 421
standardized solution, 426, 435
statistical significance, 430, 439
structural equations, 419–420
total coefficient of determination for, 424
total effects, 425, 451
t-values, 425
with unobservable constructs, 426–435, 446–449
Structure coefficients, 254
Structure loading, 92
Student's *t*-distribution critical points, table of, 458
Sum of cross products, 39
computation of, 39
sum of squares and cross products matrix, 39
Summary measures
computer procedures for, 55–57
data manipulations for, 36–42
types of, 36
Sum of squares, 38–39
for correlated contrasts, 365–366
Sum of squares and cross products matrix, 39
and between-group analysis, 42
and within-group analysis, 40–41
Symmetry, and Mahalonobis distance, 45

T

Territorial map, purpose of, 303
Tied pairs, 325–326
Total coefficient of determination
LISREL, 164
for structural equations, 424
Transformation
Fisher's Z transformation, 383
logit transformation, 383
multivariate normality test, 383
square-root transformation, 383
Triangular inequality, and Mahalonobis distance, 45
T-test, in discriminant analysis, 244–245, 246, 250
Tucker-Lewis index, 160
T-values, in structural model, 425
Two-factor model
computation of, 133–135
with correlated constructs, 147
with LISREL, 165–170
situations for use, 93–96
Two-group discriminant analysis
analytical approach to, 244–245
and classification, 242–244, 278–284
computer analysis, 245–262
discriminant function, 242, 250–254
discriminating variables, evaluation of significance, 246, 250
discriminator variables, selection of, 244–245
equality of covariance matrices, 264
Fisher's linear discriminant function, 245, 277–278
geometric view of, 237–244
identification of set of variables, 238
multiple regression approach to, 262–263
multivariate normality assumption, 263–264
new axis, identification of, 239–242
objectives of, 237, 241, 242
compared to principal components analysis, 241–242
stepwise discriminant analysis, 246, 264–273
validation of discriminant function, 273–274
Two-group MANOVA, 350–355
cell means, 351
computer analysis, 350–355
homogeneity of variances, 351
multivariate significance tests and power, 351, 353
univariate significance tests and power, 353–355
Two-stage least-squares approach, in LISREL, 152
Type I errors
nature of, 374
and violation of equality of covariance matrices, 384
Type II errors, nature of, 375

U

U-method, discriminant function validation, 273–274
Unconstrained analysis, LISREL, 171
Univariate analysis
  number of variables, 5
  objectives of, 238
Univariate effect size, 349
Univariate normal distribution
  kurtosis, of, 375
  zero skewness of, 375
Univariate normality tests, 375–380
  analytical procedures, 378
  computer analysis, 378–379
  graphical tests, 376–377
Univariate significance tests, 250
  for contrasts, 357–359, 360, 362–363
  in multivariate analysis of variance (MANOVA), 348–349
  two-group MANOVA, 353
Unobservable construct, 91

V

Validity
  of canonical coefficients, 409
  cluster analysis, 221
Variables, number for measurement scales, 3–4
Variable space, graphical representation of data in, 45–46
Variance
  computation of, 38
  generalized variance, 39, 50–51, 54–55
  situations for use, 38
  of standardized variables, 39
Varimax factor rotation, 119–120, 138
Vectors, 19–32
  addition of, 21–22
  arithmetic operations on, 25–26
  basis vectors, 25, 31
  in Cartesian coordinate system, 23–25
  centroid, 45–46
  changing basis, 31–32
  dimensionality of, 30–31
  distance and angle between two vectors, 27
  equivalent vectors, 20
  initial and terminal point, 19
  linear combination of, 26–27
  multiplication by real number, 20–21
  multiplication of two vectors, 22–23
  norm of, 20
  null vector, 21
  oblique basis, 31
  orthonormal vectors, 25, 31, 32
  projection into subspace, 28–29
  projection of one onto another, 23
  projection vector, 23, 27–28
  reflection, 21
  representation of points with respect to new axes, 32–33
  scalar product of two vectors, 20–21, 27–28
  signed length of, 28
  standard basis vectors, 25
  subtraction of, 22

W

Ward's method, hierarchical clustering method, 193, 217
Wilks' Λ test statistic
  in discriminant analysis, 246–250
  relationship to F-ratio, 348
  in stepwise discriminant analysis, 251, 266
  testing for canonical correlations, 402–403
Within-group analysis, 40–41
  sum of squares and cross products matrices, 40–41

X

χ^2 critical points, table of, 459
χ^2 statistic, null hypothesis testing, 157, 162